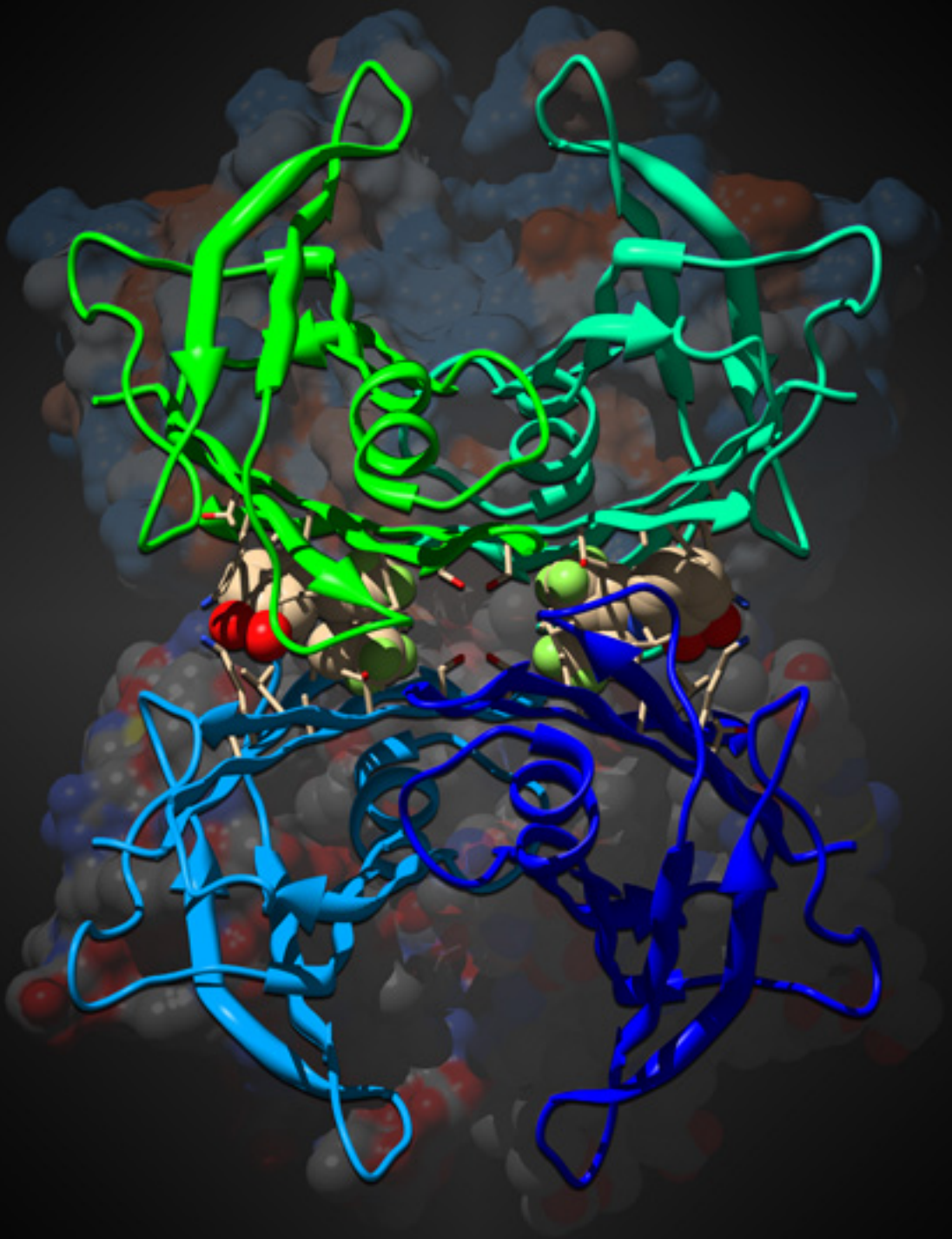
Virtual Screening with Sense and Sensibility in the Search of New Amyloid Inhibitors



Carlos J. V. Simões

UNIVERSIDADE DE COIMBRA

Virtual Screening with Sense
and
Sensibility in the Search of New Amyloid Inhibitors

Carlos J. V. Simões

A dissertation presented to the Faculty of Science
and Technology for the attainment of the degree of
Doctor of Philosophy in Biological Chemistry

October 2011

Trabalho desenvolvido sob orientação científica do Professor Doutor Rui Manuel Meireles de Brito, no grupo de Biologia Estrutural e Computacional do Centro de Neurociências e Biologia Celular, Universidade de Coimbra, e do Doutor Richard Michael Jackson, no grupo de Bioinformática do *Institute of Molecular and Cellular Biology*, Universidade de Leeds, Reino Unido.

Trabalho parcialmente financiado pela Fundação para a Ciência e a Tecnologia no âmbito do programa POCI 2010
(SFRH/BD/29357/2006)

Programa Operacional Ciência e Inovação 2010
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR

I met a traveller from an antique land
Who said: – Two vast and trunkless legs of stone
Stand in the desert. Near them on the sand,
Half sunk, a shatter'd visage lies, whose frown
And wrinkled lip and sneer of cold command
Tell that its sculptor well those passions read
Which yet survive, stamp'd on these lifeless things,
The hand that mock'd them and the heart that fed.
And on the pedestal these words appear:
"My name is Ozymandias, king of kings:
Look on my works, ye mighty, and despair!"
Nothing beside remains: round the decay
Of that colossal wreck, boundless and bare,
The lone and level sands stretch far away.

Eu encontrei um viajante de uma antiga terra
Que disse: – Duas imensas e destroncadas pernas de pedra
Erguem-se no deserto. Perto delas na areia
Meio enterrada, jaz uma viseira despedaçada, cuja fronte
E lábio enrugado e sorriso de frio comando
Dizem que seu escultor bem suas paixões leu
Que ainda sobrevivem, estampadas nessas coisas inertes,
A mão que os escarneceu e o coração que os alimentou.
E no pedestal aparecem estas palavras:
"O meu nome é Ozymandias, rei dos reis:
Contemplem as minhas obras, ó poderosos, e desesperai-vos!"
Nada mais resta: em redor a decadência
Daquele destroço colossal, sem limite e vazio
As areias solitárias e planas espalham-se para longe.

Ozymandias, by Percy Bysshe Shelley

Aos meus pais

e em memória do meu Avô J. V.

To my parents

and in memory of my Grandfather J. V.

Acknowledgements

Agradecimentos

First and foremost, I thank Professor Rui Manuel Meireles de Brito for the trust laid on me from the early stages of this project, for the enthusiasm shown throughout its completion, and for giving me the freedom to play – which is as vital as water for fish and the only way to escape the deluge of hype-based science. I also acknowledge his critical supervision and scientific contribution in the revision of all manuscripts, including this thesis. I feel privileged to have worked with him at the Structural and Computational Biology group at the Centre for Neuroscience and Cell Biology, as student of the Faculty of Science and Technology (Chemistry Department) of the University of Coimbra.

Agradeço ao Professor Doutor Rui Manuel Meireles de Brito, em primeiro lugar, a confiança depositada em mim desde as etapas precoces deste projecto, o entusiasmo demonstrado ao longo da sua compleição e a liberdade para "brincar" – que é tão vital como a água para os peixes e a única forma de fugir do dilúvio da ciência baseada em mediatismo. Agradeço ainda a sua supervisão crítica e o seu contributo científico na revisão de todos os manuscritos, incluíndo esta tese. Sinto-me privilegiado por ter trabalho com ele no grupo de Biologia Estrutural e Computacional do Centro de Neurociências e Biologia Celular, enquanto aluno da Faculdade de Ciências e Tecnologia (Departamento de Química) da Universidade de Coimbra.

I thank Doctor Richard Michael Jackson for accepting me at the Bioinformatics Group of the Institute of Molecular and Cellular Biology, for his guidance, his remarkable and continuous support, and his enthusiasm with my work. I also thank him for allowing me to participate in his own research projects, which provided me a much broader experience.

Agradeço ao Doutor Richard Michael Jackson a minha aceitação no grupo de Bioinformática do Instituto de Biologia Molecular e Celular da Universidade de Leeds, a sua orientação, o apoio notável e contínuo, bem como o entusiasmo com o meu trabalho. Agradeço-lhe ainda a minha participação nos seus projectos de investigação, oferecendo-me uma experiência mais alargada.

I thank all the members of the Leeds Bioinformatics Group, past and present, for their warm support, for the insightful discussions during the Journal Club sessions and for listening. Particular thanks go to (in alphabetical order): Nick Burgoyne, Stephen Bridgett, James Dalton, Joel Dockray, Jon Fuller, Sarah Kinnings, Mahesh Kulharia, Trishna Mukherjee and Monika Rella. Stephen Bridgett and Monika Rella deserve a special mention for the technical advice and emotional support throughout my PhD project. The following names ought to be mentioned as well: Sean Killen, Margaret Oldfield, Sheena Radford and David Westhead. These people contributed significantly to the success of this project.

Agredeço a todos os elementos do grupo de Bioinformática de Leeds, anteriores e actuais, o seu apoio caloroso, as discussões profícuas durante as sessões de *Journal Club*, e o facto de me ouvirem. Agradecimentos particulares vão para (por ordem alfabética): Nick Burgoyne, Stephen Bridgett, James Dalton, Joel Dockray, Jon Fuller, Sarah Kinnings, Mahesh Kulharia, Trishna Mukherjee and Monika Rella. O Stephen Bridgett e a Monika Rella merecem um menção especial pelos aconselhamentos técnicos e o apoio emocional ao longo do meu projecto de doutoramento. Os seguintes nomes devem ser referidos: Sean Killen, Margaret Oldfield, Sheena Radford e David Westhead. Estas pessoas contribuíram determinantemente para o sucesso deste projecto.

The work included in Chapters 6 and 7 of this thesis was developed at the Chemistry Department of the University of Coimbra. I thank all the members of the Structural and Computational Biology group, past and present, for the broad-spectrum discussions during the group's weekly meetings and for technical advice in the fields of both computational and experimental biochemistry. Particular thanks go to (in alphabetical order): Zaida Almeida, Bruno Correia, Pedro Cruz, Tiago Faria, Elsa Henriques, Catarina Jesus, Nuno Loureiro-Ferreira, J. Rui Rodrigues, Cândida Silva, and Daniela Vaz. I feel Catarina Jesus deserves a special mention for her contribution to the experimental testing of virtual screening hits presented in Chapter 7. The following names must also be mentioned: Pedro V. Alberto, Pedro Almeida, Maria João Araújo, Helena Figueiredo, Miguel Oliveira, and Sílvia Sousa. Without these people, this project would never have been completed.

O trabalho incluído nos Capítulos 6 e 7 desta tese foi desenvolvido no Departamento de Química da Universidade de Coimbra. Agredeço a todos os elementos do grupo de Biologia Estrutural e Computacional, anteriores e actuais, as discussões de amplo espectro durante as reuniões de grupo semanais e pelos aconselhamentos técnicos nas áreas da bioquímica computacional e experimental. Agradecimentos particulares vão para (por ordem alfabética): Zaida Almeida, Bruno Correia, Pedro Cruz, Tiago Faria, Elsa Henriques, Catarina Jesus, Nuno Loureiro-Ferreira, J. Rui Rodrigues, Cândida Silva e Daniela Vaz. Sinto que a Catarina Jesus merece uma menção especial pelo seu contributo no ensaio experimental dos compostos seleccionados por rastreio virtual, apresentado no Capítulo 7. Os seguintes nomes também devem ser referidos: Pedro V. Alberto, Pedro Almeida, Maria João Araújo, Helena Figueiredo, Miguel Oliveira e Sílvia Sousa. Sem estas pessoas, este projecto jamais teria sido realizado.

I thank my mother and my father for helping me keeping a level head, for their critical views of the World and of how I integrate into it, but most of all for respecting my choices. It has always been like this, really, ever since I remember. I have always felt respected, and I admire them for that.

Agradeço à minha mãe e ao meu pai o apoio na manutenção da minha sanidade mental, a sua visão crítica do Mundo e da forma como me integro nele, mas principalmente o respeito pelas minhas escolhas. Em verdade, foi sempre assim, desde que me lembro. Sempre me senti respeitado e admiro-os por isso.

I thank all my friends, past and present, for emotional support, for sharing the joy of being alive and for "stabbing" me in the front when I need it. I feel it would be neither fair nor pratical to mention names.

Agradeço aos meus amigos, anteriores e actuais, o apoio emocional, a partilha do gozo de estar vivo e as "facadinhas" pela frente quando preciso. Sinto que não seria nem justo nem exequível mencionar nomes.

Last but not least, I thank UMIC – Knowledge Society Agency and *Fundação para a Ciência e a Tecnologia* for financial support throughout this project and for a PhD grant (SFRH/BD/29357/2006).

Por último, agradeço à UMIC – Agência para a Sociedade do Conhecimento e à Fundação para a Ciência e a Tecnologia o apoio financeiro ao longo deste projecto e por uma bolsa de doutoramento (SFRH/BD/29357/2006).

Contents

Abstract

Inhibition of amyloid fibril formation by stabilisation of the native form of the protein transthyretin (TTR) is a viable approach for the treatment of severe diseases, such as familial amyloid polyneuropathy, familial amyloid cardiomiopathy and senile systemic amyloidosis. Several small organic molecules have evidenced ability to stabilise TTR and thus inhibit fibrilization in vitro, but they display impairing issues of solubility, affinity for TTR in the blood plasma and/or adverse effects. Notwithstanding, the stabilisation strategy has been gaining momentum in the field of amyloid research, as a very first therapeutic drug against such untreated diseases has now reached the market. The results of phase II/III clinical trials have revealed success in ameliorating pathological symptoms in 60% of the treated patients, a figure that highlights both the adequacy of the strategy and the prospect of further improvements.

Virtual screening (VS) relates to the use of computational techniques in the search of promising drug leads amongst virtual catalogues of thousands or millions of chemical compounds, thereby prioritizing candidates with high likelihood of being active for experimental evaluation. The first goal of this project was to evaluate a large battery of virtual screening methods and identify those that are more suited for amyloid targets, using transthyretin amyloid as a model target.

Given the high amount of structural information on TTR accessible in public repositories, several structure-based approaches were first explored in order to gain insights about the specificities of ligand binding to this target protein. Five docking programs employing a ten of different scoring functions were exhaustively tested, leading to the identification of an algorithm that is capable of predicting ligand binding modes that mimic those determined through experimental techniques. Moreover, the knowledge about strong stabilisers and their differential binding modes to the two TTR binding sites was used to build five receptor- and ligand-based pharmacophore hypotheses based on the overlapping physicochemical features of the ligands and of the TTR receptor sites.

Four protein- and ligand-based VS methods were evaluated for their ability to identify novel TTR stabilisers, making use of a benchmark set comprising known actives and carefully selected decoy molecules. Amongst the evaluated techniques are (*i*) 2D similarity searches with chemical hashed fingerprints, pharmacophore fingerprints and UNITY fingerprints, (*ii*) 3D searches based on shape, chemical and electrostatic similarity, (*iii*) LigMatch, a new ligand-based method which uses multiple templates and combines 3D geometric hashing with a 2D pre-selection process, (*iv*) molecular docking into consensus X-ray crystal structures of TTR.

The results of a benchmark for a set of high-throughput docking protocols against 40 protein targets of pharmaceutical relevance highlight the importance of a thorough validation of VS methods as an essential step to achieve meaningful results in a screening campaign against any target. The same results motivated the implementation of several academic-free docking-based protocols onto a freely-accessible distributed computing platform, as a means to allow large-scale screening campaigns against TTR amyloid with the help of volunteer citizens and to foster further drug discovery endeavours in the academia.

The potential of 22 best-performing VS protocols to retrieve promising new leads is illustrated by ranking a tailored library of 2.3 million commercially available compounds. Our predictions show that the top-scoring molecules possess distinctive features from the known TTR binders, holding better solubility, fraction of halogen atoms and binding affinity profiles combining the two binding sites of TTR. Forty-seven commercially available candidates were purchased from their respective chemical suppliers. Thirty-eight of these have already been experimentally tested for inhibition of TTR fibril formation. Thirty-one compounds showed inhibitory activity to some extent, of which five revealed to be *excellent* inhibitors, allowing no more than 40% fibril formation *in vitro* at a protein-ligand concentration *ratio* of 1:2.

To sum up, this project represents a first attempt to rationalize the utilisation of computational methods against TTR amyloid. As a result, at least two *novel* and *inventive* lead compounds that inhibit fibril formation *in vitro* have been discovered, and a new drug discovery platform based on computational techniques and tailored towards the identification of new amyloid inhibitors has been developed.

Sumário

A inibição da formação de fibras amilóides por estabilização da forma nativa da proteína transtirretina (TTR) é uma abordagem viável para o tratamento de doenças debilitantes, como a polineuropatia amiloidótica familiar, a cardiomiopatia amiloidótica familiar e a amiloidose senil sistémica. Várias moléculas orgânicas previamente identificadas evidenciaram capacidade de estabilizar a TTR e assim inibir a fibrilização *in vitro*, mas demonstram problemas de solubilidade, afinidade para a TTR no plasma sanguíneo e/ou efeitos adversos. Ainda assim, a estratégia de estabilização da TTR tem vindo a despertar atenção no campo da investigação em amilóide, com recente introdução no mercado farmacêutico de um primeiríssimo fármaco contra tais doenças. Os resultados de ensaios clínicos de fase II/III revelaram sucesso no aliviar dos sintomas patológicos em 60% dos pacientes tratados, um número que enfatiza tanto a adequabilidade da estratégia seguida como a perspectiva de futuras optimizações.

O conceito de rastreio virtual (VS) relaciona-se com a utilização de técnicas computacionais na procura de compostos-líder com potencial terapêutico de entre catálogos virtuais com milhares ou milhões de compostos químicos, "prioritizando" assim candidatos com elevada probabilidade de se revelarem activos para avaliação experimental. O primeiro objectivo deste projecto foi avaliar uma ampla bateria de métodos de rastreio virtual e identificar os mais apropriados para alvos amilóide, usando o caso de amilóide por transtirretina (TTR) como modelo de estudo principal.

Dada a elevada quantidade de informação estrutural sobre TTR acessível em repositórios públicos, várias abordagens assentes na utilização da estrutura foram primeiramente exploradas de forma a ganhar introspecções acerca das especificidades da associação de ligandos a esta proteína alvo. Entre vários aspectos intrigantes desta associação está a natureza elusiva dos fenómenos de cooperatividade observados, apesar da completa equivalência em composição dos dois locais de ligação da TTR: quando ocorre associação de um primeiro ligando a um primeiro local de ligação, a associação de um segundo ligando idêntico ao segundo local de ligação é frequentemente afectada, maioritariamente pela negativa). Alem disto, a TTR apresenta uma propensão clara para se associar a compostos hidrofóbicos e electronegativos, contendo núcleos bifenilo e/ou amplamente halogenados.

Neste projecto foram testados exaustivamente vários programas de acoplamento molecular – nomeadamente, AutoDock 4, AutoDock Vina, eHiTS, FRED e GOLD – e empregando diferentes funções de pontuação – entre as quais, a função de energia livre do AutoDock 4, a função de pontuação do Vina, a função de pontuação do eHiTS, a Chem-

Gauss3, ChemGauss4, ChemScore, GoldScore, ScreenScore e ainda o potencial de pontuação analítica (ASP). Este trabalho culminou na identificação do AutoDock 4 como o algoritmo de acoplagem com melhor desempenho na determinação computacional dos modos de ligação que mimetizam aqueles determinados experimentalmente, proximamente seguido pelo AutoDock Vina. Em acréscimo, o conhecimento reunido sobre compostos fortemente estabilizadores, bem como os seus modos de ligação diferenciais em cada um dos locais de ligação da TTR, foi usado para construir 5 hipóteses do farmacóforo baseadas em funções físico-químicas consensuais quer dos ligandos quer dos locais de ligação da TTR. Apesar de a validação teórica destes modelos ser questionável, por limitações inerentes à pontuação realizada pelo programa de pesquisas de farmacóforos disponível, eles revelaram uma utilidade crucial ao serem combinados com outras metodologias de rastreio virtual exploradas ao longo do projecto, com vista à selecção final de compostos a testar experimentalmente.

Ao longo do projecto foram ainda testadas quatro metodologias da rastreio virtual centradas quer na estrutura de ligandos quer na estrutura proteica, tirando partido de uma biblioteca vocacionada para *benchmark* e composta por moléculas activas conhecidas e moléculas inactivas ("engodos") criteriosamente seleccionadas. De entre as técnicas de rastreio virtual avaliadas encontram-se (*i*) pesquisas de similaridade bidimensionais usando impressões digitais com correspondência química, impressões digitais baseadas em farmacóforo e impressões digitais do tipo UNITY; (*ii*) pesquisas de similaridade tridimensionais baseadas na forma, em complementaridade química e em similaridade electrostática; (*iii*) o programa LigMatch, uma metodologia nova capaz de usar múltiplas moléculas de referência e de combinar alinhamentos tridimensionais de correspondência com um processo de pré-selecção bidimensional; (*iv*) acoplamento molecular contra estruturas cristalográficas consensuais da proteína TTR. Globalmente, as pesquisas de similaridade bidimensionais oferecem o melhor desempenho de rastreio virtual, quer em termos de enriquecimento precoce quer em de enriquecimento global (i.e. a capacidade de ordenar uma base de dados de compostos activos e inactivos de forma a que um maior número possível de compostos activos preencha as primeiras posições da lista ordenada). No entanto, este melhor desempenho tem como custo a selecção de compostos bastante idênticos aos utilizados como referência. Em contraste, o rastreio virtual assente no método de acoplagem molecular tem a capacidade de sugerir potenciais ligandos mais inovadores, sob o prejuízo da "prioritização" de mais falsos positivos (principalmente enviesada para compostos de elevado peso molecular). Os métodos baseados na estrutura tridimensional de compostos, e em particular o novo método LigMatch, oferecem um melhor compromisso entre desempenho e inovação.

Os resultados de um trabalho de *benchmark* exaustivo realizado ao longo deste projecto com um conjunto de protocolos de acoplamento molecular de alto débito contra 40 alvos proteicos de relevância farmacêutica enfatizam a importância de uma validação criteriosa das metodologias de rastreio virtual, como passo essencial para o alcançar de resultados de significado numa campanha de rastreio contra qualquer alvo. É bastante claro que o modelo "uma receita agrada a todos" não tem aplicação no domínio do rastreio virtual de ligandos. Ainda assim, estes mesmos resultados motivaram a implementação destes protocolos acessíveis gratuitamente a investigadores académicos numa plataforma de computação distribuída, também ela de livre acesso, como forma de permitir campanhas de rastreio virtual contra amilóide por TTR em larga escala, e de fomentar futuros esforços para a descoberta de novos fármacos na academia com a ajuda e o envolvimento de cidadãos voluntários.

Neste trabalho, realça-se ainda o potencial de 22 protocolos de rastreio virtual com melhor desempenho em *benchmark* para seleccionar compostos-líder promissores, através da ordenação de um biblioteca personalizada de 2,3 milhões de compostos químicos disponíveis comercialmente. As nossas previsões indicam que as moléculas encontradas no topo das listas ordenadas possuem características distintivas das de inibidores de TTR já conhecidos, detendo ainda melhores perfis de solubilidade, fracção de halogéneos e afinidade de ligação combinada para os dois locais de ligação da TTR. Para validar o trabalho computacional executado, foram adquiridos 47 candidatos disponíveis comercialmente aos respectivos fornecedores. Destes, 38 compostos foram já ensaiados experimentalmente para averiguar a sua capacidade de inibição da formação de fibras amilóides. 31 compostos demonstraram actividade inibitória em alguma extensão, dos quais 5 correspondem a inibidores excelentes, permitindo uma formação de fibras *in vitro* não superior a 40% a uma razão de concentrações de 1:2 entre proteína e ligandos.

Globalmente, este projecto representa uma tentativa pioneira de racionalizar a utilização de métodos de desenho de fármacos computacionalmente assistido contra amilóide por TTR. Como resultado, foram descobertos pelo menos dois compostos-líder *novos* e *inventivos*, que inibem a formação de fibras *in vitro*. Foi ainda desenvolvida uma nova plataforma para a descoberta de novos inibidores de amilóide assente em técnicas computacionais.

List of abbreviations

2D-PF	Two-Dimensional Pharmacophore Fingerprints
AD4	AutoDock 4
ADMET	Absorption, Distribution, Metabolism, Excretion and Toxicity
AUC	Area Under the (Enrichment) Curve
CADD	Computer-Aided Drug Design
CHF	Chemical Hashed Fingerprints
CPU	Central Processing Unit
EF	Enrichment Factor
FAP	Familial Amyloid Polyneuropathy
FAC	Familial Amyloid Cardiomiopathy
HBA	Hydrogen Bond Acceptor
HBD	Hydrogen Bond Donor
HPC	High Performance Computing
HTD	High-Throughput Docking
HTS	High-Throughput Screening
MD	Molecular Dynamics
NMR	Nuclear Magnetic Resonance
NSAID	Non-Steroidal Anti-Inflammatory Drug
PDB	Protein Data Bank
PSA	Polar Surface Area
PVM	Parallel Virtual Machine
QRY	Pharmacophore Query
RMSD	Root Mean Square Deviation
ROC	Receiver Operating Characteristic
ROC AUC	Area Under the (ROC) Curve
ROCE	ROC Enrichment
SBDD	Structure-Based Drug Design
STM	Single Template Mode
TTR	Transthyretin
VS	Virtual Screening
WT	Wild-Type

Thesis outline

The thesis consists of seven chapters – including a general introduction and results divided into six chapters – plus an epilogue, an appendix, and a section with references and notes. At the end of the thesis, a short list of the most relevant publications produced during the course of this PhD project is provided.

Chapter 1 is an introductory chapter. It first provides a brief overview of modern drug discovery and the role of computer-aided drug design methods in the post genomic era. The reader is introduced to the basic principles and the fundamental techniques in molecular modelling, with special emphasis on the main methodologies employed throughout the project. A comment on the critical role of academia in the moving forward of the drug discovery field softens the transition to the second half of the chapter, where the relevance of the search for amyloid inhibitors is discussed, along with the implications of new therapeutics in pathologies like Alzheimer's and Parkinson's disease. Transthyretin is introduced as a model protein target in amyloid, in view of its involvement in severe and currently untreated diseases, such as familial amyloid polyneuropathy, familial amyloid cardiomiopathy and senile systemic amyloidosis, and a short review of the currently known transthyretin amyloid inhibitors is provided.

Chapter 2 focuses on the use of structure-based approaches to study transthyretin-ligand interactions. Currently available structural data on transthyretin complexes are summarized and organised, and the results of structural quality evaluations are reported. An exhaustive study comprising five different docking programs is presented as a means to elucidate the most reliable structural models of transthyretin and the best performing algorithms in handling the intricate specificities of its binding sites.

Chapter 3 reports the results of several attempts to identify structural and chemical differences between the two equivalent binding sites of transthyretin. Five receptor- and ligand-based pharmacophore models are revealed, each one derived from a different set of transthyretin amyloid inhibitors exploring a specific region of each binding site. The results of selectivity evaluations of the models are reported and their potential use in the virtual screening of new amyloid inhibitors is discussed.

Chapter 4 is the *intermission*. The chapter provides a quick *tour* through chemical space with eyes set on its merger with biological space, and explores the concepts of molecular diversity, similarity and searches within large data repositories. The filtering of virtual libraries of chemical compounds using different molecular descriptors is approached as a primer to the construction of (*i*) a benchmarking set for evaluation of the virtual screening

performance of several methods and (*ii*) a tailored screening set of approximately 2.3 million compounds for the virtual high-throughput screening of new transthyretin amyloid inhibitors.

Chapter 5 reports the results of an evaluation of a large battery of virtual screening protocols assembled to target transthyretin amyloid. Amongst the evaluated group of techniques are two-dimensional (2D) and three-dimensional (3D) similarity search methods, a new method combining both 2D and 3D descriptors to make use of multiple templates, and docking and scoring. For each group of techniques, a best-performing virtual screening protocol is identified and its prospective application is discussed.

Chapter 6 presents a study of the application of the best-performing docking and scoring protocols assembled in Chapter 5 to a set of 40 targets encompassing five protein families of pharmacological interest. The academic-free nature and the satisfactory virtual screening performance of the selected protocols are discussed under the light of their potential implementation as part of a large-scale volunteer computing resource freely accessible to academics.

Chapter 7 reveals the implementation of a large-scale, multi-protocol, virtual screening campaign in search of new lead compounds to inhibit amyloid formation by transthyretin. The utilisation of several hardware resources, from an individual laptop, through a parallel virtual machine and two HPC clusters, all the way to a large volunteer computing platform is reviewed. The experimental validation of the developed virtual screening approaches led to the identification of potent inhibitors of transthyretin amyloid.

The **Epilogue** highlights the global conclusions drawn from this project and provides a brief overview of my own perspectives, hopes and fears in the field of computer-aided drug design and discovery, both at the academia and the industry.

Chapter 1

General introduction

"Forget the merger of the Web with TV, of phones with computers. The most powerful convergence underway today involves silicon, the substrate of computing, and carbon, the substrate of life."

[Tom Petzinger, in *The Wall Street Journal*]

1. Drug discovery in the XXIst century

The complete sequencing of the human genome has opened up exciting new avenues for groundbreaking medical research and biomedical applications. However, the excitement behind this remarkable achievement is only comparable with the frustration felt before the figures describing the lack of success observed in the drug discovery field over the last decades. In 2010, the Food and Drug Administration (USA) approved no more than 21 drugs; a modest number showing that the pharmaceutical industry has not yet escaped the so-called "target-rich, lead-poor" predicament. The research and development process leading to a new drug is a long and expensive road (**Figure 1.1**). There is an increasing demand for more effective and safer drugs that can be administered to humans for longer periods of time. Efficacy and safety alone are not sufficient. Besides producing the desired response with minimal side effects, new drugs must also demonstrate to be more cost effective than the treatments that went before them.

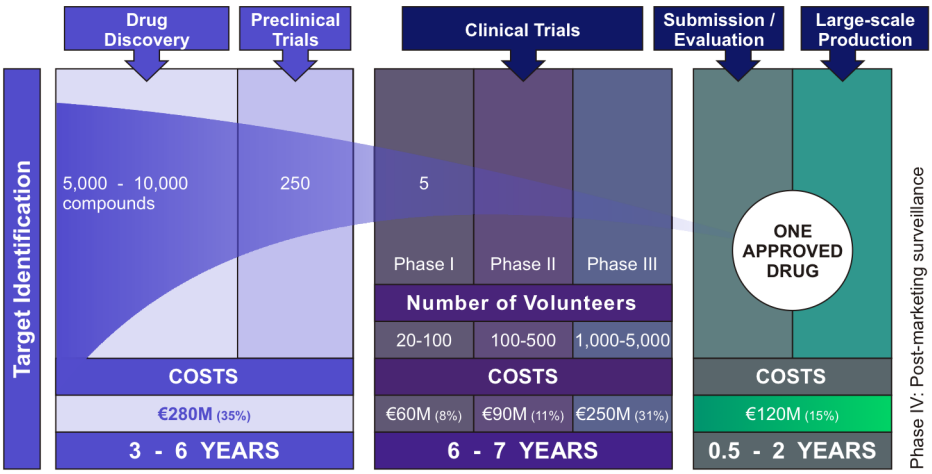


Figure 1.1. Scheme of the research and development (R&D) setting leading to the approval of a new drug.

Most modern drug discovery programs follow a similar drug discovery and development pipeline, as illustrated in **Figure 1.2**. The process begins with the identification and validation of a protein target, whose activity can be modulated to treat a particular disease. Many protein targets fall into two major classes: enzymes and receptors. The goal of drug design is thus to identify small molecules that can inhibit the activity of an enzyme, or can modulate the activity of a receptor by antagonism, agonism or inverse agonism. Antagonists inhibit the effect of the natural substrate, whereas agonists create the same or stronger effect as the substrate molecule, and inverse agonists produce an effect that is opposite to that of an agonist. Successful drugs often accomplish their activity by compet-

ing with an endogenous substrate at a protein target binding site. The development of protein-based therapeutics is also highly valuable in the pharmaceutical and biotechnology industries, but this thesis is mainly focused on small molecule drug discovery.

1.1. The drug discovery pipeline

In general lines, a drug target is first identified as being important to the pathology of interest through either biological or genetic investigations. The use of modern biology including functional genomics (such as micro-array technologies) and proteomics has revolutionised the way by which new targets are identified. Once a target has been validated and characterised, the next step is to develop an assay that can be used to determine if a compound can modulate a target's activity. The nature of the assay can vary greatly, but typically the complexity of assays will increase and throughput decrease as progress is made across the drug discovery pipeline (**Figure 1.2**). Thus, in a "lead finding" stage, a high throughput screening (HTS) campaign of hundreds of thousands of compounds (often the entire corporate collection of a pharmaceutical company) can be carried out to discover one or more *hits* with affinities in the micromolar range (i.e. with binding constants ranging from 10 micromolar to the low nanomolar range). This requires robust assays with minimum human input that can be run with large automated robots. After follow-up tests to confirm that the discovered molecules affect the target by a specific mechanism, rather than through nonspecific binding, the hits are narrowed to a limited number of new chemical entities. Further along the pipeline, these lead series enter the lead optimisation stage. They are modified using medicinal chemistry methods in order to decrease their dissociation constant (K_d) to the low nanomolar range, thereby producing potent lead molecules [1]. The process of optimising these molecules into *candidate* drugs that fulfil the desired properties of potency, absorption, bioavailability, metabolism and safety is typically the lengthiest stage in the drug discovery process. New compounds will be synthesised based on some chemical insight, knowledge of the target structure or existing active ligands. Chemical insight can include, for example, knowledge that one chemical group can often be replaced by another similar group without losing biological activity (bioisosterism). As each new compound is synthesised, its activity will be measured and the impact of the chemical modification assessed and used to drive further synthesis. The ideal output of the optimisation process is an enhanced compound series and associated structure-activity relationships (SAR) data. The lead-to-candidate stage is regarded as a multidimensional optimisation problem, for it involves exploring the somewhat limited chemical space of the congeneric series of the lead compounds.

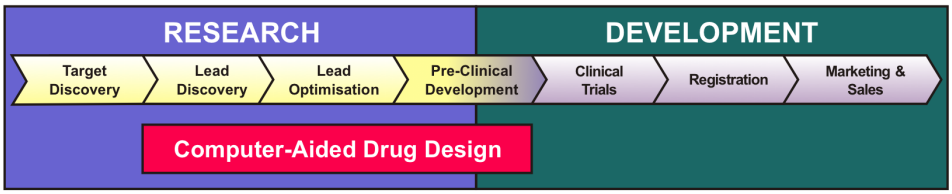


Figure 1.2. A global scheme illustrating the various stages of the drug discovery process. As disclosed, computer-aided drug design has its largest impact in lead discovery, lead optimisation and, increasingly, during pre-clinical development phases.

Once preclinical animal safety studies on the candidate molecule have been conducted, the investigational drug must take on three types of clinical trials: Phase I clinical trials involve a restricted study on healthy human subjects to confirm safety; Phase II involves a larger study on a patient population to confirm efficacy; Phase III involves a large study of patients to gather additional information about safety and efficacy. Disappointingly, more than 90% of the candidate drugs entering clinical trials will fail to reach the market, most often as a result of poor biopharmaceutical properties, toxicity, or lack of efficacy [2]. Given the attrition of so many potential pharmaceuticals during the late stages of clinical development, the successful launch of a single new drug takes around 14 years and is estimated to exceed US$800 million [3].

Computer-aided drug design has a role to play at each stage of the drug design and discovery process and is the key to increasing and exploring the molecular knowledge and chemical insight that is a prerequisite for success in modern drug discovery.

1.2. The role of computer-aided drug design

In all endeavours involving a highly technical component, from architecture, through car design, all the way to washing powder formulation, the utilisation of computers to improve quality of work and increase efficiency has been embraced. It is now difficult to imagine how a new drug could be designed in the XXIst century with absolutely no contribution from computer-aided drug design (CADD). The addition of CADD to the R&D approaches of a company, could lead to a reduction in the cost of drug design and development by more than 50% [4]. The use of CADD has led to the discovery of Indinavir, the HIV protease inhibitor. Amongst the increasing examples of compounds designed with a strong CADD component are also Dorzolamide (glaucoma treatment), Zanamivir (influenza treatment) and Lopinavir (HIV treatment). For some compounds, CADD will be the driving force in their design; for others a less pronounced but still significant contributing factor.

The most powerful and effective drug discovery strategies involve a tight integration of computational and experimental techniques at each stage of the process. In particular, structural information can be exploited from the identification of the target, all the way to the design of a bioavailable drug via structure-based drug design. Therefore, by bridging and building on resources in bioinformatics, structural biology, and structure-based drug design, structural bioinformatics can accelerate the quest for a potential drug [5].

1.3. Proteins as molecular targets in drug discovery

Proteins are astonishingly complex molecular machines involved in virtually every critical process of biology. They evolved through selective pressure to carry out all sorts of tasks, from simple, static, support functions, through the synthesis, break down or transport of other molecules, all the way to the regulation of hyper-complex processes, such as intra and extracellular communication or the control of gene expression. Built from a small *repertoire* of 20 building blocks called amino acids, proteins vary not only in function but also in size and shape, in physical and chemical properties.

Despite their vital roles to all living (and some non-living) organisms, proteins are also associated with malfunctioning and disease. The cause of malfunctioning of proteins can be extrinsic or intrinsic. Extrinsic malfunctions are often due to failure of controlling mechanisms, whereas intrinsic problems are usually the result of mutations, which affect the structure and stability of proteins and thus their activity. A deep understanding of how proteins work (and malfunction) opens up opportunities of interfering with their function in a therapeutic or regulating manner.

Most proteins exert their biological functions as components of protein complexes. In recent years and as result of the study of the crucial role of protein-protein interactions for both physiological and pathological processes, the modulation of specific protein-protein interactions has been considered of great pharmaceutical interest. The most prominent example is perhaps that of protein kinases, implicated in multiple processes related to cancer. Insoluble protein fibrils (as in amyloid) also result from deviant protein-protein interactions and ensuing assembly of conformational intermediates found along the unfolding pathway of certain proteins. These are involved in important human amyloid diseases – including Alzheimer's and Parkinson's diseases, type 2 diabetes and Creutzfeld-Jakob disease.

1.3.1. Molecular recognition

The process of molecular recognition is fundamental to biology. In many biological processes the first step consists of the binding of a ligand to a *target* receptor. The term 'ligand' derives from the Latin *ligare*, meaning *tie*, and is employed to depict a molecule that binds to a receptor molecule by non-covalent forces. Ligands, including natural substrates and drugs, bind to many different receptors, such as enzymes, antibodies, membrane-bound proteins and DNA. The driving forces for ligand binding include complementarities in shape and electrostatics between the binding site surfaces and the ligand. Thus, Coulombic and van der Waals interactions are key players in the process, aided by the formation of hydrogen bonds and solvation forces. Ligand binding to a receptor often produces a biological effect. For example, the binding of a natural substrate to an enzyme will trigger a catalytic reaction, while the binding of a drug to the same enzyme may inhibit that reaction.

Molecular recognition between a ligand and a receptor is necessary for binding to occur. Emil Fischer first proposed the theory of molecular recognition in 1894, in the form of the *lock and key* model [6]. According to this theory, high affinity for a given receptor can only be achieved by a ligand that holds a good geometric fit to that receptor. This fairly simplistic view has been extended to include the conformational changes that the two molecules may undergo during complex formation [7]. The so-called *induced fit* model proposes a mechanism whereby an open form of an enzyme binds a substrate, which induces the enzyme to undergo a conformational change and move into a closed form surrounding the substrate. Catalysis occurs in this closed form and the enzyme opens again to release the product. More recently, it was proposed that optimally fitting conformers of both receptor and ligand can be found in solution, and that there is a shift in the equilibrium towards the best-fitting conformers for both molecules upon binding [8].

1.3.2. Kinetics of binding

In solution, and under thermodynamic equilibrium conditions, a receptor, R, and ligand, L, associate to form a non-covalent, reversible, receptor-ligand complex, RL (Equation 1.1). k_{on} is the association rate constant for complex formation, whereas k_{off} is the dissociation rate constant.

$$R + L \underset{k_{off}}{\overset{k_{on}}{\rightleftharpoons}} RL$$

Equation 1.1

It is possible to determine equilibrium constants that depict the strength of the interaction by determining equilibrium concentrations of R, L and RL experimentally. The binding affinity between the ligand and receptor can either be expressed as the dissociation constant, K_d (in M), or as the association constant, K_a (in M^{-1}) (Equation 1.2). If the concentration of ligand present is less than the value of K_d, then only a small proportion of the receptor molecules will be associated with the ligand. If the concentration of ligand is the same as the K_d, then half of the receptor molecules will be in the ligand bound state. In a biological system, values of K_d generally range from 10^{-4} M (a loose association) to 10^{-16} M (a tight association).

$$K_d = 1/K_a = [R][L] / [RL]$$

Equation 1.2

The pharmacological literature frequently reports the IC_{50} value, which is the concentration of an inhibitor required to reduce the binding of a ligand (or rate of reaction) by half [9]. IC_{50} values are relatively easy to measure and are often used to compare inhibitors with one another in competitive binding assays. However, since the IC value is dependent on the amount of ligand available to the receptor, comparisons between experimental data obtained under different conditions are unfeasible.

1.3.3. Energetics of binding

The free energy of binding (ΔG_{bind}) is the *difference* in the free energy of the complex and the free energy of its components, the receptor and the ligand (Equation 1.3). It is a function of the temperature, pressure, ionic strength, *p*H, solvent, and concentrations of all of the chemical species present. Absolute free energies are not available experimentally. Instead, the free energy of a substance under a particular set of conditions is defined as its free energy difference from a reference state.

$$\Delta G^\circ_{bind} = G^\circ_{complex} - (G^\circ_{ligand} + G^\circ_{receptor})$$

Equation 1.3

ΔG°_{bind} is directly related to the experimentally measured K_d (or K_a), as shown in Equation 1.4, where R is the gas constant and T is the absolute temperature in Kelvin. It is possible to calculate ΔG°_{bind} where K_d is defined with respect to the standard state i.e. experimental conditions of 1 atm pressure and 1 M activity of the solutions. The logarithmic relationship between ΔG°_{bind} and K_d means that every 10-fold increase in potency is due to a -1.363 $kcal.mol^{-1}$ change in binding energy [10]. At 25°C, ΔG°_{bind} commonly takes values between -2.4 and -16.7 $kcal.mol^{-1}$, which is equivalent to K_d values of between 10^{-2}

and 10^{-12} M [11]. Various computational methods aim to predict $\Delta G°_{bind}$ in order to provide an estimate of K_d, and therefore ligand binding affinity.

$$\Delta G°_{bind} = -RT \ln K_a = RT \ln K_d$$

Equation 1.4

$\Delta G°_{bind}$ comprises two components: $\Delta H°$, the change in enthalpy, and $\Delta S°$, the change in entropy of the system on complex formation (Equation 1.5). When $\Delta H°$ is positive (unfavourable) and $T\Delta S°$ is positive (so $-T\Delta S°$ is favourable), the interaction is described as *entropically driven*. When the opposite situation is true, the interaction is described as *enthalpically driven*. In the case of most synthetic drugs, which are commonly rather hydrophobic and rigid molecules, binding is an entropically favourable process, often showing unfavourable enthalpy. Changes in enthalpy arise from alterations in van der Waals and Coulombic interactions as the atoms of the complex replace atoms from the solvent during complex formation. Changes in entropy reflect differences in the translational and rotational degrees of freedom for the ligand, receptor and solvent molecules, and the loss of conformational and vibrational entropy for both the receptor and ligand during complex formation.

$$\Delta G°_{bind} = \Delta H° - T\Delta S°$$

Equation 1.5

$\Delta H°$ can be determined experimentally using isothermal titration calorimetry (ITC). ITC uses sensitive microcalorimeters to measure the heat released on association of a ligand with a protein [12]. In a single ITC experiment, values of K_a, $\Delta H°$ and the stoichiometry are determined, allowing $\Delta G°_{bind}$ and $\Delta S°$ to be calculated using Equation 1.4 and Equation 1.5, respectively.

1.3.4. Protein structure determination

As the name implies, Structure-Based Drug Design (SBDD) requires the 3D representation of a protein target, ideally determined by experimental methods. Either in their *apo* (without ligands) or *holo* form (complexed with ligands), protein structures are traditionally resolved using two principal methods: X-ray crystallography and nuclear magnetic resonance (NMR).

The obtention of a crystal of the protein of interest is the fundamental step in X-ray crystallography and it can take from weeks to decades to work out the delicate balance of the appropriate conditions required to crystallise certain proteins. For example, due to their

insolubility in water, it is very difficult to obtain satisfactory crystals of membrane proteins (e.g. receptors and ion channels). In a second stage, an X-ray beam is passed through the crystal and the resulting X-ray diffraction pattern is analysed. Given the highly ordered state of the protein molecules in the crystal, it is possible to determine a 3D electron density map by means of Fourier transforms. A structural model of the protein can then be constructed based on a combination of the electron density map and knowledge of the protein sequence.

NMR captures the magnetic moment of the nuclei of individual atoms to differentiate chemical groups and determine interatomic distances and torsional angles. An intense magnetic field is applied to a purified solution of the protein, resulting in the alignment of certain nuclei spins with the field. Exposure to a brief electromagnetic pulse causes these nuclei to become excited momentarily, and emit/absorb radiofrequency radiation that can be detected and recorded. The excited atoms may also excite their neighbours, a phenomenon referred to as the Nuclear Overhauser Effect (NOE). The NOE is exploited to produce NMR spectra that are used to detect atoms within close proximity of one another. By combining these and other spectra with knowledge of the protein sequence, it is possible to determine a map of the 3D arrangement of atoms within the protein, from which a series of models can be constructed. The main advantage of NMR is that, unlike X-ray crystallography, it does not rely on obtaining a crystal of the protein of interest. However, a significant limitation of NMR is the relatively small size (< 50 kDa) of the proteins that might be analysed.

1.4. Computer-aided drug design: the basics

The concept of Computer-Aided Drug Design (CADD) encompasses the application of informatics methods to the discovery, design and optimisation of biologically active compounds. Indeed, *in silico* methods comprise a wide terrain, including bioinformatics, where drug targets are derived from genomic data, docking studies, where the binding of a ligand or drug to a particular target protein is studied computationally, and chemoinformatics, where activity and structure are correlated using statistical means.

In this subsection, a basic introduction to some of the methodologies most commonly employed in CADD is provided. For further information, I point the reader to references [13,14]. The following subsections approach the methods used in this project in greater detail.

1.4.1. Structure prediction and homology modelling

Frequently, drug discovery projects need to target proteins for which no structure has been experimentally determined by either X-ray crystallography or NMR. In these cases, structure prediction methods may be employed to obtain a suitable structure for SBDD.. The most prominent examples of this are perhaps G protein-coupled receptors (GPCRs), a group of therapeutic targets linked with ca. 50% of the most recently marketed drugs [15]. The most accurate structural prediction method is homology modelling [16], which relies on the fact that all proteins adopt one of a finite number of folds, for which there may already be a representative with a solved structure [17].

By definition, homology modelling requires a template sequence of a known 3D structure that is evolutionarily related to the target sequence. Indeed, since protein structure is conserved to a greater degree than protein sequence, it can be assumed that proteins with significant sequence similarity will also share structural similarity [18].

Sequence comparison is usually carried out by the Basic Local Alignment Search Tool (BLAST) [19], which assigns pairwise sequence similarity based on local alignments, rather than global alignments. In this way, it is possible to identify sequences that have weaker, yet still significant, similarity to the target sequence. BLAST identifies a series of short, non-overlapping sub-sequences ('words') in the target sequences, which are then matched to candidate sequences. Since only the most significant word matches are evaluated, BLAST is able to provide a very rapid result [20]. The extent of sequence similarity between the target and template generally determines the accuracy of the model. A sequence identity of greater than 50% usually allows a high modelling accuracy (backbone RMSD < 2 Å) to be achieved, whereas a sequence identity of 30-50% typically results in a medium modelling accuracy (backbone RMSD ~ 2-4 Å), and below 30% sequence identity, model quality is often poor (backbone RMSD > 4 Å) [21,22].

The next step in homology modelling is to find a solved protein structure that is homologous to a query sequence, based on sequence similarity. This structure then serves as a template onto which a model of the query protein can be constructed. The model, which is guided by the alignment of the two protein sequences, is initially built using only the backbone atoms of the structure as a template. Where there are insertions or deletions in the sequence alignment, the backbone of the model is generated according to the spatial constraints of the adjacent residues. Once the backbone of the model has been constructed, side chains are added, starting with residues that are identical between the two sequences. When this is the case, side chain conformations are adopted from the equivalent residues on the template. Where residues differ between query and template, they are modelled based on a combination of common side chain conformations and local spa-

tial constraints. Once the side chains have been added, unfavourable contacts are removed using an energy minimisation process. Although homology modelling has the advantage of avoiding extensive experimental studies, its main drawback is that it relies on the existence of a suitable template. However, recent advances in structural genomics have led to increased coverage of the fold space, meaning that it is becoming increasingly possible to generate an accurate homology model for any given protein sequence [23].

1.4.2. Molecular Mechanics (MM): the force field

Molecular mechanics (also addressed to as *force field* methods) ignores electrons and their motions and considers the atoms in a molecule as spheres connected by springs. Typically, changes in energy due to bond stretching, angle bending, torsional energies and non-bonded interactions (electrostatic and van der Waals interactions) are computed using parameters such as bond lengths (l), angles (θ), torsions (ω), charges (q) and Lennard-Jones parameters (the collision parameter σ and the well depth ε), which, together with a set of equations, constitute the force field [24]. A very simple molecular mechanics force field is represented Equation 1.6 and the main energy contributions illustrated in **Figure 1.3**.

Molecular Mechanics force fields are derived by studying distinct types of molecules. The most prominent and commonly employed force fields are AMBER [25], CHARMM [26], GROMOS [27] and OPLS [28], which were primarily designed for proteins and lipids. Other force fields, such as GAFF [29] and MMFF [30] are more suited for small molecules. Energy calculations using a MM force field quantify the extent to which bond lengths, angles and torsions deviate from the *ideal* values defined in the force field.

$$V = \sum_{bonds} \frac{k_i}{2}(l_i - l_{i,0})^2 + \sum_{angles} \frac{k_i}{2}(\theta_i - \theta_{i,0})^2 + \sum_{torsions} \frac{V_n}{2}\left[1 + \cos(n\omega - \gamma)\right] +$$

$$\sum_{electrostatic} \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} + \sum_{van\ der\ Waals} 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right]$$

Equation 1.6.

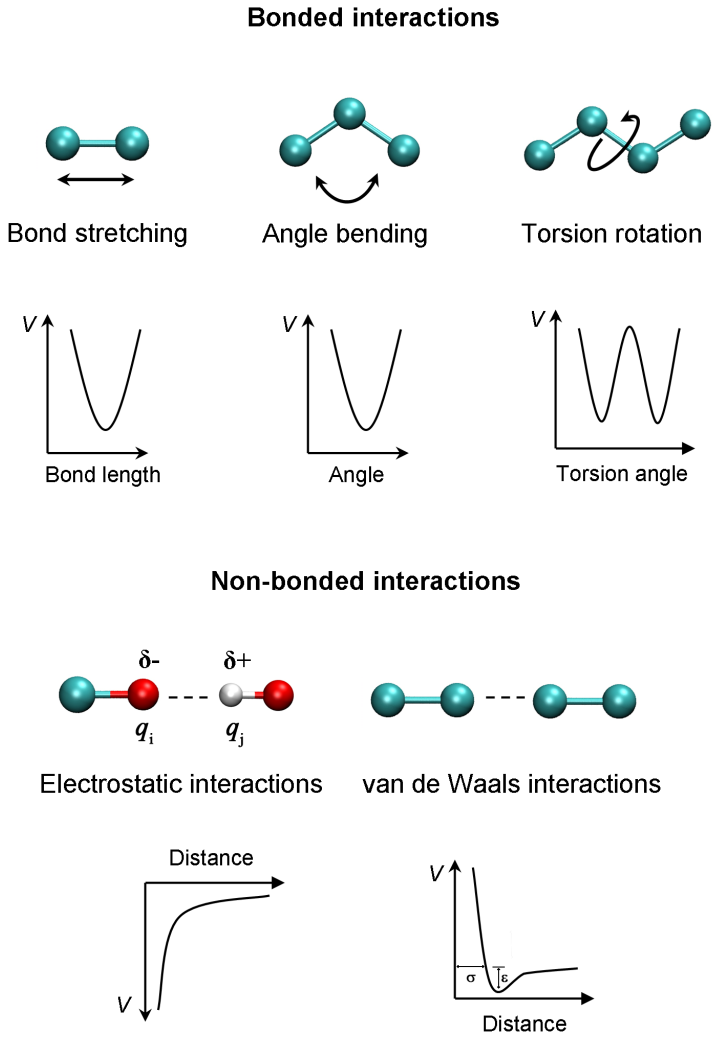


Figure 1.3. Schematic representation of the main energy contributions to a force field. Variations in potential energy are illustrated. Adapted from reference [24].

The iterative process of changing bond lengths, angles and torsions to generate *better* structure conformations is called Energy Minimization (EM) [31]. At each iteration, the energy of the new conformation is measured and the changes that lower the energy are retained for further optimisation until no additional improvement is possible. This point is referred to as the *local* energy minimum. As depicted in **Figure 1.4**, step-wise energy minimization tendentiously lets a system trapped at a local minimum conformation. This energy saddle prevents further progression rather than allowing the system to reach a global minimum (the *global* minimum) and the corresponding lowest-energy conformation. To help crossing energy barriers, methods like simulated annealing, which performs Metropolis walks in the search space by progressively lower temperatures [32], or, more broadly, stochastic conformational searches are used.

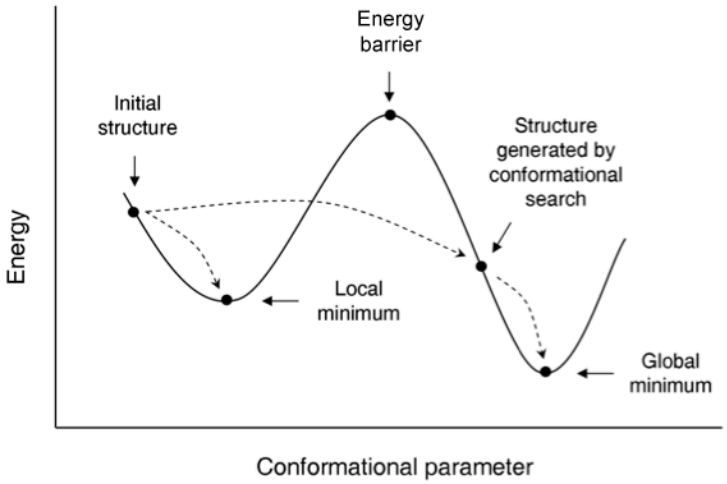


Figure 1.4. Schematic representation of a one-dimensional energy surface. Energy minimization methods move downhill to the nearest minimum. Energy barriers are crossed using conformational search, eventually leading the most stable molecular conformation.

1.4.3. Conformational space sampling

Conformational searches comprise a set of techniques that can be used to generate large *ensembles* of conformers by stepwise bond rotation, molecule rotation and translation, and/or energy minimization. Contrarily to energy minimization, these methods can cross energy barriers and reach conformations that will be close to the global energy minimum. Several computational tools can be used to perform conformational searches.

The Monte Carlo (MC) technique and Genetic Algorithms (GA) are stochastic methods in nature and are commonly applied to problems like molecular docking [33–36]. MC methods randomly generate all possible conformations by rotation about a bond, before minimizing their energies and retaining the structure if it is the lowest energy conformation observed at that point in the simulation. The longer the simulation runs the more effective conformational sampling is. On the other hand, GAs allow a more thorough conformational search. Selective pressure is applied during each iteration in order to encourage high-scoring conformational features to be carried over to the next generation. New conformations are introduced via random *mutations*, while new conformational alternatives are generated by the *crossover* of random pairs of conformations. More detailed explanations about MC methods and GA are provided below, under the subject of Molecular Docking. Finally, and by contrast, Molecular Dynamics (MD) is a deterministic technique that also generates conformations by simulating the time-dependent movement of the molecules (briefly described in a subsection below). As with EM, however,

MD searches tend to get stuck in local minima and their application as a high-throughput solution is very limited.

Molecules do not inhabit in void space. Generally, and especially on living systems, they are found in aqueous environments that will affect their conformation. Thus, even though energy minimization and conformational analysis can be undertaken in vacuum, it is often more adequate to explicitly account for the interactions of the solute with the solvent [37].

More precise calculations of small molecule conformations require the use of quantum mechanics (QM), but these are far more expensive in terms of computing time [38]. QM calculates properties of the molecules using quantum physics to evaluate the interactions between electrons and nuclei. QM methods can be divided into two main methods. The most rigorous comprises *ab initio* techniques that do not require the use of pre-defined parameters. These methods are however dependent on the selection of the basis set, functions to describe the orbitals of the system. On the other hand, semi-empirical methods are faster than *ab initio* methods and applicable to compounds with larger molecular weight. Semi-empirical quantum chemistry methods consider only electrons in the outer shells of the atoms and use predefined parameters.

1.4.4. Molecular Interaction Fields: the grid

As with many biological processes happening in nature, ligand-receptor binding is mostly determined by non-covalent forces. A Molecular Interaction Field (MIF) describes the interaction energy between a target molecule and a certain chemical probe hopped along a 3D grid around the target. Chemical probes come in many sizes and flavours and try to mimic the chemical properties of a binding partner. MIFs can thus provide a detailed picture of the energetic conditions between two or more molecules approaching each other [39]. Computer graphics are able to display MIFs as 3D isoenergy contours. Contours of high positive energy represent regions from which the probe would be repelled, while contours of high negative energy indicate energetically favourable regions for binding.

MIFs analyses are commonly applied in a wide range of molecular modelling studies [40–45]. The chosen approach varies with the type and amount of information available for the macromolecular targets and/or their ligands. When the 3D structure of a protein target is known, the MIFs are used to locate favourable regions for the ligand binding. These regions can then be taken as a starting point for the design of stronger binders for that receptor. Often, however, limited or no structural data of the receptor is available. In such cases, MIFs can help generating reasonably detailed representations of the potential re-

ceptor binding site by taking advantage of available information on active ligands. A prerequisite for this strategy is that all ligands explore the same binding mode within the active site.

1.4.5. Quantitative structure-activity/property relationships

The characterization of Structure-Activity Relationships (SAR) is a critical task in any medicinal chemistry project. Precise and descriptive SAR provide guidance to select further compounds to be synthesized, by suggesting chemical modifications with higher likelihood of yielding better molecules than a random modification. Quantitative structure-activity relationships (QSAR) and quantitative structure-property relationships (QSPR) attempt to statistically relate modifications to the structure of a series of compounds to an experimentally-determined biological activity or physical/chemical property, respectively. The initial step in QSAR or QSPR modelling involves the choice of descriptors to depict the physicochemical properties of the compounds under study. Constitutional and topological descriptors are often combined with more complex descriptors, representing the 3D structure of the molecules. These include several geometric descriptors, and most importantly a variety of electrostatic, QM and MM-derived descriptors. The mathematical models thus built can be utilised to predict the activity/property of similar compounds [46]. QSAR and QSPR models are empirical equations that relate several descriptors of the molecular structure to a single property:

$$\text{Activity / Property} = f\,(\text{descriptor1}, \text{descriptor2}, \text{descriptor3}, \ldots)$$

The concept of three-dimensional quantitative structure-activity relationships (3D-QSAR) is a further expansion of QSAR. The most frequently used 3D-QSAR methods, comparative molecular interaction field analysis (CoMFA) [47] and related approaches [48], assume that the most important features of a molecule, when binding to a protein target, are its steric and electrostatic properties, and its molecular interaction fields. During a 3D-QSAR analysis, a lattice box is built around each compound and a probe atom is placed at each lattice point. The interaction between the ligand and the probe is measured to define properties such as shape, hydrophobicity and hydrogen bond donor-acceptor capacity. The structural alignment of compound series within the lattice box is the most demanding, time-consuming and subjective step in any 3D-QSAR study [49] (see **Figure 1.5**). As a solution to this problem, alignment-independent descriptors have been recently developed [40].

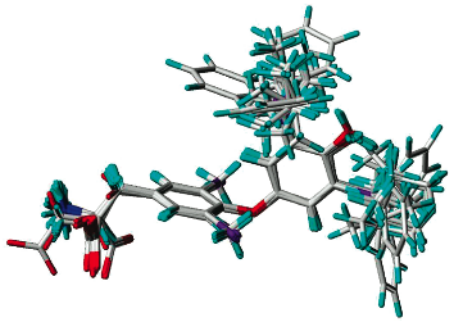


Figure 1.5. A structural superposition of thyroxine and analogs, disclosing a neat alignment of the scaffolds. Adapted from reference [50].

There are well-known strengths but also pitfalls to QSAR and a number of cautions must be taken in order to retrieve the expected outcomes of the method [51]. Ligand-based receptor site models (or 3D-QSARs, or pseudoreceptors) continue to serve as important tools in drug design and new developments have been continuously witnessed, from the early attempts to account multiple (ligand binding) modes [52–55], all the way to the inclusion of multiple species in both receptor- and ligand-based QSAR studies [56–59]. Just recently, Natesan et al. validated a promising new multimode-multispecies method using a ligand-based approach on a set of thyroxine analogs, withdrawing models with good predictive ability of binding affinities for the protein target transthyretin [50] (**Figure 1.6**).

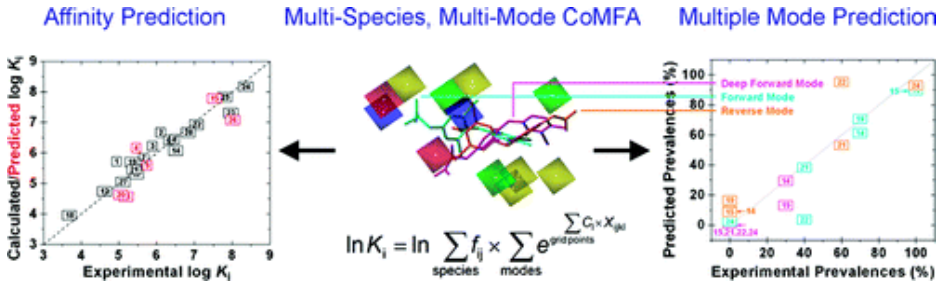


Figure 1.6. Illustration of the 3D-QSAR studies employing a multimode-multispecies method on a set of thyroxine analogs. Recently published by Natesan et al. [50].

1.4.6. Molecular Dynamics simulations

Molecular Dynamics (MD) simulations are the archetypal computational technique for conducting atomic-resolution studies of biological systems [60]. MD is a MM-based theoretical approach whereby successive configurations of a system are produced by integrating the Newton's law of motion ($F = ma$). The resulting trajectory describes how the positions and velocities of particles in the systems vary with time [61]. MD simulations of

biological macromolecules provide atomic detail on the internal motions of these systems. Constant improvements to the methodology and computing power extended the use of MD studies to much larger systems including, for example, very large macromolecular complexes, explicit solvent and/or membrane environment, greater conformational changes and longer time scales [60]. The original MD simulations were less than 10 picoseconds in length, but current simulations can be 100,000 times longer (microsecond timescale) and millisecond-scale MD simulations have already been reported [62,63].

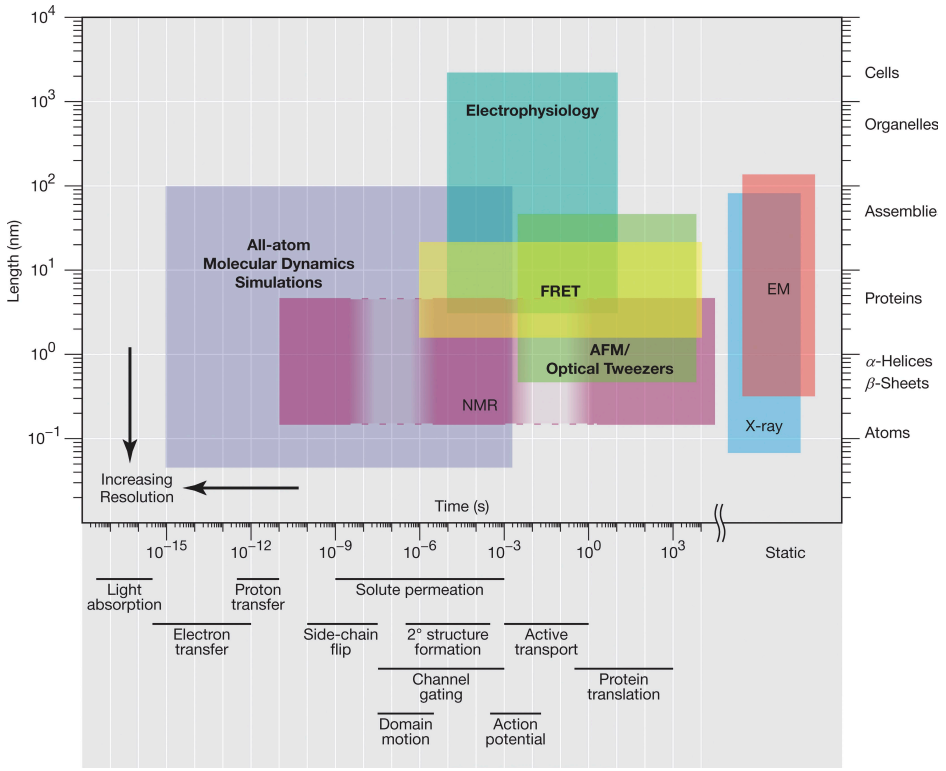


Figure 1.7. Spatiotemporal resolution of MD simulations compared with various biophysical techniques. The temporal (abscissa) and spatial (ordinate) resolutions of each technique are indicated in coloured boxes. Techniques capable of yielding data on single molecules (as opposed to only on ensembles) are shown in bold. NMR methods can probe a wide range of timescales, but they provide limited information on motion at certain intermediate timescales, as indicated by the lighter shading and dashed lines. The timescales of some fundamental molecular processes, as well as composite physiological processes, are indicated below the abscissa. The (spatial) resolution needed to resolve certain objects is shown at right. AFM, atomic force microscopy; EM, electron microscopy; FRET, Forster resonance energy transfer. Extracted from reference [64].

Computational models arising from MD simulations are useful because they can provide critical mechanistic insights that may be difficult or impossible to garner otherwise. Drug design applications of MD simulations include the estimation of target selectivity predictions [65–67], generation of multiple conformations for flexible docking [68,69], devel-

opment of dynamic protein-based pharmacophores [70–72], metabolism prediction [73] and refinement of protein homology models [74]. A great deal of effort, however, has been centred on the use of MD-based methods for the accurate determination of binding free energies and relative binding free energies, largely inspired by some of the first applications of free energy methods to protein–ligand binding [75]. Relative binding free energy calculations (commonly referred to as FEP calculations) constitute the majority of protein–ligand binding calculations conducted in academic and industrial settings, and a variety of techniques have been developed to improve their efficiency and accuracy. This field of work has been thoroughly reviewed and I point the reader to references [76–79]. For a broader review of recent theoretical and computational advances in the modelling of binding affinities, I recommend references [80–84].

1.5. Molecular Docking

Molecular docking is a method for predicting whether and how two or more molecular entities can interact to form an intermolecular complex. Most docking algorithms can generate many possible structures, and so they also need a means to evaluate each structure and discriminate the most valuable ones. Therefore, docking algorithms typically consist of both a search algorithm to explore different conformations of the interacting partners and a scoring function to estimate their binding affinities. The so-called "docking problem" globally refers to the challenge of generating and scoring plausible structure of intermolecular complexes [85].

There are six degrees of translational and rotational freedom of one molecule relative to the other, along with the conformational degrees of freedom within each molecule. Several algorithms have been devised to tackle the docking problem, mostly classified by the degrees of freedom they disregard. The earliest approaches for docking small molecules treated both the ligand and the protein as rigid bodies, exploring only the six degrees of rotational and translational freedom [86]. Yet, more recently, the Monte Carlo method and genetic algorithms have been explored for an efficient sampling of ligand (and sometimes protein) configurational space.

1.5.1. The Monte Carlo algorithm

The heart of the Monte Carlo (MC) algorithm is a random number generator, combined with the ability to compute the energy of the system for a particular set of coordinates. A MC simulation begins with a suitable set of coordinates for all particles of a system. This set of coordinates is perturbed in a random manner and the new configuration of the sys-

tem is accepted as a starting point for the next perturbing step if it is lower in energy than the current. If the new geometry is, however, higher in energy, the Boltzmann factor $e^{-\Delta E/kT}$ is calculated and compared with a random number between 0 and 1. If $e^{-\Delta E/kT}$ is larger than this number, the new conformation is accepted; else, the old configuration is (again) added to the sampling, and a new perturbing step is attempted. The key variation of MC methods is how the perturbing step is accomplished. For a system comprised of spherical particles (such as atoms), the only variables are the centre o mass of each particle, and the trial moves are simple translations of particles. For rigid non-spherical particles, the three rotational degrees of freedom must also be sampled, while for flexible molecules it is also of interest to sample the internal degrees of freedom (conformations, vibrations).

A central aspect of MC methods is to ensure that the chain of configurations results from a *symmetric* probability decision. In this context, symmetry means that each step is reversible, i.e. the probability of undoing a step by the next move is equal to the probability of generating the step – a condition often referred to as the *detailed balance* condition. If this is not fulfilled, the properties derived from the resulting ensemble can (but do not necessarily) display systematic errors, which are usually hard to trace. Generating random moves corresponding only to translation in the positive direction, rather than both positive and negative directions, will almost be sure to lead to artefacts. From a computational perspective, a MC simulation is by far less demanding than an MD simulation. Indeed, if the detailed balance condition is verified, only a single random particle is subjected to a single random perturbation in each step. This procedure is usually more efficient than a trial step involving all particles: when only a single particle is moved, only the change in the energy related to this particle is required, not the whole energy function.

One of the shortcomings of MC methods is their inability to account for correlated motions. Exploring the conformational space of a larger molecule, such as a protein in a solvent, is inefficient, since several concurrent perturbations of torsional angles are required for generating acceptable conformational changes. Such correlated movements are difficult to generate by random perturbations, and almost impossible if only single particle movements are employed in each trial step. On the other hand, one of the advantages of MC methods is the ease with which they can be implemented in computer programs, where they can be used to explore the translational and rotational space of relatively small molecules and their internal degrees of freedom [36,87].

1.5.2. Genetic algorithms

Genetic algorithms (GAs) are utilised as robust search algorithms in several optimisation problems, for they can adapt and improve based on survival of the fittest, mixed with a directed, yet randomised, search procedure. A GA simulates the process of evolution and natural selection by the generation and manipulation of artificial genetic systems [88]. They accomplish this by manipulation the so-called chromosomes, which are represented as strings that can undergo reproduction, crossover and mutation. Just as the total set of chromosomes makes up the genotype of a species, a collection of strings, or in the simplest case just one string, are termed the structure of an artificial system. This structure encodes a set of parameters or points in solution space, similar to the genotype encoding the phenotype of an organism. Chromosomes are a collection of genes, each represented by an allele and location, whereas strings comprise features, each associated with a value and location.

From the docking viewpoint, each chromosome represents a possible solution to the ligand-docking problem. Chromosomes are treated as individuals and as part of population (of fixed size) where each member is evaluated for fitness. Parent chromosomes are then randomly chosen but biased towards fitness and subjected to reproduction operators, producing child chromosomes. Their fitness is evaluated and, if novel, they replace the least fit individual in the population. The whole process, including operator and parent selection, is repeated unless an acceptable solution is found. In the extended version of the algorithm, populations are split into sub-populations and additional genetic operator migration is introduced, allowing individuals to move across sub-populations. Crossover recombines two fittest parent chromosomes whereas mutation changes a value at random. “Survival of the fittest” is achieved over time, moving the population to the best solution for the docking problem. The fitness function plays an essential role in the selection process and determines how accurately it can predict the binding conformation. Starting ligand poses (chromosomes) are generated at random. Each chromosome contains protein-ligand mappings of interaction points (hydrogen bonds, hydrophobic points, conformation around rotatable bonds) and is given a fitness score according to the evaluation of the scoring function.

In spite of the stochastic nature of GAs, they are generally highly reproducible. With certain targets, however, they may require more or longer runs to obtain a match to the crystallographic binding mode.

1.6. Classification of scoring functions

Structure-based drug design methods, such as molecular docking or *de novo* design, are often used with the ultimate goal of predicting ligand-binding affinities, based on the positioning and interactions established within an active site. Besides trying to identify correct ligand poses that match X-ray determined conformations, scoring functions attempt to correctly predict the experimental free energy of the association between a ligand and a protein. In fact, even though many scoring functions report scores using abstract units, some scoring functions were specifically design to estimate the Gibbs free energy of binding. Some studies have even claimed the ability to predict binding affinities within a 1.7-2.4 kcal.mol^{-1} error [89]. Scoring functions can be divided in three main categories: (*i*) force field-based functions (e.g. AutoDock 4 Free Energy Function, GoldScore), (*ii*) empirical free energy or regression-based functions (e.g. ChemScore, X-Score), or (*iii*) knowledge-based potentials (e.g. DrugScore, Potential Mean Force).

1.6.1. Force field-based scoring functions

Force field-based scoring functions (also called Molecular Mechanics- or first principle-) approximate binding affinity by summing individual contributions in a master equation. The terms used for different interaction types are based on the physicochemical theory and should not be cross-correlated with each other. These terms are often combined with solvation and entropic terms. As docking methods are concerned, DOCK 3.0 score [90] is one of the earliest scoring functions that covers the principal contributions to binding: shape and electrostatics are dealt with by means of a van der Waals term and an electrostatic potential term. These separable terms are combined into a grid-based AMBER force-field scoring function that is computed at specific grid points as the sum of ligand atom interactions at the grid points (using a interpolation scheme) assuming additivity of individual terms.

1.6.2. Empirical scoring functions

Empirical scoring functions are developed more specifically for protein-ligand docking by fitting experimental binding affinities using a training set of protein-ligand complexes and are thus dependent on the training set. The free energy of binding is approximated by summing up individual energy terms, which are often simpler but related to molecular mechanics energy terms. Weights or coefficients for each term are derived by regression analysis. Different functions implement various types of energy terms and can include entropic and desolvation terms (although these are still approximations).

ChemScore [91], for example, comprises four simple terms: two contact terms for lipophilic and metal interactions, a hydrogen bonding term and a penalty term depending on the number of rotatable bonds. The weights were derived by regression based on a training set of 82 protein-ligand complexes with known binding affinity and their robustness assessed by cross validation. The design concept involved reduction of the total number of terms and exclusion of those that showed intercorrelation. Additionally, all terms and coefficients should be physics based and interpretable. The scoring was later applied to *de novo* designed compounds that were synthesised and tested [92]. The scoring function was found to be valuable, however, it overestimated binding affinity in several cases and subtle changes between close analogues were not predicted with accuracy.

1.6.3. Knowledge-based scoring functions

Knowledge-based scoring functions are derived by statistical analysis of the frequency distributions within a set of protein-ligand structures from which pairwise atomic interaction potentials are deduced. As such, they reproduce observed preferences of functional group binding, i.e. experimental structures rather than binding affinities. Like with empirical scoring functions, these functions try to overcome the problem of insufficient description of a complex binding event due to the lack of explicit parameters. Well-known examples are DrugScore [93] and Muegge's Potential Mean Force (PMF) [94]. With the increase in available crystal structures (and therefore knowledge), these scoring functions are expected to further improve in the future. The scoring functions differ in respect to their chosen reference distribution, an important term influencing the distance-dependent pair potentials. For example, PMF sets the cut-off at 12 Å for sampling atom pair contacts, while DrugScore at 6 Å. The larger PMF cut-off value was chosen to include implicit solvation effects, whereas DrugScore considers specific interactions.

1.7. Pharmacophore modelling and searches

A pharmacophore is the group of steric and electronic features that are necessary to ensure the optimal supramolecular interactions with a specific biological target and to trigger (or block) its biological response [95]. Using this representation is a useful way to identify new active compounds [96]. The new molecules are most typically studied in 3D so that the pharmacophore model captures both the nature of the functional groups but also their relative orientation to each other.

Because functional groups in a pharmacophore model are not usually considered at an atomic level, but as broader interaction properties, pharmacophore searches are often

successful at scaffold hopping, i.e. at retrieving molecules that are quite dissimilar to the reference ligands. The most typically used features include hydrogen bond acceptors, hydrogen bond donors, positive ionizable groups, negative ionizable groups and hydrophobic regions.

1.7.1. Ligand-based pharmacophore modelling

Ligand-centric approaches can be undertaken in order to identify the most important functional groups and thus develop a pharmacophore model. A set of known active compounds are structurally aligned in the 3D space and the shared functional groups identified, as well as their spatial arrangement [97]. Often, distance tolerances are applied to the distance values, in order to compensate for small displacements that might be allowed within the binding site. Training sets with highly active but conformationally restricted compounds avoid time-consuming and potentially misleading superimpositions. Moreover, in order to reproduce the bioactive conformation, pharmacophore models are usually restricted to low-energy geometries, although not necessarily an energy minimum [98].

Pharmacophore queries can also include a measure of the required shape of the molecules, based on the shape of the compounds used to build the model, in order to improve complementarity to the biological target and avoid steric clashes. Possible adverse steric interactions are prevented with the use of excluded volume spheres, forbidden volumes that compounds cannot map. A complementary approach is the use of inclusion volume spheres selected on the basis of the shape of highly active compounds.

Using queries to search in 3D requires a method to consider conformational flexibility. The simplest approach is to generate and store multiple conformations of all ligands in the database. The software will then test rigidly all conformations of the ligand and return those that fit. Alignment methods require a quantitative measure to assess the degree of overlap between the ligands and the pharmacophore. Typically in point-based methods, the optimisation algorithm attempts to reduce the root mean square deviation of the pharmacophoric features by least-squares fitting. Employing additional 2D substructure filters improves the efficiency of these approaches and reduces the computational time required.

1.7.2. Receptor-based pharmacophore modelling and combined approaches

In the fairly long history of pharmacophore modelling, the utilisation of protein structures is a somewhat recent feature [99]. Nevertheless, several studies have already shown good performances of combined protein structure- and ligand-based approaches in pharmacophore modelling [100–102]. Several programs can nowadays be explored to manually build receptor-based or receptor- and ligand-based models, such as MOE and DS ViewerPro. The program LigandScout, on the other hand, offers automatic pharmacophore perception from 3D complexes [103]. Alternative approaches include structure-based focusing (SBF) [71] and MUSIC [70,72], and can be followed even when only unbound (apo) structures of the protein target are available. These methods combine accessible interaction sites identified using geometric or energetic criteria, and may even account for protein flexibility during the construction of pharmacophore hypotheses.

In Chapter 3, pharmacophore modelling and searches will be covered in detail, as a mixed receptor- and ligand-based approach is exploited in attempt to devise a set of specific and selective pharmacophore hypotheses for a target of pharmaceutical interest.

1.8. Molecular similarity and similarity searches

The terms molecular *diversity* and molecular *similarity* comprise several meanings, depending on the chosen criteria, but, in essence, every small molecule-based approach to identify novel active compounds exploits the concept of molecular similarity to some extent. The so-called "Similar Property Principle" states that molecules that are globally similar should have similar biological activity [104]. In chemoinformatics, the measure of diversity and similarity involves the use of three main components: the descriptors, the coefficients and the weighting scheme. These will be covered in detail in Chapter 4. A myriad of methods that compute diversity of molecules has been developed over the last 50 years and new methods are continuously being presented. Each method defines a different configuration of diversity space, and uses its own descriptors to accomplish its mission. Methods of classification and selection of descriptors are increasing in popularity and are employed to optimise the measures of molecular similarity and diversity.

Maldonado et al. offered an excellent review on this topic, covering several approaches on molecular similarity and diversity, classification and selection methods, and different approaches for comparative analysis [105]. Several types of molecular descriptors, classification and selection methods have been exploited throughout this project. They will be described in great detail in Chapter 4.

1.8.1. Molecular fingerprints

Molecular fingerprints are bit-string representations of molecular structure and/or properties. Simple search strings and atom count vectors represent some of the most basic forms of one-dimensional fingerprints and are often powerful search tools [106,107]. Nevertheless, over the last decades several types of two-dimensional (2D) fingerprints have been developed (see an example in **Figure 1.8**). State-of-the-art 2D fingerprints include hashed connectivity pathways [108], structural dictionary-based [109] and layered atom environment fingerprints [110]. For historical reasons, daylight fingerprints have been used as a standard for benchmarking [108]. However, from a scientific point of view it is difficult to recognize any 2D fingerprint as a standard for similarity searches.

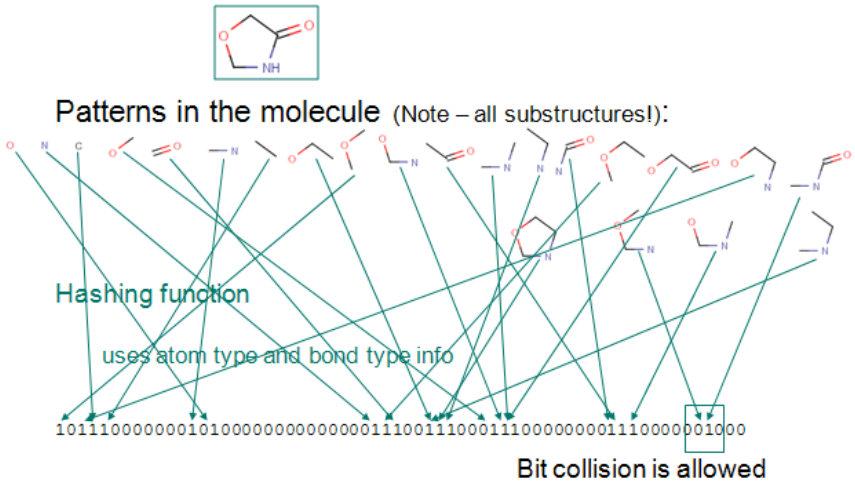


Figure 1.8. Schematic representation illustrating the process of ChemAxon's 2D fingerprint generation. Each unique connected fragment sets a bit 'on'. Fragments can be from 1 to 7 in length, and map out the whole molecule in all directions (adapted from ChemAxon's website).

Two-dimensional fingerprints were originally developed for similarity searching employing single reference compounds. Yet, more recently it was demonstrated that the use of multiple template molecules could enhance search performance [111]. Normally, every template is a known bioactive compound. However, molecules found to be most similar to a single reference compound in a first similarity search can be used, regardless of its activity. This is known as 'turbo' similarity searching [112]. Further studies resulted in attempts to boost fingerprint search performance by following strategies to either scale [113] or average [114] fingerprints, as well as on the assessment of alterative scoring protocols – namely the nearest neighbour methods [114] and data fusion [115]. These techniques individually determine similarity indexes for the database compounds against each reference molecule; the similarity score is either computed as the average similarity

against a pre-specified number of nearest neighbours in the reference set or as the maximum [114]. The latter strategy is usually referred to as 1-NN or 'sum fusion rule' and has yielded the best results in comparative benchmarks [115]. Notwithstanding, nearest neighbour methods, namely 1-NN, seem to show less ability to detect structurally distinct active compounds than methods employing multiple compound information as a whole [116]. Peter Willet as offered an excellent review where data fusion methods are extensively discussed [117].

1.8.2. 3D similarity methods

Similarity search algorithms often employ either 2D or 3D molecular descriptors. However, certain methods rely exclusively on 3D descriptors, including shape-matching algorithms [118], shape-based fingerprints [119], fuzzy, 3D-feature representations derived from cluster analysis of molecular conformations [120], molecular field descriptors [121] and pharmacophore fingerprints [122]. These fingerprints attempt to transform local molecular views into a global view by systematically scrutinizing potential pharmacophore arrangements in molecules. Ensemble pharmacophore techniques render 3D-similarity searching independent of exhaustive guessing of the bioactive conformation of a compound, which persists to be a big bottleneck for important applications of several 3D descriptors and methods.

1.9. Virtual Screening

Experimental high-throughput screening (HTS) is nowadays one of the standard techniques used in the pharmaceutical industry for lead discovery. HTS implies fully automated assays, allowing fast biological evaluation of a large number of compounds for a given target. However, the costs associated with such screenings are high and the hit rates low. This problem might be avoided by not screening complete databases experimentally, but only small subsets enriched in the most promising compounds.

As opposed to HTS, virtual screening (VS, or vHTS) describes any computational approach used to prioritise compounds for experimental screening, predicted to have better probability of being active than compounds selected by chance. Large virtual catalogues may be used, typically with thousands to millions of structures, corresponding to compounds available in-house, commercially available compounds or chemically feasible compounds.

Virtual screening can follow a structure-based approach, a ligand-centric strategy or combination of both structure- and ligand-based methods. Structure-based VS typically uses molecular docking to select compounds predicted to bind with high affinity to the binding site of the target protein. Ligand-based VS uses information from known active compounds to select other molecules. A training set of these structures is the starting point to build pharmacophore queries, as well as similarity or substructure searching procedures. As mentioned above, structure-based pharmacophore models and searches are also being applied with success. Mixed approaches try to make use of knowledge on both the target receptor structure and the known active ligands to define the pharmacophore queries.

Virtual screening can be applied to millions of molecules in a relatively short timeframe. However, the real value of VS is in its complementarity to experimental evaluation techniques. Both ligand- and structure-based approaches are able to identify sets of compounds, which must then be tested in an adequate *in vitro* assay. The main disadvantage of virtual screening is that if tight constraints are used on the type of compounds selected, this will limit the possibility of finding novel and unexpected molecules due to serendipity.

2. Drug discovery in academia

The drug discovery process requires a fundamental step of identification of a few small molecules (lead compounds) that can serve as a basis to further optimisation and the development of a new drug. While in the past this step was mostly accomplished by serendipity, experimental screening methods are today crucial to the generation of new leads. High-throughput screening (HTS) has indeed become the archetypal technique used by the industry to maximize the productivity of the drug discovery process on its early stages. Yet, and despite the success stories of HTS [123–126], the complexity and the quality of the assays [127–129], along with the high-costs involved in their implementation and application, greatly limit its use in small pharmaceutical companies, let alone by academic researchers. Even so, this need not hinder the involvement of academic researchers in drug discovery campaigns for targets of limited commercial value, such as rare diseases and diseases common in developing countries. In a time when the number of target protein structures is sharply increasing versus limited numbers of new therapeutic solutions being introduced in the market, the academic setting is an important niche for the “bubbling” of new ideas in the field of drug discovery.

Considering the great impact in costs reduction of CADD at the pharmaceutical industry, the utilisation of *in silico* methods is becoming a reality also at the academia. In fact, many of the available applications in current use are either open-source software or distributed free-of-charge under a general public license (GPL). Proprietary CADD software is often provided for free or negotiated at reduced costs to the academia. The range of applications currently accessible to academic researchers is immense [130] and their integration in drug discovery programs can foster the involvement of Universities. Moreover, with the ever-increasing computer power and new calculation methods, computational screening is becoming more accessible and is much cheaper than HTS. Academic researchers can design and apply their own algorithms and applications to guide screening campaigns, even if they can only afford the experimental testing of a few tens of compounds.

In recent years, a dramatic increase in the availability of large-scale data for drug discovery has been witnessed [131]. Indeed, the number and dimension of screening and bioactivity data repositories has expanded significantly. The ChEMBL database [132], a tremendous resource that has been recently transferred from the private sector into the public domain, now supplements the existing databases, such as the BindingDB [133], BindingMoad [134] or PDSP Ki [135]. Chemical structure repositories, such as PubChem [136], ZINC [137] and GDB-13 [138], provide access to tens of millions of compounds for applications such as virtual screening [139]. It is also worth mentioning other public domain databases containing useful data for drug discovery. DrugBank [140] and DailyMed [141] provide information regarding approved drugs, ClinicalTrials.gov gathers data on clinical-stage experimental drugs, and DSTTox [142] and TOXNET [143] collect toxicity information from a broad range of public sources. In the field of toxicity, a number of public screening initiatives are also in progress, such as the EPA ToxCast [144–146] and Tox21 [147] projects. Even though these efforts are mainly focused on environmental chemicals, the resulting knowledge may still be of relevance in a drug-discovery context.

Furthermore, a meeting of different commercial and publicly accessible chemical informatics, databases and social networking tools is underway and promises to change the way research collaborations are started, preserved and extended [148]. A community-based platform that combines conventional drug discovery informatics with Web2.0 features in secure groups is believed to be the solution to allow for profitful, immediate collaborations involving sensitive drug discovery data and intellectual property. Assorted chemical and biological data from low-throughput or high-throughput experiments are stored, mined and then selectively shared either just securely between elected colleagues or openly over the Internet [149,150].

The increasing availability of public data, CADD resources and software, and computer power coincides with initiatives in the pharmaceutical industry aimed at reducing costs, for example via increased outsourcing and engaging in precompetitive activities. In principle, one can expect that the coupling of new theoretical approaches developed at the academia with experimental screening techniques exploited by industry will improve the productivity of drug discovery [151]. The establishment of cooperative efforts between the two settings is highly desirable, as it can decrease the high-risk/high-reward ratio of a drug discovery project, which is more critical in the industrial realm for the approval and marketability of new drug candidates.

3. Amyloid as a target for drug discovery

The term "amyloid", which means 'starch-like' and derives from the Greek *amylon* (for starch), was originally coined by Rudolf Virchow in 1854 [152]. The name resulted from a mistaken association of the amyloid substance with carbohydrates, motivated by their common iodine-staining properties. Amyloid actually refers to insoluble proteinaceous aggregates that exert cytotoxic activity and destroy the tissue architecture in certain target organs. Amyloid formation mostly results from abnormalities of the process of protein folding and unfolding, causing proteins to lose their native structure and self-assemble into insoluble, extracellular fibrils. The number of diseases caused by such abnormalities, known as the amyloidoses, is growing. Important examples of such pathologies are Alzheimer's, Parkinson's and Huntington's diseases, type 2 diabetes, familial amyloid polyneuropathy (FAP) and cardiomiopathy (FAC), infectious spongiform encephalopathies, such as Creutzfeldt-Jacob's disease, and the so-called 'aging diseases' like senile systemic amyloidosis (SSA).

The pathogenesis of the amyloidoses requires contribution of multiple factors, of which only a few are well established: the conformational modification of the amyloidogenic protein, post-translational modifications and co-deposition of glycosaminoglicans and serum amyloid P. Yet, in parallel with the exponential growth of biochemical data regarding the key events of the fibrillogenic process, several reports have shown that small molecules, through the interaction with either amyloidogenic proteins or with the common constituents, can modify the kinetics of formation of amyloid fibrils or facilitate amyloid reabsorption. To date, however, only one inhibitor of amyloid formation is known to have reached the drug market: an "only-child" shading light over a perhaps overlooked niche for potential new therapies.

3.1. Amyloid and amyloid diseases

Amyloid disorders are those resulting from the formation and deposition of amyloids. They are usually divided into two categories, depending on the distribution of the amyloid deposits. In localized amyloidosis, amyloid is restricted to a single tissue or organ, usually in the surroundings of the cells responsible for the synthesis of the precursor protein. In systemic amyloidosis, the amyloidogenic proteins are usually derived from circulating precursors that are either in excess, abnormal or both. **Table 1.1** provides a brief summary of the most well-known amyloid diseases.

Table 1.1. Summary of known amyloidosis in humans.

Amyloid protein	Protein precursor	Amyloidosis	Syndrome and/or involved tissues
Aβ	Aβ protein precursor (AβPP)	L	Alzheimer's disease Aging and Familial (prototype Dutch)
APrP^{sc}	Prion protein	L	Spongiform encephalopathies
ACal	(Pro)calcitonin	L	C-cell thyroid hormones
AIAPP	Islet amyloid polypeptide	L	Islets of Langerhans Insulinomas
AANF	Atrial natriuretic factor	L	Cardiac atria amyloidosis
APro	Prolactin	L	Aging pituitary Prolactinomas
AIns	Insulin	L	Iatrogenic diabetes mellitus
ALac	Lactoferrin	L	Cornea amyloidosis
AMed	Lactadherin	L	Senile aortic amyloidosis
AKer	Kearto-epithelin	L	Cornea
AL	Immunoglobulin light chain	S, L	Primary Myeloma-associated
AH	Immunoglobulin heavy chain	S, L	Primary Myeloma-associated
ATTR	Transthyretin	S	Familial, Senile, Systemic
AA	(Apo) serum AA	S	Secondary, reactive
Aβ2M	β2-microglobulin	S	Chronic haemodialysis
AApoAI	Apolipoprotein AI	S	Familial
AApoAII	Apolipoprotein AI	S	Familial
AGel	Gelsolin	S	Familial (prototype Finnish)
ALys	Lysozyme	S	Familial
AFib	Fibrinogen α-chain	S	Familial
ACys	Cystatin C	S	Familial (prototype Iceland)
ABri	BRI gene product	S	Familial dementia, British

L – Localised; S – Systemic. Adapted from reference [153].

3.1.1. Localized and systemic amyloidoses

The most well-known example of localised amyloidosis is that of Alzheimer's disease (AD), affecting more than 5% of the population over the age of 65 years and approximately 25 million people worldwide [154,155]. In AD, the major component of the amyloid deposits is a peptide, the amyloid β-peptide (Aβ). In neuritic plaques of AD patients, this peptide is 42 to 43 amino acid residues-long, while in vascular amyloid it is 39 to 40 residues long [156,157]. The Aβ peptide is derived from the larger precursor protein (APP) as a normal cleavage product [158,159].

Several other common types of localized amyloidosis have been described, including amyloidosis restricted to certain tumours of endocrine glands. In medullar carcinoma of the thyroid, amyloid fibrils are composed of procalcitonin [160,161], whereas in patients with islet cell tumours associated with diabetes mellitus type 2, the amyloid subunit deposited is termed islet amyloid polypeptide or amylin [162,163]. Transmissible spongiform encephalopathies in human (or prion diseases) are characterised by neuronal spongiform degeneration, astrocytic gliosis and parenchymal amyloid deposition in the form of amyloid plaques [155].

Systemic amyloidosis can be further subdivided into non-hereditary and hereditary amyloidosis. In humans, the most frequent forms of non-hereditary systemic amyloidosis are amyloid A associate amyloidosis (AA) and light and heavy immunoglobulin chains associated amyloidosis (AL and AH, respectively). On the other hand, the most common forms of hereditary amyloidosis are related to six different proteins: transthyretin, apolipoprotein AI, apolipoprotein AII, lysozyme, gelsolin, cystatin C and fibrinogen.

3.1.2. Transthyretin-associated amyloidoses

The protein transthyretin is associated with the most prevalent type of hereditary systemic amyloidoses. The pathologic conditions include familial amyloid polyneuropathy (FAP) and familial amyloid cardiomiopathy (FAC). A non-hereditary condition is also related to TTR: senile systemic amyloidosis (SSA), affecting about 25% of people over 80 years of age. In SSA, the deposits occur mainly in the heart and are composed of wild-type TTR. FAP is related to a peculiar form of hereditary autosomal dominant polyneuropathy. Corino de Andrade first described the disease in 1952 [164] in the Portuguese population, mainly from the northern part of the country. The age of onset of the disease is usually between 20 and 35 years of age, characterised by systemic deposition of amyloid and with a special involvement of the peripheral nerves. Progression to death is fast, usually within 10 to 15 years.

Clinically, FAP is characterised by early impairment of temperature and pain sensation in the feet, and autonomic dysfunction leading to paralysis, malabsorption and emaciation. Painless injury to the feet complicated by ulcers, cellulitis, ostemyelitis and Charcots joints may also occur [165]. Motor involvement occurs with disease development, causing wasting and weakness, accompanied by a progressive loss of reflexes. Upper limbs involvement may occur years after lower limbs manifestations, progressing in a similar way. The amyloid deposits can occur extracellularly in any part of the peripheral nervous system, including the nerve trunks, plexus and sensory and autonomic ganglia. In the peripheral nerves, they affect the epineurium, perineurium and especially the endoneu-

rium. In the endoneurium, deposits are usually closely opposed to Schwann cells or collagen fibrils.

The heart is another organ frequently involved in FAP. Clinically, the cardiomiopathy may express as an arrhythmia, heart block or hear failure. Electrocardiographic abnormalities with Q-wave and T-wave repolarisation changes and various conduction disturbances are some of the clinical features. Echocardiography shows a restrictive cardiomiopathy with thickened interventricular septum and ventricular walls [166], although changes consistent with a hypertrophic cardiomiopathy have also been described [167]. In FAP, the amyloid deposits are widely distributed and other clinical features additional to the main ones already described can occur: nephropathy and more rarely pulmonary [168] and bone involvement [169].

3.1.3. Amyloid formation pathway: the hypotheses

While the involvement of the amyloid substance with several impairing diseases is more or less consensual, different theories have been proposed to explain the formation of amyloid. In fact, the three main hypotheses are, in many respects, complementary in nature:

- **The conformational hypothesis**

  Soluble globular proteins undergo destabilisation and conformational changes into partially unfolded states before self-assembling into amyloid fibrils. In particular, many mutations seem to favour protein destabilization. Unlike globular proteins, however, amyloidogenic peptides like the Aβ peptide are mostly composed of random-coiled structure in their native states. Still, these peptides can adopt partially structured conformations that are stabilised by oligomerization. Walsh et al. witnessed a transient increase in α-helix during the conversion of an extensively unfolded Aβ to amyloid fibrils enriched in β-sheet [170].

- **The proteolytic hypothesis**

  Proteolysis has been deemed as a critical process in most forms of amyloidosis. Indeed, amyloidogenic proteins are often released from their precursors via proteolysis. Aβ, for instance, results from cleavage of the amyloid precursor protein (APP) by a combination of β- and γ-secretases [171,172]. Moreover, TTR peptides, besides the intact protein, have been detected in amyloid fibrils of SSA and FAP patients, leading to the hypothesis that proteolysis could underlie the formation of fibrils by releasing amyloidogenic fragments [173]. Both experimental and computational works have

been conducted to assess the amyloidogenic propensities of different TTR fragments [174–176]. However, earlier studies showed that no fragments are found in samples of Val30Met TTR fibrils, and the collected evidence seems to indicate that the release of peptides is not a requirement for amyloid formation [177,178]. Therefore, the role of proteolysis in TTR amyloid is still uncertain.

- **Nucleation and seeding**

  A number of studies have shown that amyloid formation is often a nucleation-dependent process, whereby ordered nuclei are formed after a lag phase [179]. Fast growth of the fibril takes place after nucleation and is followed by an accumulation time course (sigmoidal in shape) [180]. According to this hypothesis, fibril growth requires the formation of an oligomeric *nucleus* that is the highest energy species along the pathway, and the rate of fibril formation is increased by addition of preformed aggregates (or *seeds*), which effectively circumvent the nucleation step. TTR amyloid has long been believed to follow a nucleation-dependent polymerization. However, this view is not entirely consensual. Hurshman et al. demonstrated that, under denaturating conditions, the aggregation of TTR monomers is not accelerated by seeding and suggested a downhill polymerisation process [181].

3.1.4. Amyloid fibrils: an unique superstructure

The 3D structure of amyloid fibbers has been extensively investigated using various biophysical methods, such as X-ray diffraction [182–184], solid-state NMR [185–187], electron microscopy [188], Fourier-transformed infrared spectroscopy [189] and atomic force microscopy [190,191]. Regardless of their protein precursor, these polymers share in common a remarkable number of features: they are built of uniform, non-branched fibers with a diameter of ca. 100 Å and variable lengths. The molecular structure is marked by an extended β conformation of polypeptide chains stacked through hydrogen bonds and arranged in sheets that run parallel to the main axis of the fiber (see **Figure 1.9**). The β-strands are disposed perpendicularly to the main axis and are separated by distances of 4.7 Å; the distance between the sheets varies from approximately 10 to 15 Å.

Before growing into mature extended fibrils, the amyloidogenic proteins self assemble into small β-rich oligomers, presenting a globular shape and the tinctorial properties of amyloid fibrils. Typically, these oligomers are soluble in physiological buffers and play a key role in the cytotoxic mechanisms of disease-associated amyloidogenic proteins [192,193].

The highly ordered supersecondary structure of protofibrils and fibrils is specifically recognized by dyes like thioflavin and Congo red, which are widely utilised in the diagnosis of amyloid disease and important in the study of ligand binding to amyloid fibrils [194–196].

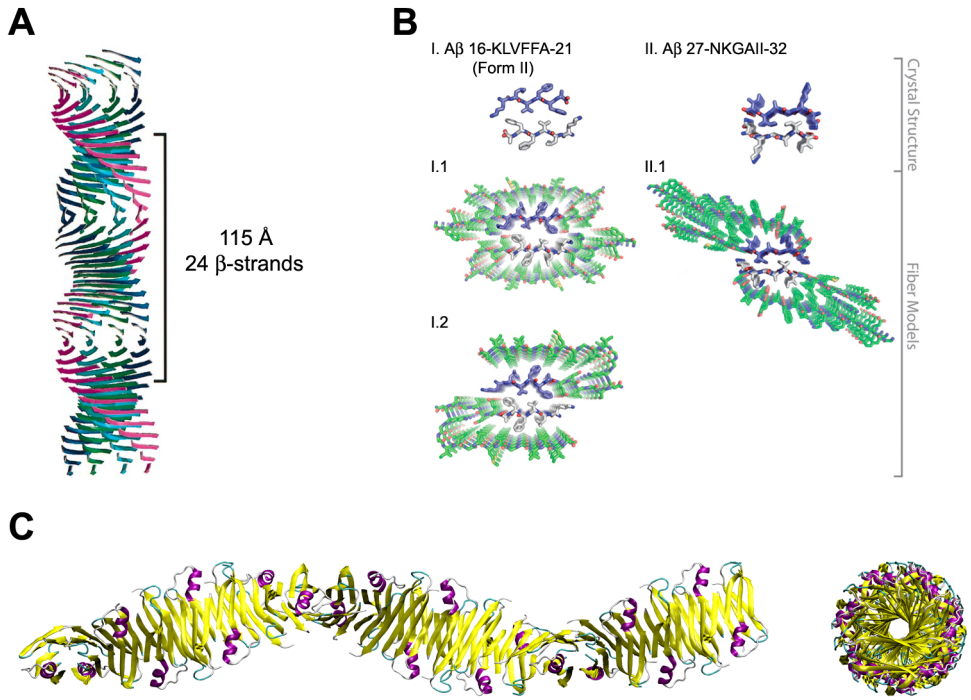


Figure 1.9. An illustration of different studies conducted to shade light on the possible structure of amyloid protofibrils. (A) Molecular model of the generic amyloid fibril derived by synchroton X-ray diffraction, showing the common core protofilament structure [197]. (B) Models of Aβ protofilaments associations via different steric-zipper interfaces. The interface between two β-strands within the crystal structures of (*I.*) A$\beta_{16\text{-}21}$ Form II or (*II.*) A$\beta_{27\text{-}32}$ was used to model interactions between two protofilaments [198]. (C) A theoretical model of a transthyretin protofilament obtained by molecular docking and driven by experimental data [74].

3.2. Transthyretin (TTR): a model target in amyloid

Transthyretin (TTR) was first discovered in 1942 during an electrophoresis experiment with cerebrospinal fluid (CSF) [199] and shortly after detected in serum [200]. It was first called component X and later named "prealbumin", a term based on its characteristic migratory electrophoresis pattern in front of albumin. TTR is an evolutionarily conserved homotetrameric protein synthesized in the liver and the choroid plexus of the brain, which can be found in the blood plasma and cerebral spinal fluid. In 1958, TTR was shown to be a carrier of the thyroid hormone thyroxine (T_4) [201] and, ten years later, to bind retinol-binding protein (RBP) [202]. In 1982, it became clear that TTR was not re-

lated to albumin, and the current name was adopted by the International Union of Biochemists to avoid misunderstandings [203].

TTR is also implicated in the deposition of fibrils in peripheral nerves and heart tissue in FAP and FAC, respectively. This process involves an initial step whereby the native TTR tetramer dissociates to monomers with altered tertiary structure, followed by the self-assembly of these monomers to form cytotoxic oligomers [204,205]. The initial step was shown to be rate limiting for amyloid formation and can be modulated by the binding of thyroxine-like compounds to the two identical thyroxine-binding sites of tetrameric TTR [181,206–208]. Even today, TTR stands as one of the very few targets available to the therapeutic targeting of amyloid fibril formation. Our interest in human TTR as a model target is linked not only to its involvement in disease [74,204,209], but also to the many peculiarities that render it a particularly challenging target [210].

In the late 1990s, it was shown that ligand binding to TTR results in stabilization of the tetrameric fold and inhibition of amyloid formation *in vitro* [211–213]. As will be described in a subsection below, mostly over the last decade, tens of analogues and derivatives have been discovered and designed, showing potential as drug candidates [206,214–222]. Although several stabiliser molecules discovered to date have shown efficacy at inhibiting fibrilization *in vitro*, the optimal balance between affinity and selectivity for TTR in the blood plasma is still a goal to achieve. Furthermore, the typical TTR stabiliser (comprised of two aromatic rings often substituted with halogen atoms) is associated with undesirable properties, such as poor solubility and low polar surface areas, which are in turn linked with important adverse-effects and accumulation in the body.

3.2.1. Gene Structure and Expression

Human TTR is encoded by a single-copy gene, which has been assigned to chromosome region 18q11.2-q12.1 by Sparkes et al. [223]. The gene spans approximately 7.0 kilobase pairs with four exons (each with approximately 200 bases), three introns, a TATA box-like sequence at nt –24–30 and a CAAT box-like sequence at –95–102 [224,225]. Given the association of FAP with a mutation in TTR, it can be assumed that the gene for this pathology is also mapped to 18q11.2-q12.1. The genomic structures of TTR have also been reported in the rat [226] and the mouse [227]. Investigations on transgenic mice identified two regions that are sufficient for directing TTR gene expression in the liver [228], but little is known about the regulatory mechanisms for the abundant expression in the choroid plexus of several species.

Complementary DNA (cDNA) coding for TTR of more than 10 different species has currently been cloned [229]. It generally consists of a 5' untranslated region (14-30 nucleotides), a coding region of 127-130 amino acids (varying across different species), and a 3' untranslated region (115-181 nucleotides) preceding the poly(A) tail. Changes introduced in the splice sites of intron 1 resulted in the expression of TTR with a shorter and more hydrophilic N-terminal amino acid sequence [230]. While the direct impact of these changes on TTR function is elusive, binding studies performed with plasma from birds and mammals revealed that TTR preferentially binds T_3 in birds and T_4 in mammals. Oddly enough, the amino acids that are involved in the interaction of TTR with thyroid hormones are entirely conserved across vertebrate TTRs [231]. Chang et al. proposed that the change in the N-terminal could be responsible for the change in hormone affinity [232]. The shorter, more hydrophilic N-termini correlates with preferential binding to T_4, whereas the longer, more hydrophobic N-termini with preferential binding to T_3 [232,233]. More recently, this hypothesis was quantitatively confirmed through further binding affinity studies, and the influence of the N-terminal region on the binding of T_3 and T_4 to TTR became more evident [234].

Figure 1.10 shows a multiple alignment of the amino acid sequence from representative species, revealing the very high degree of conservation that has occurred in TTR between eutherians, marsupials, birds and lizards (65-85%). The sequence conservation between TTR from the latter species and amphibian and fish is significantly lower (47-48% identity). Nonetheless, comparison of the protein sequences considering conservative amino acid substitutions discloses higher analogy between sequences, implying great conservation of TTR properties at the chemical level.

Even though TTR is encoded by a single gene yielding a single protein product, great sequence heterogeneity is found as result of mutations in single residues at various positions of the wild-type sequence. The mutations may be non-pathogenic or cause severe pathologies such as FAP. Despite the even distribution of pathogenic mutations along the protein sequence, multiple mutation *hotspots* focused on localised regions of the protein structure have been suggested [229]. Strikingly, some of the mutations that cause disease in humans are a normal feature of the protein in other species. For example, the genetic mutations that lead to the Val30Leu and Ile84Ser TTR variants in humans are normally present in sea bream (see **Figure 1.10**). Comparison of human TTR variants linked to disease with TTR from other species reveals that 15 of the 36 substituted amino acids in human TTR have been 100% conserved in TTR of all other species, which may imply that these particular residues are of structural or functional relevance and there has been strong evolutionary pressure to conserve them [235]. The remaining 21 amino acids exhibit low conservation across species.

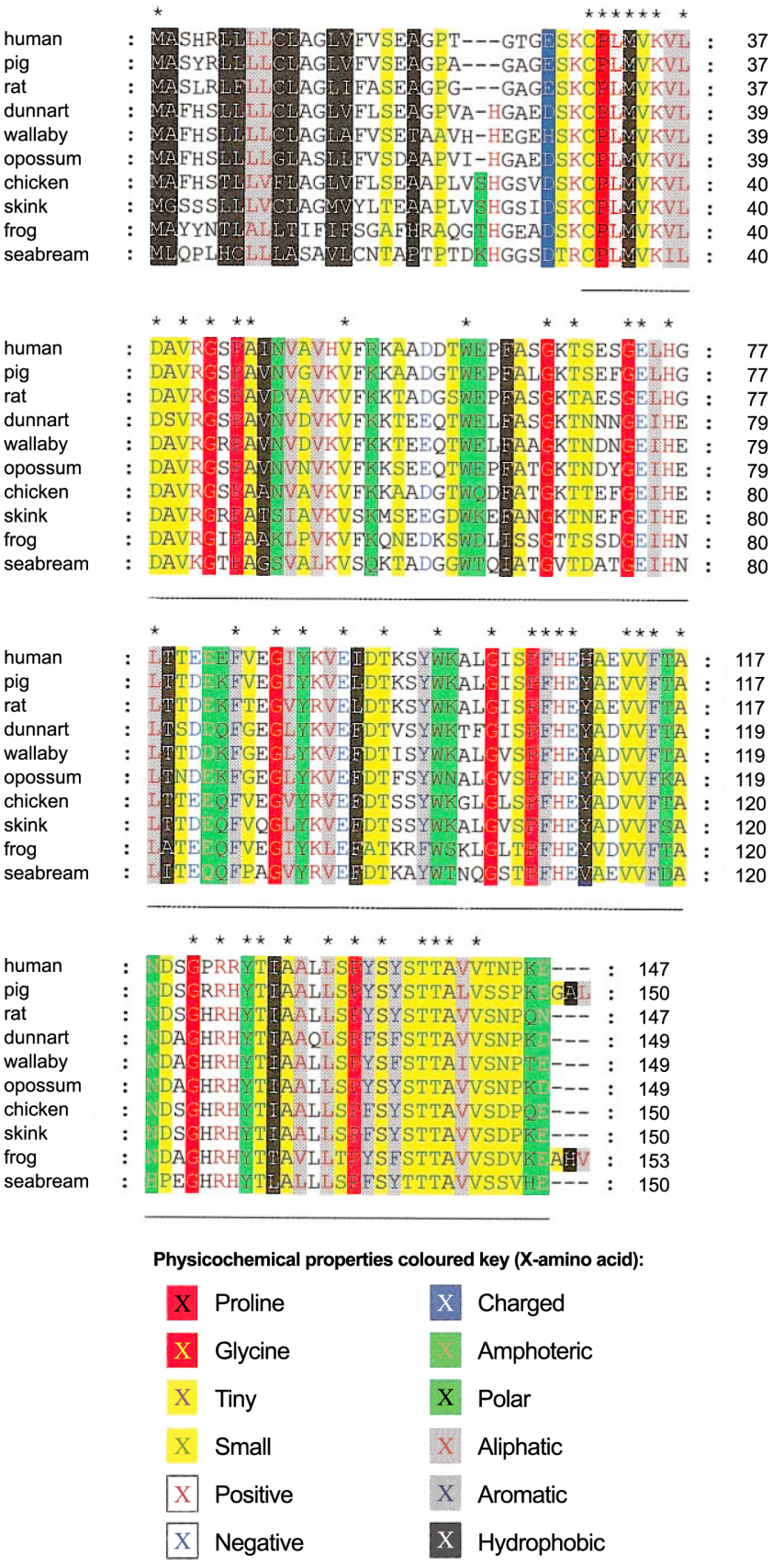


Figure 1.10. Multiple sequence alignment carried out using Clustal X [236] **of transthyretin with the prepeptide from representative species of all vertebrates.** Ad-

apted from reference [235]. Human [224], pig (*Sus scrofa*) [237], rat (*Rattus norvegicus*) [238], dunnart (*Sminthopsis macroura*) [224], wallaby (*Macropus eugenii*) [239], opossum (*Monodelphis domestica*) [224], chicken (*Gallus gallus*) [240], skink (*Tiliqua rugosa*) [241], frog (*Rana catesbeiana*) [242], and sea bream (*Sparus aurata*) [243]. Coloured blocks refer to the physiochemical properties of the amino acids according to the key. The underlined residues correspond to the region modelled by homology (and using rotamer-based methods) in reference [235]. Residues marked with an asterisk are identical in all species analysed.

Gene expression of TTR is significantly restricted to the liver and the choroid plexus. TTR synthesised in the liver is secreted into the blood, whereas that from the choroid plexus epithelium is secreted into the brain cerebrospinal fluid, where it is the main carrier of thyroid hormones [244]. TTR gene expression has also been found in the eye of cattle and sheep [245–247]. Small amounts of mRNA are expressed in the pancreas of rat and humans [248]. In vitro cultures of pigment epithelium of the rat retina show production and secretion of TTR into the interphotoreceptor space of the retina [249]. Presence of TTR in the visceral yolk sac during fetal rat development [250], in the developing rat eye [251], and in the developing chicken heart [252] have also been reported. There have also been cases where no expression was detected in the choroid plexus and TTR expression occurs exclusively in the liver. These include premetamorphic tadpoles of the frog [253] and juvenile sea bream [243]. The significant expression of TTR in the intestine and the heart of adult sea bream raises questions about the physiological function of the protein at these sites. Without any doubt, however, the diverse location of TTR gene expression provides evidence of evolutionary deviations and adaptations in TTR function.

3.2.2. Molecular Structure

The molecular structure of transthyretin was first described by Blake et al. [254] and further refined at 1.8 Å resolution [255]. TTR is a multimeric protein composed of four identical 127-residues monomers and functioning as a 55-kDa homotetrameric unit. Each subunit is characterised by an extensive β-structure composed of eight strands, arranged into two four-stranded sheets (D-A-G-H and C-B-E-F), and one short α–helix between strands E and G (**Figure 1.11**). The assembly of the two β-sheets largely determines TTR's tertiary structure. Critical contributions to the tertiary structure are made by:

- a cluster of seven aromatic residues;
- three tyrosines and one aspartic acid involved in side-chain-main-chain interactions;
- one histidine (buried in TTR's dimer) integrating an internal network of water molecules.

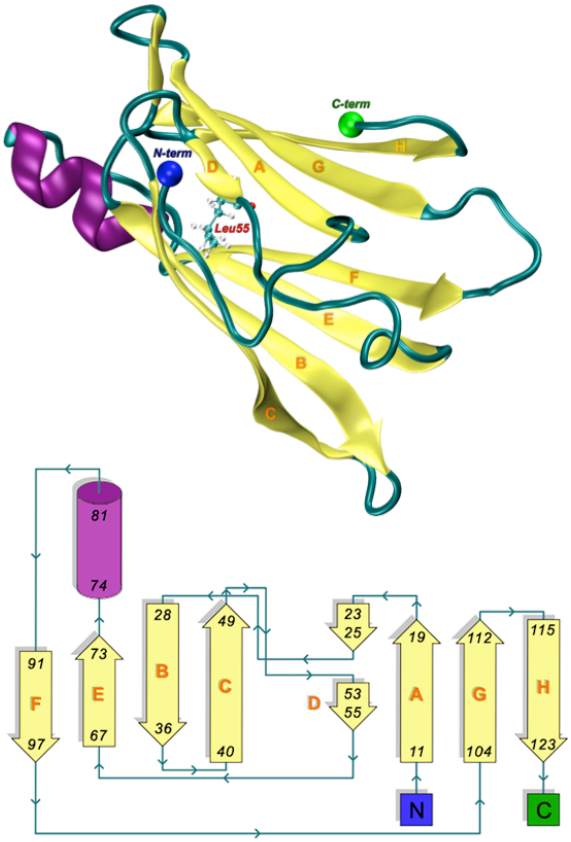


Figure 1.11. Molecular structure of one TTR subunit. The coordinates were retrieved from the Protein Data Bank (PDB entry 1tta), the 3D ribbon representation was produced using VMD and the 2D topology diagram was generated using HERA/PDBsum [256,257].

Quaternary interactions take place at two sets of interfaces proximally arranged around two of the three molecular 2-fold axes in the tetramer. The exclusive monomer-monomer interface encompassed by the F, F', H and H' strands is mainly orchestrated by hydrogen bond interactions in an antiparallel arrangement of the β-strands that extend the two four-stranded sheets in the monomers to eight-stranded sheets in the dimer. The two dimers then assemble by hydrophilic and hydrophobic interactions between the AB and the GH loops at interfaces that entail all four monomers (see **Figure 1.12**A).

The native tetramer of TTR features two funnel-shaped binding sites deeply buried in a narrow cylindrical channel situated at the interface of two dimers. The channel has a diameter of approximately 8 Å, and the proximity of paired side-chains of Leu-110, Ser-115 and Ser-117 causes a strong constriction near the centre. The structurally-equivalent binding sites display negative cooperativity for the binding of thyroxine [258] and hold characteristics typically associated to protein-protein interfaces [259]: featureless and flat [260], and mostly comprised of apolar residues. The residues involved in T_4 binding seem to have been conserved between the human and the chicken TTR [240]. The chemistry of the binding sites is characterised by three linearly-arranged elements:

- a hydrophilic patch formed by the hydroxyl side-chains of Ser-117 and Thr-119;
- a hydrophobic patch formed by the methyl groups of the Leu-17, Thr-106, Ala-108, Leu-110, Thr-119, and Val-121 pairs;
- a group of charged residues including the paired side-chains of Lys-15 and Glu-54.

The characteristics of TTR binding sites and the nature of cooperative effects in ligand binding will be further discussed in a subsection below and in Chapters 2 and 3 of this thesis.

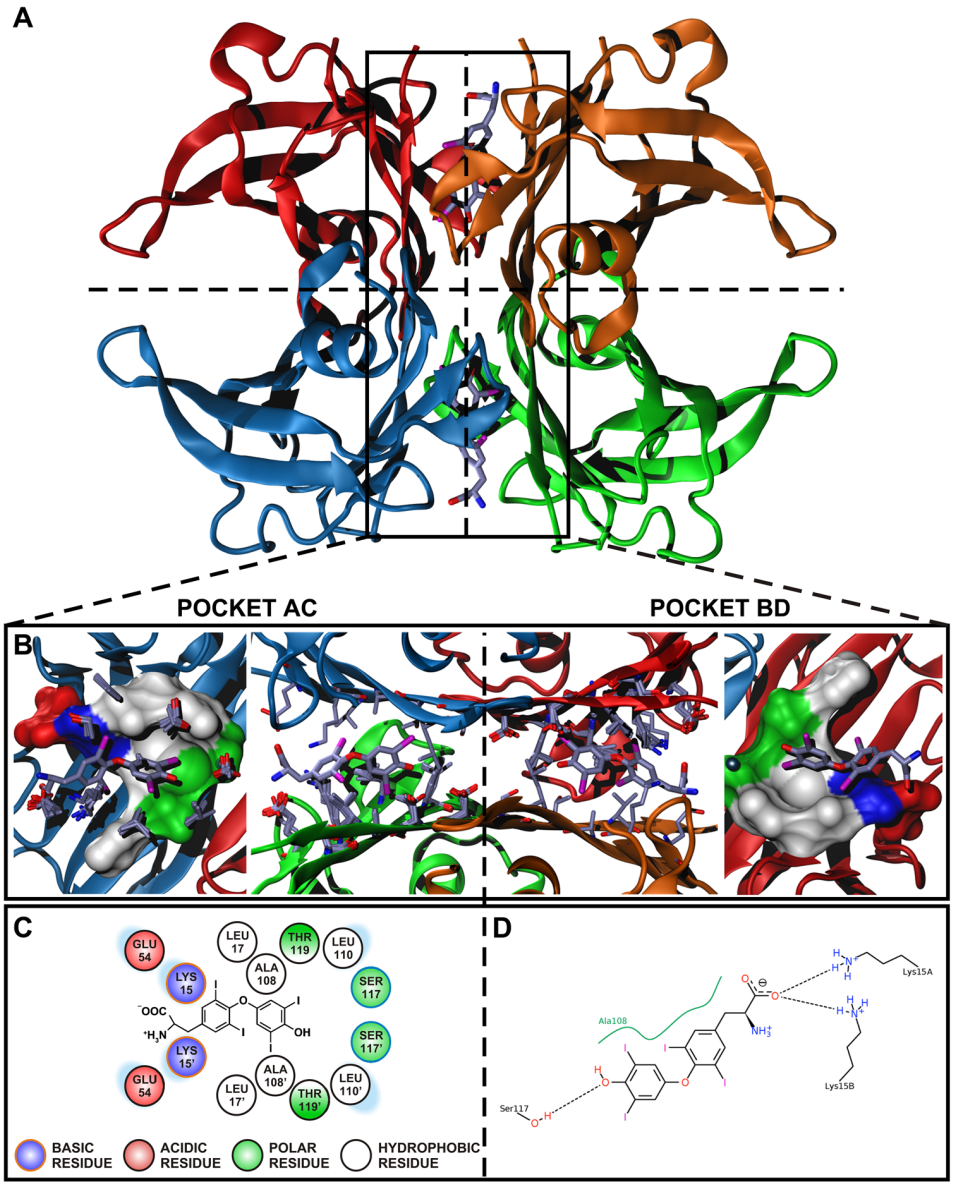


Figure 1.12. Molecular structure of human transthyretin (TTR) bound to its endogenous ligand thyroxine (T_4). (A) A 3D ribbon representation of the X-ray structure (PDB code 2rox), with T_4 positioned within the two symmetrical binding sites created at the dimer-dimer interface (vertical dashed line). (B) Expanded views of the binding sites from

two different perspectives. In the central panels, the result of a multiple structural alignment of 30 TTR complexes is displayed as licorice drawings for the side chains of one monomer. At the edge panels, the molecular surface of binding site residues is shown. (C) A 2D topology diagram of the binding sites' features. (D) An alternative 2D diagram of the interactions between TTR and T_4 generated using PoseView.

Under physiological conditions, TTR can also be found in complex with the retinol-binding protein (RBP), a 21 kDa protein composed of 182 residues that transports all-*trans*-retinol [261,262]. In theory, each TTR tetramer presents four anchor points for the binding of RBP molecules, one per monomer (see **Figure 1.13**). However, like with thyroxine [258], the association between TTR and RBP displays negative cooperativity, i.e. the binding of a first RBP molecule to TTR alters the affinity of binding of a second molecule and so forth. In practice, crystallographic studies have established the stoichiometry of the TTR-RBP complex as two RBP molecules per TTR tetramer [263–265]. The concentration of RBP in the plasma is limiting, and the complex isolated from serum is composed of TTR and RBP in a 1 to 1 stoichiometry.

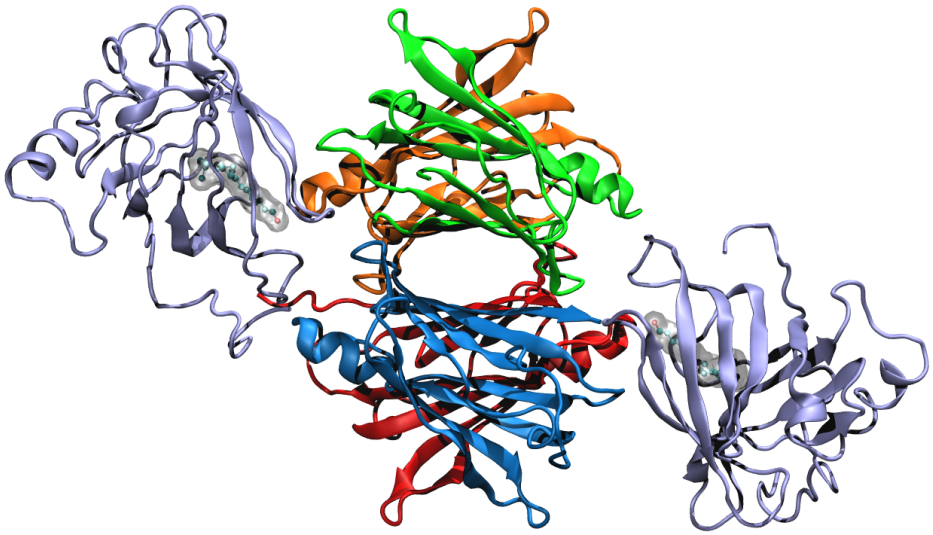


Figure 1.13. Molecular structure of TTR bound to two retinol-binding protein (RBP) molecules in a 3D ribbon representation. Retinol is shown in a Corey-Pauling-Koltun (CPK) representation; the presence of retinol bound to RBP is essential for the formation of a stable TTR-RBP complex. The coordinates were retrieved from the Protein Data Bank (PDB entry 1qab) and the image was generated with VMD.

RBP binds at a 2-fold axis of symmetry in the TTR tetramer, and therefore the recognition site itself has 2-fold symmetry: four TTR amino acids (Arg-21, Val-20, Leu-82, and Ile-84) are contributed by two monomers. Residues Trp-67, Phe-96, and Leu-63 and Leu-97 of RBP are flanked by the symmetry-related side chains of TTR. The structure also discloses

an interaction of the carboxy terminus of RBP at the protein-protein recognition interface.

3.2.3. Function

The name TransThyRetin describes the most prominent physiological "roles" of TTR, i.e., the *trans*port of *thy*roxine [266] and *retin*ol-binding protein [267]. The implication of such name was that the physiological roles of TTR were sufficiently well understood to replace the prior descriptive term of "thyroxine binding prealbumin" by the more functional nomenclature. Currently, there seems to be no room for doubt that the assumption that the biology of TTR was fully known was slightly precipitated. As pointed out by Buxbaum and Reixach, "semantic clarity leading to undeserved observational certainty is a constant risk when trying to rationalize scientific nomenclature" [210].

TTR is suspected to result from a duplication event in a gene of the vertebrate lineage encoding 5-hydroxyisourate hydrolase (HIUase), a member of a wide family of enzymes involved in purine metabolism [268]. Given their high sequence similarity to TTR, HIUases were initially believed to have transport functions and became known as TTR-like proteins (TLPs) or TTR-like proteins (TRPs) [269,270]. Only more recently it was recognized that both prokaryotic and eukaryotic members of this family are involved in the degradation of uric acid to (S)-allantoin [270]. Both the sequence and the structure of TTR and TRPs have been highly conserved throughout evolution, suggesting important biological roles. However, TTR knockout mice have normal fetal development, are phenotypically normal, viable and fertile, even though they have no detectable retinol in plasma and a decreased level of thyroid hormone [271]. These contradictory notions corroborate the fact that the function of TTR is still poorly understood.

Until recently, TTR had established clinical significance in two settings. TTR serum concentration has been used as a marker of nutritional and inflammatory status in several conditions [272,273]. Under normal circumstances, TTR is present in human plasma at a concentration of 0.25 $mg.mL^{-1}$ [274,275] and with a fast turnover rate of 2 days half-life [276]. More relevantly, as an etiologic agent, TTR is amongst ca. 30 proteins linked to the important amyloid diseases. Besides its activities as a carrier protein, TTR has also been shown to hold protease activity [277]. The "puzzling" nature of this proteolytic role stems not only from its apparent disconnection to TTR transport roles but also from the lack of structural protease determinants in the protein [278].

Since TTR is also synthesized in the brain (by the choroid plexus), it is plausible to deem a function in the biology of the nervous system. Indeed, using TTR gene knockout mice it

was possible to establish implications in sensorimotor performance [279], depressive-like behaviour [280], neuropeptide maturation [281], nerve regeneration [279] and spatial reference memory [282]. Interestingly, TTR has also been shown to be able to sequester amyloid β protein [283,284] and to play a protective role in transgenic murine models of Alzheimer's disease (AD) [285,286]. In fact, the lack of TTR seems to yield a discrete behavioural phenotype, even in the absence of the pathologic Aβ gene [280]. However, evidence of these observations in humans is still missing. While the exact mechanisms by which TTR operates in the brain remain elusive, it is clear that the function of this protein extends beyond its systemic role as a carrier of T_4 and retinol.

TTR's ability to bind several families of compounds has also been hypothesized as a way of acting as an endogenous detoxifier of molecules with potential pathologic effects. This has also been proposed for albumin, the most abundant protein in the plasma (35-50 $mg.mL^{-1}$), which presents four hydrophobic binding regions that are able to bind ions, fatty acids, drugs, hormones, and prevent the formation of free radicals [287]. The many similarities between serum albumin and TTR – such as their molecular weight, low isoelectric point or ability to bind hydrophobic compounds – paired with their complementary secondary structures – albumin being mostly α-helical and TTR mostly β-sheet – makes it reasonable to advocate that these proteins may share responsibilities like the removal of potentially toxic dietary substances or by-products of metabolism [210]. Recently, Liz et al. provided an exceptional review focused on the association of TTR with various types of ligands and substrates, which also attempts to establish links to putative biological roles [288].

3.2.4. Variants and involvement in disease

Several studies have shown that amyloidogenic mutations destabilise the native structure of TTR and thus accelerate a cascade of events that involves tetramer dissociation followed by partial unfolding of the released monomers into an aggregation-prone conformation by which oligomers, soluble aggregates, insoluble amorphous aggregates, and amyloid fibrils are formed in a downhill polymerization process. Holding highly conserved sequence and structure across human and non-human species, TTR alone is associated with more than 90 amyloidogenic mutations, which are all implicated in the deposition of fibrils [205]. Of these, V30M, L58H, T60A, I84S and V122I mutations are responsible for the majority of familial amyloidotic polyneuropathy (FAP) and familial amyloidotic cardiomiopathy (FAC) cases [289].

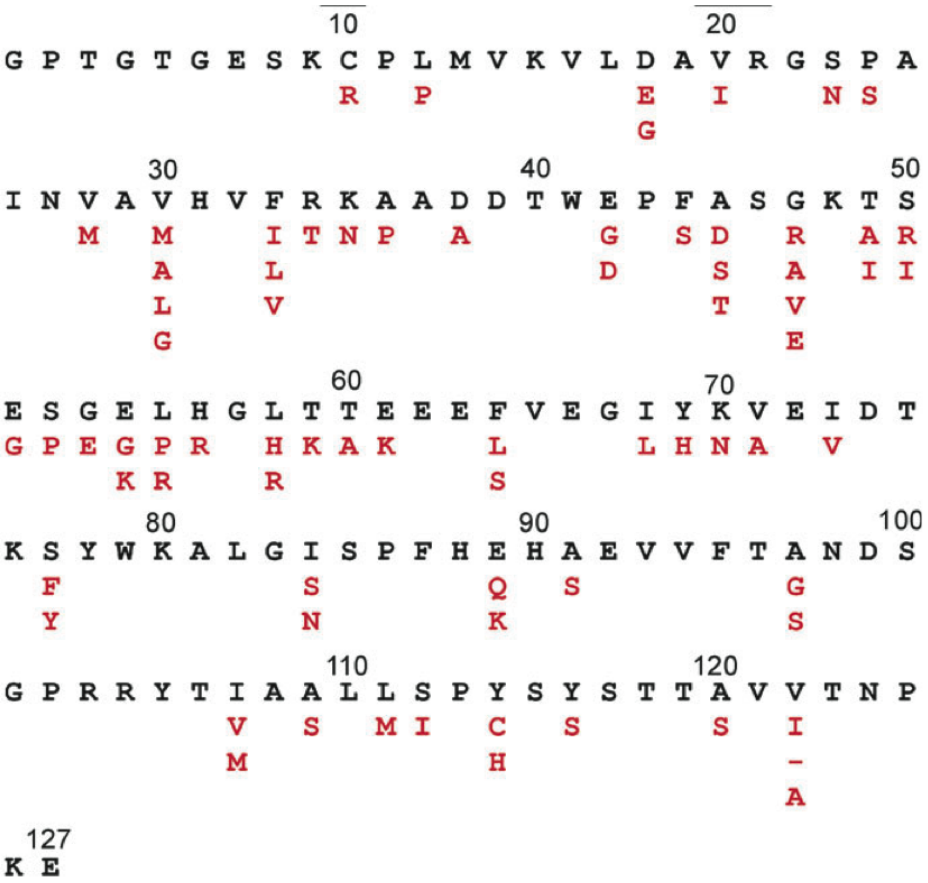


Figure 1.14. Human TTR amino acid sequence revealing the position of several amyloidogenic mutations (in red). Citations for each mutation can be found in the online database maintained by C. E. Costello at the Boston University School of Medicine. (http://www.bumc.bu.edu/msr/ttr-database/, accessed 12/09/2011)

Val30Met TTR (or V30M – where residue Valine in position 30 is mutated to a Methionine) was identified as a common underlying genetic variant of FAP in 1984 [290]. A large number of mutations have been detected since then, many of them associated with FAP, which are evenly distributed over the TTR sequence [291–294] (see **Figure 1.14**). V30M-TTR has been found in many kindreds around the world and is known to be the most common amongst the amyloidogenic TTR mutations [293–295]. In 1994, 1233 patients had received a FAP diagnosis at Centro de Estudos de Paramiloidose in Porto and more than 1500 patients were registered in Portugal almost ten years later (M. J. M. Saraiva, PhD, oral communication, September 2003). An epidemiological survey conducted by Holmgren et al. in Northern Sweden estimated that the number of ATTR V30M gene carriers in a total population of 500,000 was approximately 7500 [296]. In Japan, more than 394 FAP patients had been found by 2005. These were geographically scattered across the country and were genetically independent [295].

Figure 1.15. Distribution of FAP around the world. Locations of foci of patients with FAP amyloidogenic V30M-TTR are presented. The size of the circles is related to the number of patients at each location (extracted from reference [295]).

The question of whether there is a common origin in the foci for a mutant allele is still a matter of debate. Portugal, Sweden and Japan are geographically distant and no consanguineous links between foci have been detected. However, inspired by renowned historical relations between Europe, Japan, North and South America, and Africa, Continho hypothesized that a mutant allele in the Portuguese kindred could be the single origin of the mutation for FAP foci around the world [297]. This hypothesis has not been scientifically tested, though. Ohmori et al. compared haplotypes in multiple foci of FAP patients and found out that a common founder could be conceivable for Japanese and Portuguese patients, as well as for Portuguese and Spanish patients, but not for Swedish and other patients [298]. Additional investigation of genotypes and phenotypes are still required.

Besides V30M-TTR, several other TTR variants are implicated in different forms of amyloidosis characterised by diverse symptoms of variable severity and age of onset. The L55P-TTR variant, for example, albeit rare, is related with an extremely aggressive form of TTR amyloidosis, involving the heart, peripheral nerves and other organs, with patients dying at earlier ages and 5 to 10 after diagnosis [299]. The V122I-TTR variant is the most common amyloidogenic mutation worldwide. It is responsible for the onset of FAC, predominantly in individuals of African descent. It is estimated that approximately 4% of African Americans (1.3 million people) are heterozygous for the V122I allele [300]. FAC does not result from loss of TTR function (due to aggregation); it seems to be caused by tissue-selective TTR deposition in the heart [300,301]. The age of onset of FAC is similar to that of senile systemic amyloidosis (SSA) – typically >60 years of age – but FAC patients are considerably more prone to suffer cardiac failure (especially V122I homozygotes) [300,302,303].

A few TTR variants have been shown to play a protective role against amyloid fibril formation, when coupled with amyloidogenic mutations (compound heterozygotes), by interallelic transsupression. Mutations like R104H and T119M stabilise the TTR tetramer by increasing the energy barrier for dissociation. In the T119M-TTR variant, the stabilization caused by the mutation is so high that dissociation is almost completely prevented [304].

3.2.5. Amyloidogenic determinants and intermediates

Amongst the intrinsic structural changes necessary for amyloid formation by TTR upon tetramer dissociation, the displacement of the D-strand-loop-C-strand from its native position in the TTR monomer is believed to be essential to expose a new interface comprised by strands A and B, in turn required for monomer assembly into aggregates [182,305–307]. Site-directed spin labelling combined with electron paramagnetic resonance (EPR) [308] and hydrogen/deuterium (H/D) exchange studies by NMR spectroscopy have supported this hypothesis [309] and a computational model of an amyloid protofilament of TTR was proposed by Correia et al. [74]. In this work, proper subunit-subunit docking was possible only after full displacement of the D-strand–loop–C-strand segment from the β-sandwich core of the TTR monomer (see **Figure 1.11**).

Further computational studies by Rodrigues et al. using high-temperature MD simulations have suggested that disruption and displacement of β-strand D, followed by separation of β-strand C from the core of the TTR monomer, exposes key residues and a hydrophobic interface – embodied by β-strands A and B – that may promote monomer-monomer interactions [209]. This event occurs earlier in simulations of the L55P-TTR variant and is only witnessed in simulations WT-TTR after extensive unfolding of the entire monomer occurred. These observations are in agreement with the amyloidogenic propensities of these two TTR variants. The simulations have also evidenced a pronounced unfolding of the α-helical motif of TTR into coils and turns in L55P-TTR, contrasting with the unfolding behaviour of WT-TTR. Loss of α-helical content and transitions from α-helix to β-sheet are known to be a determinant of some amyloid diseases, namely BSE and Creutzfeldt-Jakob [310].

The atomic-scale events observed by Rodrigues et al. provide a (possible) picture of what the intermediate structures in the unfolding pathways of TTR may look like. Whether these potentially amyloidogenic species are on- or off-pathway intermediates to fibril formation is still a matter of debate, but these simulations show that the intermediates can populate multiple unfolding routes in relatively high amounts [209]. These computa-

tional insights are useful to the understanding of the molecular mechanisms behind TTR amyloid and, thus far, could not be offered by any experimental methods. Structural bioinformatics methodologies have indeed aided efforts taken towards a deeper understanding of the determinants of TTR amyloid and amyloid in general. Another remarkable example has been recently provided by Cendron et al., where the authors applied a battery of protein stability prediction algorithms to analyse a diverse set of TTR variants [311]. Collectively, their results hold up the notion that the high amyloidogenic propensity of pathogenic TTR variants is mostly determined by destabilization of their native structures, rather than by a higher β-aggregation tendency.

3.2.6. Mechanism of neurotoxicity

The biochemical basis by which amyloidogenic proteins exert neurotoxic effects is still elusive [312]. It is plausible that neurotoxicity is a common feature to all types of amyloid. It even seems possible that the amyloid conformation itself may be the toxic principle. There is limited evidence implying specific amino-acid sequences for the toxic effects [193]. In fact, deposition of gelsolin and apolipoprotein AI, which hold little or no amino-acid sequence similarity to TTR, has been proposed to cause FAP [295]. Thus, toxicity does not seem to be linked to the presence of specific amino acid sequences, so much as to specific conformational motifs of structure-rich protein aggregates.

Several studies have been undertaken in order to understand the mechanism of neurotoxicity in FAP. Hou et al. presented a model of the mechanism of TTR-induced neurotoxicity, according to which TTR monomers, low-molecular-mass nuclei, oligomers or protofibrils are the major toxic species [313] (see **Figure 1.16**). Their studies show that these low-molecular-mass diffusible species can bind to lipid membranes. This process causes disruption of the structure of the lipid rafts, thus inducing alterations in the membrane. This in turn leads to activation and calcium entry through voltage-gated calcium channels (VGCC). Moreover, TTR may also bind to a receptor for advanced glycation end-products (RAGE) to affect MAP kinase signalling [314] and induce endoplasmic reticulum (ER) stress, with release of calcium from intracellular stores [315]. ER stress is potentially cytodestructive, and RAGE receptors regulate cascades that are implicated in mitogenesis, cellular injury, death, and apoptosis [316]. The larger amyloid deposits are less toxic than the low-molecular-mass diffusible species but may constitute a local pool of TTR that may dissociate back into toxic species.

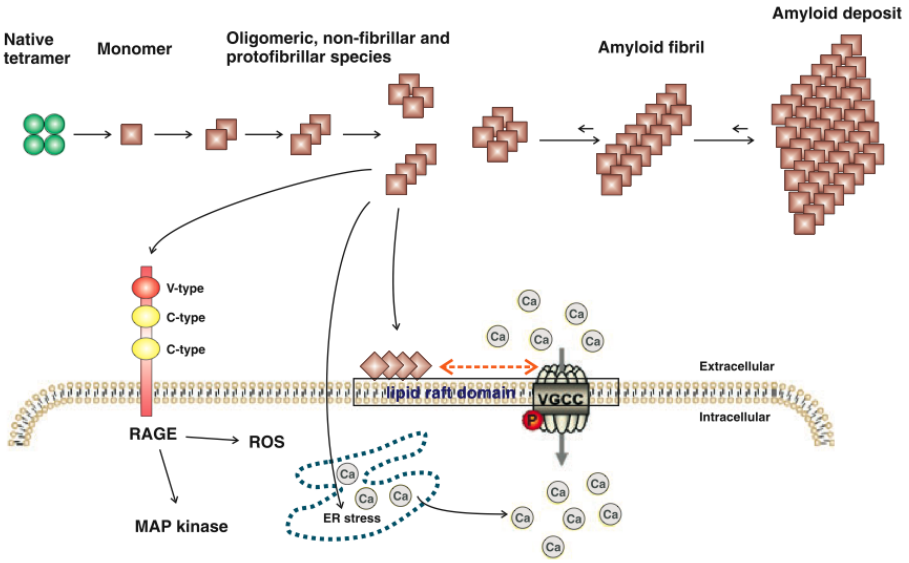


Figure 1.16. Hypothetical mechanism illustrating how TTR may cause neuronal dysfunction. ROS, reactive oxygen species; V-type, V-type binding domain on RAGE; C-type, C-type binding domain on RAGE. This figure was extracted from reference [313].

3.2.7. Extrinsic factors in TTR amyloid

The most accepted model for FAP pathogenesis is grounded on the presence of point mutations resulting in instability of the native TTR tetramer, followed by early dissociation to non-native monomeric species with propensity for self-association [204]. Nonetheless, this model does not seem to account for a few important aspects surrounding TTR amyloid. First and foremost, non-mutated TTR can also form amyloid in vivo, leading to SSA in elders. Moreover, TTR keeps accumulating into amyloid deposits even upon liver transplantation [317]. The fact that carriers of the same amyloidogenic TTR variant develop disease symptoms in times varying by decades [318] also remains unexplained. Even though TTR is mainly produced in liver, deposition of amyloid fibrils typically occurs in a number of different tissues in different patients. With these examples in mind, it is clear that non-genetic/endogenous factors are required to trigger TTR deposition in a specific tissue. Besides the conformational modification of the amyloidogenic protein, the most established factors in the field of amyloid are the co-deposition of glycosaminoglicans (GAGs) and of serum amyloid P component (SAP), and post-translational modifications.

There is evidence that glycosaminoglicans (GAGs) may play a role in TTR deposition. GAGs represent a diverse group of highly sulfated carbohydrates that controls several physiological processes [319]. GAGs are normally detected in proteoglycans, attached to the proteinaceous core. Amongst the most well known GAGs are heparan sulfate (HS),

dermatan sulfate, keratan sulfate and chondroitin sulfate (CS), whose carbohydrate scaffolds vary considerably in extent of sulfation. GAGs are also found attached to amyloid deposits, including TTR aggregates. In FAP, amyloid deposits often occur in the endoneurium [320], which is rich in extracellular matrix proteins, namely CS proteoglycans [321]. Amyloid deposition is also frequently found in association with HS proteoglycan-rich basement membranes [322]. In fact, HS can also bind to amyloid and induce fibrillogenesis [323], and as been suggested as a therapeutic target in amyloidogenesis [324–326].

Amongst other non-genetic factors, several post-translational modifications may directly affect protein structure, stability and function. Glycation, in particular, is a common link between amyloidogenic pathologies, including Alzheimer's and Parkinson's diseases and FAP [324–326]. Methylglyoxal is one of the strongest glycation agents *in vivo* and it is produced in all living cells as a non-enzymatic glycolysis by-product [327]. Methylglyoxal glycation was indeed identified in Alzheimer's disease [328–330], Parkinson's disease [331], dialysis-related amyloidosis [332], prion diseases [333], hemodialysis-related β2 microglobulin amyloidosis [334], and murine ApoAII amyloidosis [335]. Likewise, a report showed that amyloid fibrils extracted from FAP patients are glycated by methylglyoxal [336]. Lately, Costa et al. found out that fibrinogen is a specific glycation target with increased glycation in FAP patients, besides being one of the main TTR interacting proteins in plasma [337]. The authors observed a significant reduction of the chaperone activity *in vitro* upon glycation by methylglyoxal, and concluded that fibrinogen prevents plasma TTR thermal-induced protein aggregation. A new model for TTR amyloidogenesis *in vivo* was thus proposed, wherein increased glycation of fibrinogen reduces its chaperone activity, reducing TTR tetramer stability and triggering the pathway to aggregation, amyloid formation and disease (**Figure 1.17**).

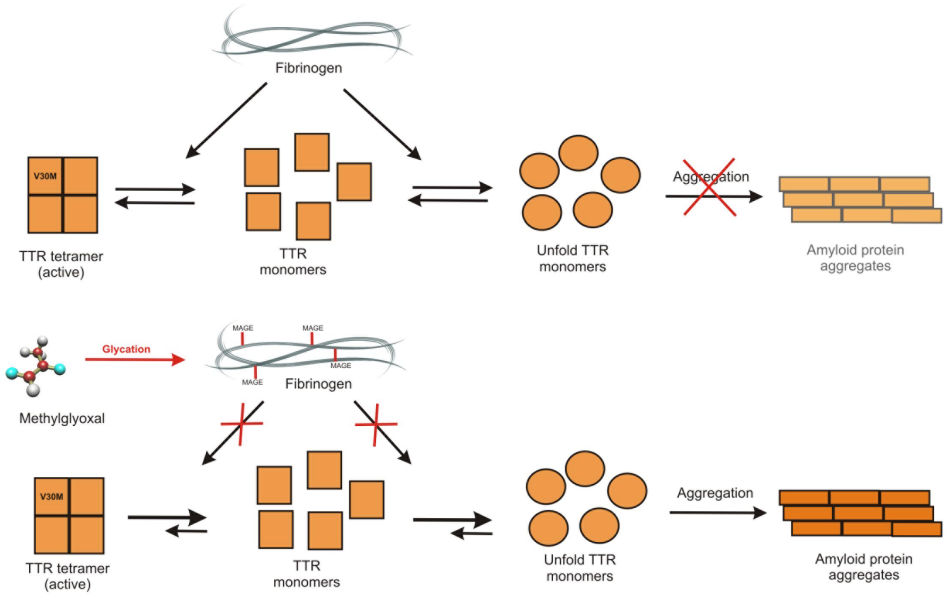


Figure 1.17. Schematic representation of the molecular effects of fibrinogen glycation *in vivo*. *In vitro* studies showed that amyloidogenic TTR variants hold higher propensity to follow an aggregation route towards the formation of amyloid. Considering the findings by Costa et al. showing that fibrinogen interacts with TTR in the plasma, it was proposed that fibrinogen chaperone activity decreases TTR's tendency to form fibrils. This figure was adapted from reference [337].

3.3. Therapeutic strategies in amyloid

In parallel with the exponential growth of biochemical data on key events of the fibrillogenic process, several reports have shown that small molecules, through interaction with either the amyloidogenic proteins or with the common constituents, can modulate the kinetics of amyloid fibril formation or facilitate amyloid reabsorption. These molecules can be categorized on the basis of their target and mechanism of action, according to the following features: (*i*) compounds that stabilise the amyloidogenic protein precursor; (*ii*) compounds that halt fibrillogenesis by interaction with partially unfolded intermediates and/or with low molecular weight oligomers populating the initial stages of fibril formation; (*iii*) compounds that act on mature amyloid fibrils and decrease their structural stability. Besides these three categories, there is a fourth group of compounds that can displace critical co-factors of the amyloid deposits, such as GAGs and SAP, and thus promote the dissolution of fibrillar aggregates.

Several reviews can be found in the literature describing in great detail both potential strategies and compounds discovered to inhibit amyloid [338,339]. For the sake of concision, over the following subsections I will focus only on the first three mechanisms of therapeutic action, based on the model for amyloid fibril formation by TTR originally

proposed by Quintas et al. [204] (see **Figure 1.18**), and point the reader to references [340–343] (and references therein) for more information on the fourth kind of strategies.

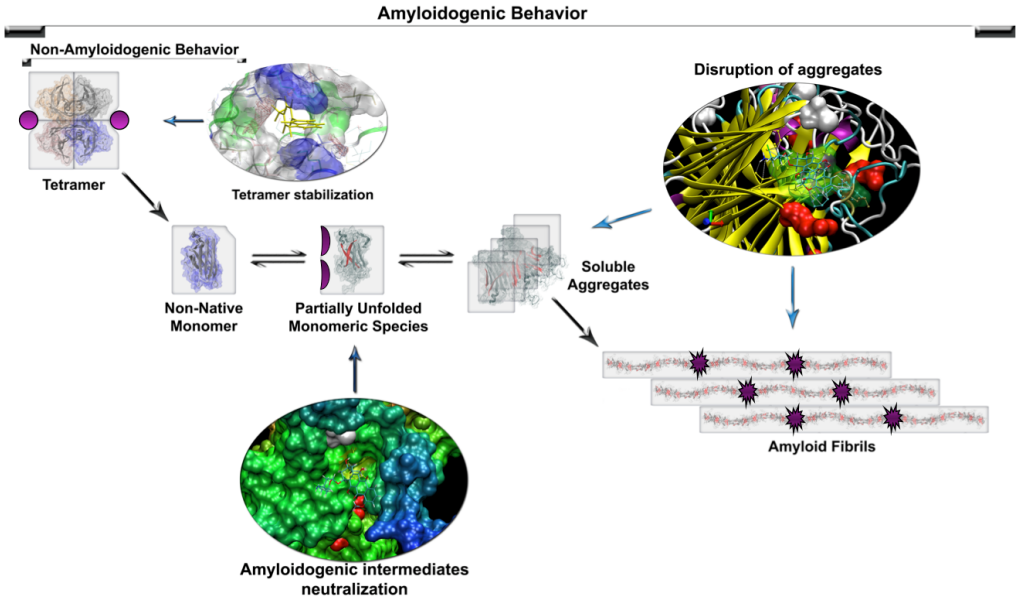


Figure 1.18. Model for amyloid fibril formation by TTR under common physiological conditions. Three main points of therapeutic intervention are highlighted by the icons in purple and the elliptical insets. This figure was inspired by reference [204].

3.3.1. Stabilization of the native protein

A rule-of-thumb in protein biochemistry is that ligands that bind to a protein in a specific conformation are able to stabilise that particular conformation. This implies that binders of native-state proteins have the capacity to prevent (or at least postpone) protein dissociation and/or unfolding, which are triggering events of the fibrillogenic pathway. Transthyretin is an outstanding illustration of this concept. The observation that thyroxine not only binds but also stabilises the TTR tetramer and inhibits fibril formation [206] paved the way to the identification of several classes of structurally diverse inhibitors of fibrillogenesis. A number of examples is given in the next subsection, starting on page 56. The fact that the binding of thyroid hormones to TTR stabilises the protein against fibril formation in the brain [206] is a thrilling indication that the inhibition confirmed *in vitro* may also be valuable *in vivo*.

As will be explained in the last section of this chapter, this therapeutic strategy was followed throughout my PhD project, while seeking to identify novel and safe amyloid inhibitors holding high selectivity for TTR in the human plasma and adequate (drug-like) properties for this target compartment. More details about the techniques for studying

the binding of small molecules to TTR and for evaluating its inhibitory effect of fibril formation will be provided in subsequent sections and chapters of this thesis.

3.3.2. Neutralization of intermediate species

Compounds that stabilise a certain precursor protein are unlikely to produce an anti-amyloid effect in other amyloidogenic proteins. Conversely, given the common structural link that unites amyloid aggregates, compounds that neutralize or inhibit aggregation into cytotoxic intermediates are most likely to exert similar anti-amyloid activity towards many types of amyloid.

Several families of structurally uncorrelated compounds have been investigated for their ability to interfere with the self-aggregation process of peptides and proteins driving the amyloid fibril formation. The vast majority of such molecules was tested *in vitro* on Aβ peptides – mostly Aβ 1-40 and Aβ 1-42. However, this approach to the development of new therapeutics for Alzheimer's disease seems to suffer from at least one of several shortcomings: uncertainty about the mechanism of action, lack of specificity, limited central nervous system penetration, or compound intrinsic toxicity. Indeed, for many of these compounds it cannot be ascertained what step of the aggregation process they are targeting. The use of synthetic Aβ peptide in *in vitro* experiments is itself associated with a high degree of batch-to-batch variability, which affects the reproducibility and reliability of the results [344].

Techniques for studying the interaction between small organic molecules and protein intermediates and aggregates, as well as characterizing the kinetics of the polymerization process, include dye staining, fluorescence and mass spectrometry, dynamic light scattering, circular dichroism, NMR and X-ray diffraction, along with electron and atomic force microscopy for imaging data.

The Congo red dye was found to inhibit Aβ and amylin fibril cytotoxicity, either by preventing fibril formation, through increasing soluble peptide levels, or binding to preformed fibrils [345] (see **Figure 1.19**). Later on, it was also shown that Congo red stabilises Aβ monomers and thus inhibits Aβ 1-40 polymerization [346]. Unfortunately, this compound is unsuitable to become a drug candidate, not only due to toxicity issues but also to a highly variable concentration-response detected *in vitro* with different amyloidogenic peptides and proteins [347,348]. The aggregation kinetics, conformational changes and thermodynamics of the model system immunoglobulin light chain variable domain all lead to think that, at relatively low concentrations, Congo red might bind and promote a partially unfolded state of the protein, whereas at higher concentrations it

might favour a denaturated state, therefore respectively accelerating or stalling fibrillogenesis [349].

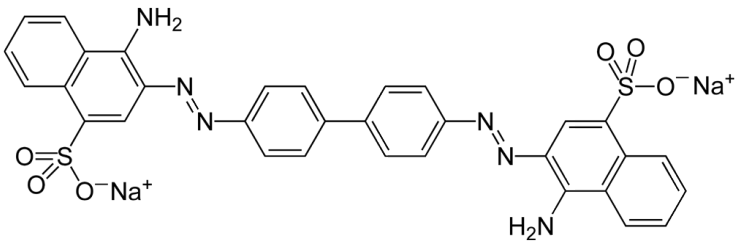
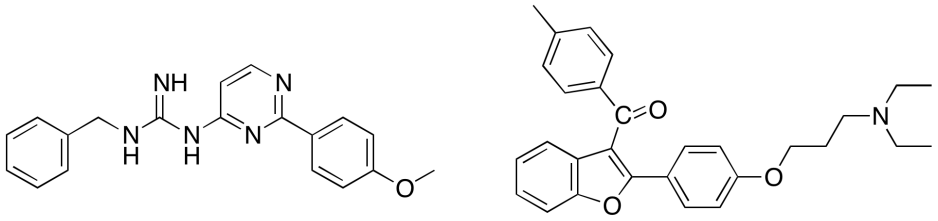


Figure 1.19. Chemical formulae of the Congo red dye. A model for the interaction with polypeptides in amyloidogenic conformation was originally proposed by Li et al. [350]. An antiparallel dimer of Aβ 1-42 may interact with Congo red through hydrophobic interactions between the central part of the peptide and the biphenyl groups of the dye, while two salt bridges would be formed between two opposed Lysine residues and the two sulfonate groups.

Effective inhibitors of Aβ self-assembly have been widely described, typically in two defined categories: peptide-based and small molecule inhibitors [351–354]. Amongst a number of small molecules that have been reported to inhibit Aβ fibril formation are rifampicin [355,356], melatonin [357], an antioxidant compound called nordihydroguaiaretic acid (NDGA) [358] and β-cyclodextrins [359]. A class of inhibitors including apomorphine and its analogues was shown to interfere with Aβ 1-40 fibrillogenesis and to promote accumulation of stable protofibrillar intermediates, thus blocking fibril formation [360]. More recently, Sood et al. has been studying the role of molecular chirality in the action Aβ self-assembly inhibitors using enantiomeric organofluorine compounds [361].

Information about small molecule inhibitors of the self-assembly of partially unfolded TTR monomers is scarce. Cardoso and Saraiva identified two interesting scaffolds that were able to inhibit, to an extent of approximately 60%, the association between intermediate species leading to TTR fibril elongation [362] (see **Figure 1.20**).



SKF 99130 SKF 64346

Figure 1.20. Chemical formulae of two compounds identified by Cardoso and Saraiva to inhibit the extension of L55P-TTR fibrils. SKF 99130 (left pane) was able to promote approximately 57% inhibition and SKF 64346 (right pane) approximately 64% [362].

3.3.3. Disruption of amyloid fibrils

Therapeutic approaches against amyloid can also include disrupting or dissolving mature amyloid fibrils. Several compounds have been shown to directly target amyloid deposits and/or to promote amyloid disassembly.

The anthracycline 4'-iodo-4'deoxy-doxorubicin (IDOX), a derivative of doxorubicin, is a good example of a compound falling in this category (see **Figure 1.21**). Originally developed by Pharmacia, IDOX is an antineoplastic agent that was serendipitously found to induce amyloid resorption in patients with immunoglobulin light chain amyloidosis (AL) [363]. Unlike its parent doxorubicin, IDOX binds strongly to several types of natural amyloid fibrils tested: amyloid A, Aβ-peptide, AL, β_2-microglobulin and insulin [364]. These results suggested that this compound specifically recognizes common structural features of the amyloid fibrils and that iodine plays an important role in determining amyloid targeting. Moreover, and again by contrast with doxorubicin, IDOX can also inhibit amyloid fibril formation by binding to insulin fibrils synthesized *in vitro*, mediated by key hydrophobic interactions at two separate binding sites [364]. IDOX has also been shown to be capable of disaggregating TTR amyloid fibrils *in vitro* into amorphous material [365]. Studies of the interaction of IDOX with an "amyloid-like" oligomer isolated throughout the crystallization of L55P-TTR revealed that this compound induced rapid dissociation of the crystals [366].

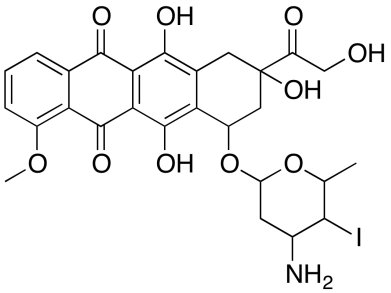


Figure 1.21. Chemical formulae of 4'-iodo-4'deoxy-doxorubicin (IDOX).

While interesting in essence, this potential therapeutic strategy is haunted by concerns that compounds able to solubilise the amyloid fibrils could favour the release and accumulation of highly toxic oligomers [367–369]. Surprisingly, Cardoso et al. have shown that the fragments resulting from IDOX-induced disruption of TTR fibrils are not cytotoxic, and also reported a group of tetracyclines and nitrophenols that exhibit fibril disruption activity [370]. Together with IDOX, these molecules may be used as templates towards the identification of a pharmacophore defining their interaction with mature amyloid fibrils. Notwithstanding, further theoretical and experimental studies need to be conducted before disruption of fibrils becomes an appealing strategy to be pursuit in the development of novel therapies for FAP and other amyloid diseases.

3.3.4. An invasive approach: the case of TTR amyloid

The idiosyncrasies and peculiarities that characterise FAP as the most prominent TTR-related amyloidosis extend to the unique therapeutic strategy of established efficacy that has been in place until the imminent launching of tafamidis meglumine, the first drug against FAP. Since TTR is mainly synthesized in the liver, liver transplantation is a crude form of genetic therapy that removes the pool of the amyloidogenic precursor. The first liver transplantation on a FAP patient happened in Sweden in 1990 and the approach has then been followed all over the world.

Even though liver transplantation has been reported to slow down the progression of FAP [371], with improvements in autonomic function, gut symptoms, nutritional state [372] and detectable regression of the TTR amyloid deposits 1-2 years upon transplantation [373], the procedure does not always result in amelioration of symptoms, and progression of cardiomiopathy after transplantation in FAP patients has been reported [317]. Furthermore, liver transplantation is very costly and several problems are known to arise from the procedure [374,375] (common to other organ transplant procedures). Timing is key and yet one of the major setbacks: firstly, FAP gene carriers who show no clinical symptoms cannot undergo liver transplantation before disease onset; secondly,

clinical complications present before the surgery will remain unchanged after the procedure is taken.

3.4. Pharmacology of TTR amyloid inhibitors

The study of TTR-T_4 interactions may well represent one of the first landmarks of the use of computer graphics towards a detailed understanding of protein-ligand interactions. Indeed, it was back in 1982 that Blaney et al. pioneered molecular modelling studies of T_4 and several analogues by calculating their molecular surfaces along with TTR binding sites [376] (see **Figure 1.22**). More than a decade later, the identification of flufenamic acid, a known anthranilic acid with non-steroidal anti-inflammatory activity, as a potent TTR tetramer stabiliser was the first stepping stone in the quest for strong TTR amyloid inhibitors [377]. Contemporarly, Kelly's group found out that diflunisal had comparable activity in terms of inhibiting TTR fibril formation. They tested 78 structurally diverse compounds, including benzophenones, anthraquinones, flavones, stilbene derivatives and adamantane derivatives; an eye-opener endeavour that singled out the flavone apigenine and the stilbene derivative resveratrol, along with flufenamic acid, as the most promising inhibitors [212]. In 2000, Klabunde et al. offered an exceptional summary of a first decade of preliminary efforts to identify promising drug leads against TTR amyloid [215].

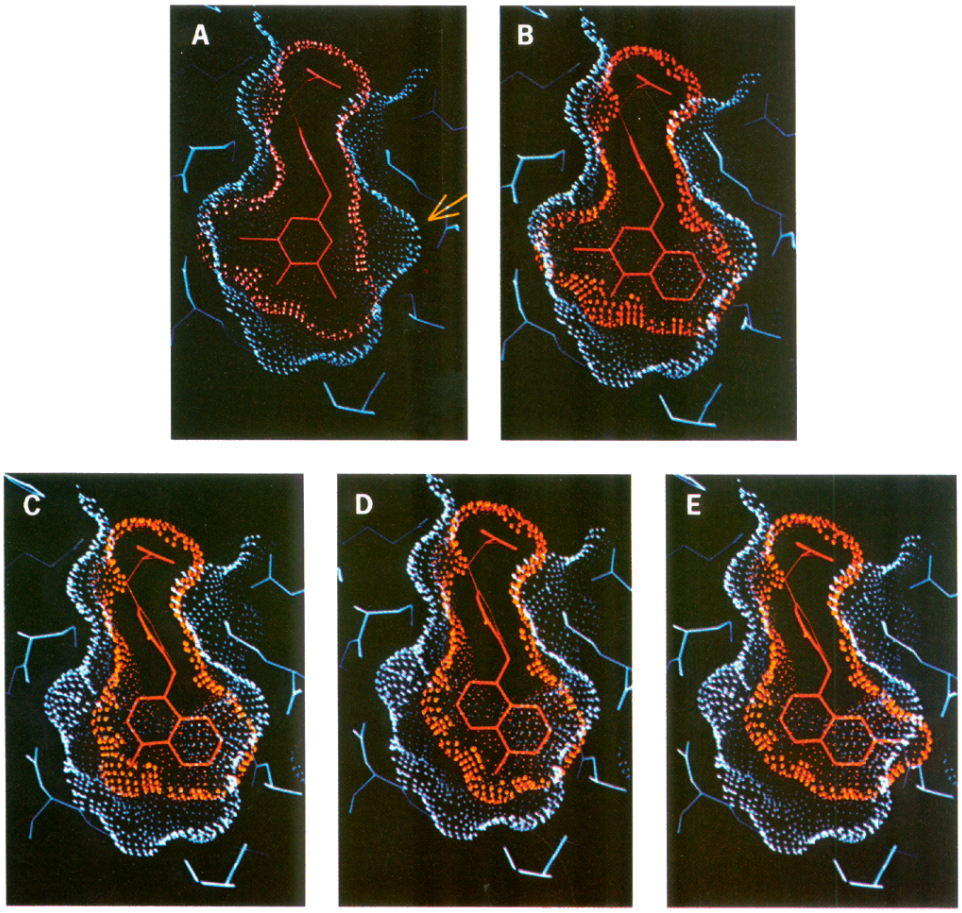


Figure 1.22. Early molecular representations of the interactions between TTR and T_4 analogues. The different panes correspond to cross-sections of T_4 analogue-TTR binding surfaces: (A) L-thyroxine and four naphthyl analogues (B, C, D and E). The orange arrow indicates an empty pocket. This figure was extracted from reference [376].

Either by serendipity or design, over the past decade a number of new effective amyloid inhibitors have been discovered. Lately, fragment-based approaches have been employed to identify the most efficient combinations for the substructural elements composing the classic binder [378–380]. Amongst the most common threads to TTR stabilisers are:

- two aromatic rings, often linked by a spacer;
- one carboxylic group, either in *meta* or *para* position;
- one or several halogen atoms attached to the central scaffold.

From the pharmacology viewpoint, a critical aspect associated to most TTR binders is indeed the recurrent occurrence of halogenated substituents. As will be discussed in greater detail in Chapter 4, possibly due to the presence of two charged Lysines at both entrances of the binding channel, TTR binding sites seem be more electropositive than all

other surfaces of the protein. This may explain TTR's natural propensity to bind electro-negative/halogenated compounds.

The use of halogen atoms in Medicinal Chemistry is a quite controversial subject. In a personal communication at the University of Marburg (Germany), in 2006, Professor Hugo Kubinyi showed great scepticism and concern about bioactive molecules containing halogens, opening up exceptions for only a few molecules bearing fluorine and one very specific molecule bearing one bromine atom. Still, while some of the most reputable medicinal chemists show concerns about halogens' potential for bioaccumulation and toxicity, discouraging their use in drug design [381], other researchers highlight the versatility of these exotic atoms [382].

A second critical aspect from the pharmacology and drug design viewpoints is that TTR binding sites display cooperativity effects in ligand binding. Indeed, it has been reported for several ligands that the binding of the first molecule to one TTR binding site causes a decrease in affinity of binding of the second molecule to the unoccupied site (negative cooperativity). This is the case for T_4, flufenamic acid and perhaps most other known TTR ligands. The true nature of this *allosteric* effect is hard to pin down. Without exception, all crystal structures of TTR complexes available in the PDB show two occupied pockets within the asymmetric unit; none was able to capture a single ligand-binding event. In parallel, and as pointed out before, the two TTR binding sites are completely identical.

The first attempts to understand cooperativity effects in TTR were made by Cheng et al., where, based on binding data at several pH values, the authors suggest that an electrostatic interaction may contribute to negative cooperativity [383]. As a follow up to this work, Irace and Edelhoch tried to tackle conformational changes in TTR binding sites upon T_4 binding. The authors followed ligand-binding effects on aromatic chromophores by ultraviolet fluorescence and, since the perturbation was very different for the two binding sites, suggested that the binding of the first ligand produces the same conformational effect, thus maintaining the symmetry of TTR [384]. More than a decade later, Neumann et al. presented a notable attempt to shade light on the structural basis of cooperativity by comparison of distances between corresponding β-strands on each binding site [385]. Despite the authors' claims, given the limited size of their sample (10 TTR apo+holo structures) and the lack of experimentally-determined atom coordinates of the full biological unit, their results look statistically inconsequential.

All these aspects will be covered in greater detail in Chapters 2 and 3 of this thesis.

3.4.1. Endogenous binders

Thyroid hormones and retinol-binding protein are the major endogenous binders of TTR. Thyroxine (T_4), or 3,5,3',5'-tetraiodothyronine (**Figure 1.23**), is the principal hormone secreted by the follicular cells of the thyroid gland. It is synthesized via the iodination and covalent bonding of the phenyl group of tyrosine residues found in a primary peptide, thyroglobulin, and then secreted into thyroid granules. These iodinated biphenyl compounds are cleaved from their peptide backbone when stimulated by thyroid-stimulating hormone. T_4 is transported in blood mainly bound to thyroxine-binding globulin (TBG), to an extent of near 99.95%, and, to a lesser extent, to transthyretin and serum albumin. The half-life of thyroxine once released into the blood circulatory system is approximately 1 week.

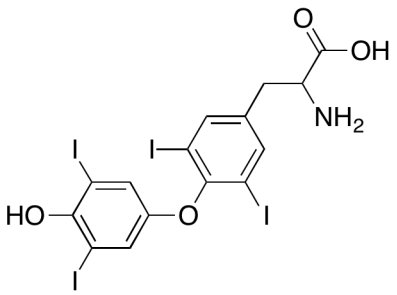


Figure 1.23. Chemical formulae of thyroxine (T4) the major endogenous binder of TTR.

TTR-T_4 interactions are directed by non-polar contacts between the central hydrophobic portion of T_4 and the side chains of Leu-17, Ala-108, Leu-110, Thr-119 and Val-121, and between the charged groups of the ligand's tail and the side chains of Lys-15 and Glu-54. Binding site complementarity, in terms of shape, is far from ideal: even though the crystal structures of TTR-T_4 complexes show that one T_4 molecule binds deeper to one of the sites than the other, T_4 does not bind deep in TTR binding sites (PDB entry 2rox). This is due to steric hindrance of the four iodine atoms that render T_4 a bulky molecule. In fact, the reported affinity of T_4 for TTR is modest (K_a's of 1×10^8 and 9.6×10^5 M^{-1}, negative cooperativity) compared to its affinity for thyroxine binding globulin ($K_a = 6 \times 10^9$ M^{-1}), the main carrier of T_4 in the plasma.

3.4.2. Natural products

Some natural products have been shown to bind TTR and reduce TTR fibril formation. Resveratrol, or trans-3,4',5-trihydroxystilbene, is an antioxidant found in red wine and grape juice that was shown to inhibit platelet aggregation and play a role in the preven-

tion of congestive heart failure due to coronary thrombosis [386,387]. The anti-apoptotic activity of resveratrol is mediated by direct inhibition of the main arachidonate metabolizing enzymes 5-lipoxygenase, 15-lipoxygenase and cyclooxygenase (COX) [388]. The X-ray structure of TTR in complex with resveratrol shows good binding site complementarity at a minimum energy ligand pose (PDB entry 1dvs). Due to the two-fold symmetry of TTR's binding sites, resveratrol presents two binding modes that are related by 180° rotation along the channel. TTR-resveratrol interactions are dominated by non-polar contacts between the stilbene moiety and the side chains of Lys-15, Leu-17, Ala-108, Leu-110, Thr-119 and Val-121. The apparent rigidity of the stilbene scaffold provides a minimum entropic penalty upon binding, favouring association. Yet, the strong potency of resveratrol derives from the enthalpic contribution of two hydrogen-bonding interactions between its hydroxyl groups and the side chain Nε of Lys-15 and the hydroxyl side chain of Ser-117. Because resveratrol extensively reduces TTR fibril formation, stilbene derivatives may be regarded as promising leads for the development of new drugs against TTR-associated amyloid diseases.

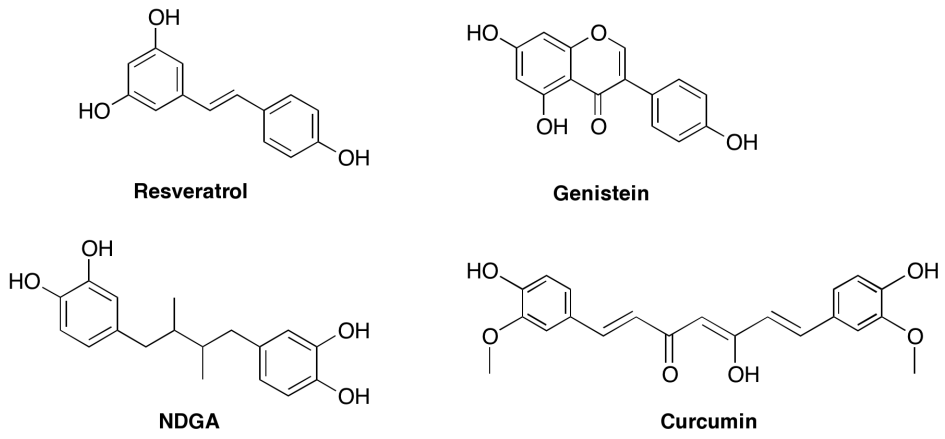


Figure 1.24. Chemical formulae of four natural products identified to bind TTR and inhibit fibril formation.

Genistein is the major isoflavone present in soy and has been evaluated in preliminary trials for the treatment of breast, prostate and uterine cancers, osteoporosis, cardiovascular disease and menopausal symptoms. Genistein was shown to bind TTR (K_{d1} = 40 nM, K_{d2} = 1.4 μM), prevent tetramer dissociation and reduce acid-mediated fibril formation to < 10% of that exhibited by TTR alone, including the V30M and V122I variants responsible for FAP and FAC, respectively [389]. Moreover, this natural product displays high selectivity for TTR in the human plasma. Based on this discovery, it is plausible to hypothesize that FAP and FAC patients could benefit from just increasing their intake of soy products.

In a recent study, Ferreira et al. studied the modulation of TTR fibrillogenesis by selected polyphenols [390]. In specific, they showed that both nordihydroguaiaretic acid (NDGA) and curcumin bind to TTR and stabilise the tetramer. Yet, while NDGA reduces TTR aggregation only to a limited extent, curcumin sturdily suppressed TTR amyloid fibril formation. The authors proposed that this strong inhibition is due to the generation of small "off-pathway" non-toxic oligomers. This is in line with curcumin's chemical structure (see **Figure 1.24**), which more easily resembles that of compounds shown to interfere with TTR intermediate species (see on page 52) rather than the structure of a classic TTR tetramer stabiliser.

3.4.3. Pollutants

The halogenated aromatic hydrocarbons, such as chlorinated dioxins, furans, biphenyls, naphthalenes and benzenes are amongst the most relevant environment pollutants, mostly due to their widespread occurrence and persistence in the environment and concentration in the food chain. In particular, the polychlorinated biphenyls (PCBs) are surrounded in concerns about their accumulation in the human body. The study of the interactions of PCBs with TTR started even before the effect of tetramer stabilization due to ligand binding became known. Rickenbacker et al. used computer graphics and predictions regarding binding affinities, and showed, through competition binding studies with labelled thyroxine, [^{125}I]-L-T_4, that the PCBs could bind TTR more effectively than T_4 [391]. Ironically, certain hydroxylated PCBs are today amongst the strongest TTR binders and fibril formation inhibitors [219]. In fact, the most potent inhibitor known to date, 4,4'-dihydroxy-3,3',5,5'-tetrachlorbiphenyl (PDB entry 2g5u), binds to TTR with the highest affinity ($K_{d1} \approx K_{d2} \approx 3.2$ nM) and displays positive cooperativity [378].

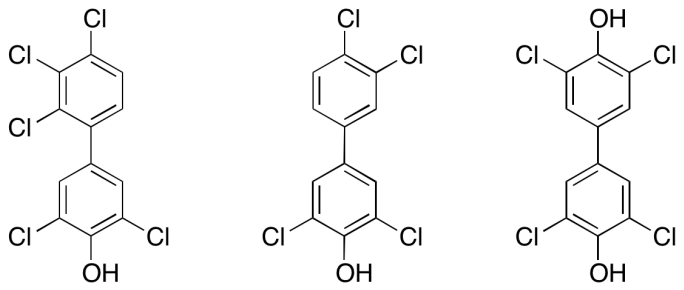


Figure 1.25. Chemical formulae of three polychlorinated biphenyls (PCBs) with high affinity for TTR. PCB18 is the most potent TTR amyloid inhibitor found to date and one of the few compounds known to exhibit positive cooperativity. The numbering is in accordance with reference [219].

TTR's "sweet and sour" attraction for the PCBs epitomises an important lesson in the field of drug discovery. *Potency* is not everything and *safety* should be well bear in the mind of chemists and modellers from the early stages of any drug design program.

3.4.4. Non-steroidal anti-inflammatory drugs

The non-steroidal anti-inflammatory drugs (NSAIDs) comprise a wide group of structurally diverse drugs that share the ability of controlling inflammatory processes, pain and fever. This is accomplished via the inhibition of different forms of an enzyme called cyclooxygenase, responsible for the production of prostanoids (protaglandins, prostacyclins and thromboxanes) involved in the inflammatory response. Flufenamic acid (FLU) and diflunisal were the first NSAIDs discovered to stabilise TTR against tetramer dissociation and amyloid fibril formation [212]. Soon after, Klabunde et al. showed that other NSAIDs, including diclofenac, flurbiprofen, fenoprofen and indomethacin, could also halt TTR fibrillogenesis, though to a smaller extent [215] (**Figure 1.26**). Later on, Miller et al. tested several NSAIDs along with a few designed analogs, for their ability to prevent acid-induced fibril formation and chaotrope-induced tetramer dissociation of the most common disease-associated TTR variants – V30M, V122, T60A, L58H and I84S [289]. All these were significantly stabilised by flufenamic acid and one of its analogs, and to a lesser extent by diflunisal.

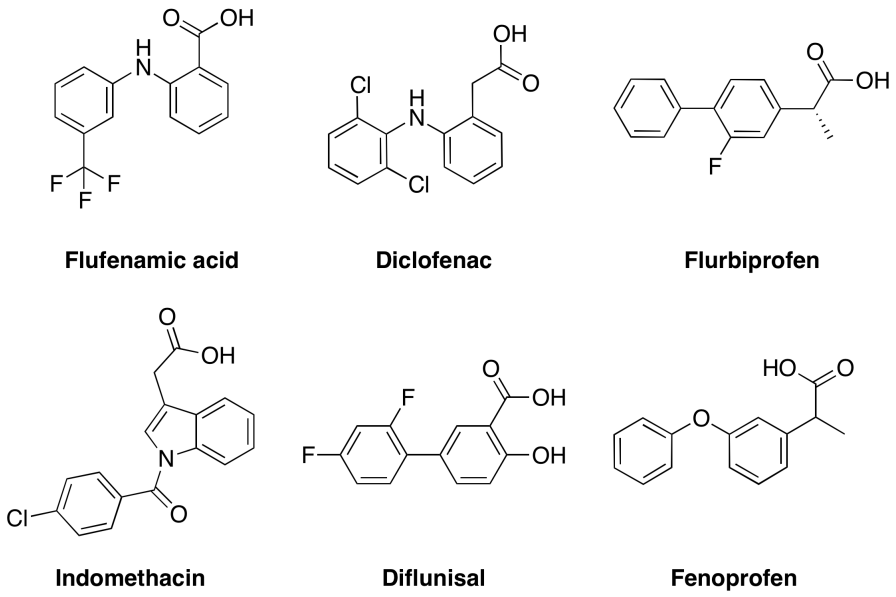


Figure 1.26. Chemical formulae of six nonsteroidal anti-inflammatory drugs (NSAIDs) identified to bind TTR and inhibit fibril formation.

In general, the crystal structures of TTR-NSAID complexes show that these ligands mediate inter-subunits hydrophobic interactions and hydrogen bonds, and that the additional

van der Waals contacts stabilise the tetrameric fold of TTR (PDB entries 1bm7, 1dvx, 1dvt, 3d2t). Interestingly, besides the two binding modes related by 180° rotation along the TTR binding channel, FLU exhibits two distinct binding conformations defined by rotation about one of its two aryl-amino bonds (PDB entry 1bm7). In spite of what may be seen as an entropic penalty to ligand binding, of all NSAIDs found to interfere with amyloid formation by TTR, FLU is the most potent binder (K_{d1} = 30 nM, K_{d2} = 255 nM) [377]. Thus, it was subsequently taken as a reference lead compound to the design of new inhibitors.

3.4.5. NSAID derivatives and designed inhibitors

Like with most NSAIDs, a prolonged intake of flufenamic acid would cause gastrointestinal ulcers, due to inhibition of isoform 1 of the enzyme cyclooxygenase (COX-1). COX-1 is responsible for the synthesis of prostaglandin, a substance with a critical role in gastric mucosal protection. Therefore, Klabunde et al. set out to discover new TTR amyloid inhibitors with specific binding to TTR and high gastric tolerability [215]. They used molecular modelling to study the interaction between TTR and flufenamic acid in crystal structures, and thus put forward a modified FLU (an *ortho* isomer) and a number of tricyclic scaffolds with high affinity for TTR – the dibenzofurans [215] and the phenoxazines [392] (**Figure 1.27**).

Even though *o*-FLU has shown better binding site complementarity than FLU (PDB entry 1dvs), it did not offer higher affinity for TTR than its isomer. It did however show higher selectivity and inability to inhibit COX-1 [215]. Modelling studies based on the TTR-FLU complex led to the hypothesis that tricyclic compounds could mimic the interaction of the two symmetry-related binding modes of FLU simultaneously. To test it, dicarboxydibenzofuran-4,6-dicarboxylic acid (DDBF) was evaluated and displayed excellent inhibitor activity. The tricyclic ring system of DDBF binds between residues Lys-15, Val-17 and Ala-108, stretching across the outer region of TTR binding sites (PDB entry 1dvu). These promising results encouraged Klabunde et al. to design similar compounds with a substituted-phenyl ring that could fill the inner cavity of the sites. A series of 15 N-phenyl-substituted phenoxazine-4,6-dicarboxylate were synthesized and tested, and m-trifluoromethyl-substituted N-aryl phenoxazine (PHENOX) was found to be the most promising inhibitor: K_{d1} = 78 nM, K_{d2} = 235 nM (PDB entry 1dvy) [215].

Isothermal titration calorimetry (ITC) studies revealed an interesting difference between PHENOX and FLU. The binding of FLU to wild type TTR is mainly enthalpically driven: for the first binding event, ΔG= -10.25 ±0.25 kcal mol^{-1}, ΔH= -11.4 ± 0.7 $kcal.mol^{-1}$. As expected though, given the observed multiple binding modes of FLU, the entropy associated

with binding and the conformational changes in the protein is unfavourable (TΔS= -1.15 ± 0.7 kcal.mol^{-1}). Conversely, there is an important favourable entropic contribution for PHENOX binding: for the first binding event, ΔG= -9.67 ±0.14 kcal mol^{-1}, ΔH= -4.52 ± 0.4 kcal mol^{-1}, TΔS = 5.15 ±0.45 kcal mol^{-1}. This can be due to the lower flexibility and larger size of PHENOX, which result in a minor conformational entropy penalty and in the ability to displace more water molecules from TTR binding sites, respectively [215].

Additional studies conducted in Kelly's group used structure-based principles to synthesize and evaluate several dibenzofuran derivatives, focusing more on the ability to bind TTR in the plasma with high affinity and selectivity [222]. In the meanwhile, other research groups have joined the race. Saraiva's group has proposed the use of iodine in compounds like diflunisal, as a means to attain more specific and effective drugs [393,394]. Surolia's group reported a new class of inhibitors based on biphenyl ethers (BPE) [395]. It is noteworthy that besides stabilizing TTR tetramer, the strongest BPEs were also shown to inhibit fibril elongation and disrupt pre-formed fibrils [395]. More recently, the mechanisms behind this polyvalent action were further investigated in Surolia's group, using quantum dots (QD) technology [396]. According to the authors, the exceptional optical properties of BPE-QD nanoparticles allowed following the slow association of the small organic molecules with mature amyloid fibers and a new mechanism of disruption was proposed.

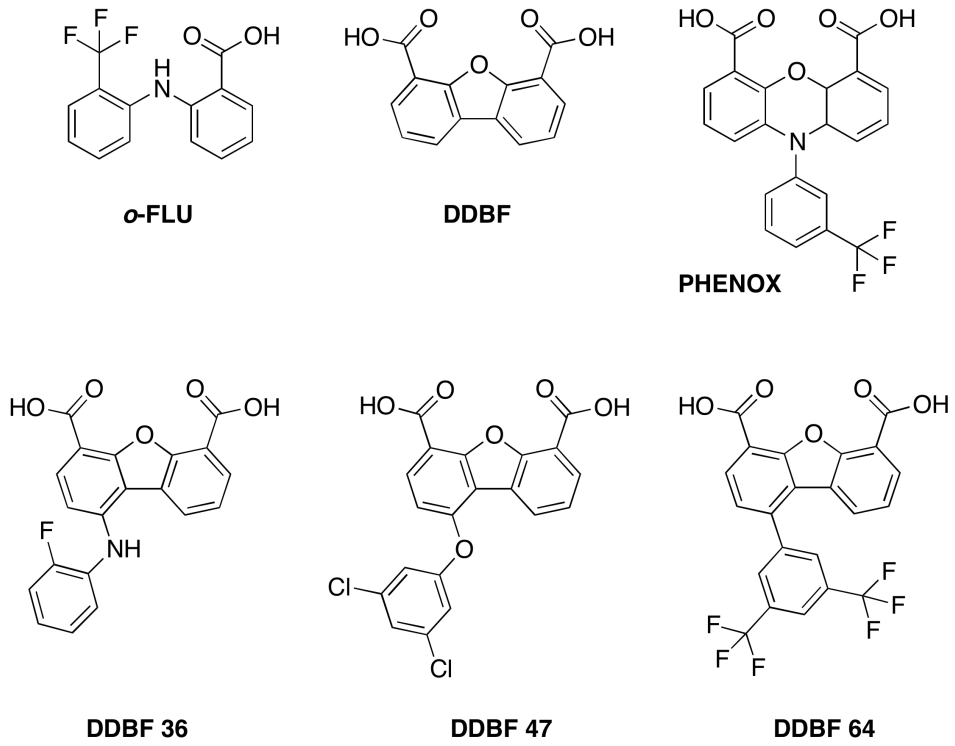


Figure 1.27. Chemical formulae of six TTR amyloid inhibitors inspired by the structure of flufenamic acid. Top row: from left to right, *ortho*-tri-fluoromethylphenyl an-

thranilic acid (differs from FLU in the positioning of the CF_3 substituent); DDBF is dicarboxydibenzofuran-4,6-dicarboxylic acid (a lead compound); PHENOX is N-*meta*-trifluoromethylphenylphenoxazine-4,6-dicarboxylic acid. Bottom row: three DDBF derivatives representing 3 sub-families: DDBF 36, a biarylamine with high inhibitory activity; DDBF 47, a biaryl ether with the highest affinity for TTR in plasma, DDBF 64, a biaryl showing good compromise between affinity for TTR in plasma and inhibitory activity.

One decade after the release of the crystal structure of flufenamic acid in complex with TTR, more than one hundred effective TTR amyloid inhibitors have been reported, a few of which have been protected by patent laws. The majority of the potent and selective amyloid inhibitors identified to date can be clustered into three main structural families: bisaryloxime ethers [397], biphenyls, and 1-aryl-4,6-biscarboxydibenzofurans [215] (~20, 30 and 40%, respectively). The remaining 10% are based on 2-phenylbenzoxazole [378], biphenylamine and biphenylether substructures [395] or others [398]. The most potent and selective bisaryloxime ethers display poor chemical stability [397]. On the other hand, the biphenyl system has been referred to as one of the most promiscuous moieties in nature (personal communication by Bernard Testa, 2009, Porto) and several biphenyl derivatives have been associated with rodent toxicity and presumed toxicity in humans [219]. The dibenzofuran inhibitors and 2-phenylbenzoaxole inhibitors are therefore the most promising candidates identified to date and notable lead optimisation efforts have been taken [222,378]. The benzoxazole shown in **Figure 1.28**, in particular, is worth highlight. Initially referred to as FX-1006A, today this compound is better known as Tafamidis and represents the very first amyloid inhibitor to successfully complete all phases of clinical trials and reach the European drug market.

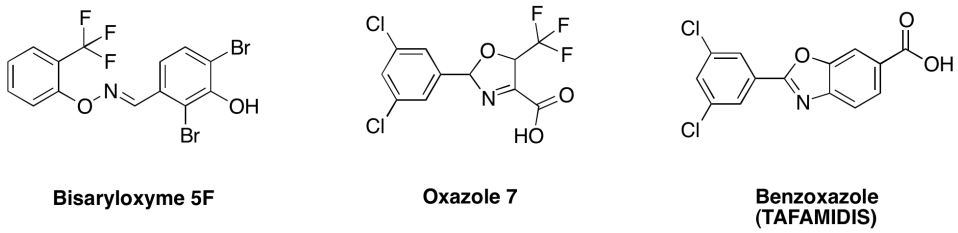


Figure 1.28. Chemical formulae of the latest and strongest TTR amyloid inhibitors rationally developed to date. The compound on the left represents a family of bisaryloxyme inhibitors affected by chemical instability; the compound in the middle represents a recent class of oxazole inhibitors; the compound on the right is Tafamidis, the first and sole anti-amyloid compound to have reached the drug market.

3.4.6. *Exotic* TTR stabilisers

Imagination knows no limits and over the last decade some unconventional approaches to TTR tetramer stabilization have been explored. Both Kelly's and Mark Pepys' groups

have reported interesting results obtained with palindromic[1] bivalent ligands that bind TTR prior to full tetramer assembly [399,400] (see **Figure 1.29**). These selective "super-stabilisers" were shown to be irreversibly trapped within the native TTR tetramer and, besides its unusually high molecular weight, display oral availability in mice. Curiously, according to Kolstoe et al. the presence of chlorine atoms in both mds84 and 4ajm15 is essential for the binding of the bivalent ligands by TTR [400].

Recently, a new family of designed stilbenes that covalently modify TTR has been presented [401]. When bound, these compounds chemoselectively react with the ε-amino group of Lys-15, yielding an amine bond. As stressed by the authors, however, in order to be translated into potency *in vivo*, the integration of covalent binding into the TTR stabilization mechanism requires ligands to be highly selective for TTR in plasma (which has been demonstrated), to react with TTR very rapidly comparative to its short half-life (which has also been demonstrated in vitro), and to exhibit a plasma distribution, concentration and half-life that sustain a near-quantitative conjugation yield (yet to be demonstrated) [401].

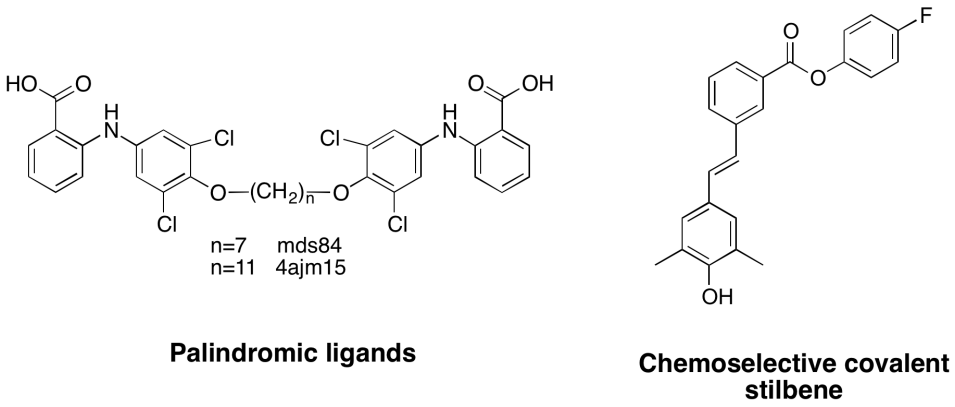


Figure 1.29. Chemical formulae of representatives of two compound families exploring unconventional mechanisms of binding to TTR tetramer. On the left, two bivalent ligands that bind TTR before its full tetrameric unit is assembled, filling both binding sites simultaneously. On the right, a designed stilbene that selectively and covalently modifies TTR in the human plasma.

4. Objectives of the project

Before July 2011, there were no approved dugs for the treatment of familial amyloid polyneuropathy (FAP) or familial amyloid cardiomiopathy (FAC) in the market. Until then, the only treatment of established efficacy was orthotopic liver transplantation. The

[1] The term "palindromic", deriving from "palindrome", refers to a word, phrase, number or other sequence of units that can be read in the same way in either direction.

results of phase II/III placebo-controlled clinical trials, released in late 2010, confirmed the therapeutic efficacy of tafamidis meglumine (FX-1006A), by halting the progression of FAP in approximately 60% of the treated patients. These results highlight the success of the followed strategy – the stabilization of the tetrameric form of transthyretin (TTR) through the non-covalent binding of small organic molecules – but disclose room for improvement. Tafamidis represents today the very first pharmacological solution for FAP, yet evidence on its efficacy and safety is still limited. The central motivation of this project was to identify new functional lead compounds, combining TTR amyloid inhibitory activity and adequate drug-like properties, which could open avenues for the development of alternative therapeutic solutions for this impairing disease.

From the viewpoint of the design of organic molecules to interfere with amyloid formation, TTR is both an appealing target and a challenging one. The elucidation and characterization of novel, structurally diverse and selective binders can also shade light on some of the most intriguing idiosyncrasies surrounding ligand-binding to TTR that remain poorly understood: first, the negative cooperativity observed for most known binders and, second, its propensity to bind biphenyls and molecules bearing multiple halogen atoms.

The availability of structural data of TTR bound to a number of ligands is extensive, with more than 60 complexes of human TTR at 2 Å resolution (or better) deposited in the Protein Data Bank (as of July 2010). This strongly encourages the exploitation of structure-based virtual screening (SBVS) methods like molecular docking and receptor-site pharmacophore models to identify compounds establishing optimal interactions within the receptor sites of TTR. However, ligand-based VS (LBVS) approaches such as 2D fingerprint similarity searches, 3D shape-matching or 3D pharmacophore searches have provided good results, often outperforming molecular docking, and may be valuable to identify compounds with appropriate shape, chemical topology and electrostatic properties similar to those of reference TTR stabilisers, yet with potentially better affinities and/or ADME/Tox profiles.

As a means to reach the central end of discovering novel and safe TTR stabilisers, the key objectives of this project were to evaluate the performance of a wide battery of state-of-the-art *in silico* methods and, through the combination of multiple computational algorithms, to develop a technology for the virtual screening of amyloid inhibitors.

Chapter 2

Evaluation of docking methods against TTR: where quality meets accuracy

"All that glisters is not gold."

[William Shakespeare, in *The Merchant of Venice*]

1. Introduction and theory

Structure-based drug design (SBDD) has been actively sought ever since the emergence of structural biology. With the ever-growing availability of detailed three-dimensional structures of target proteins determined by X-ray crystallography and NMR, the design of highly selective drugs should become tractable. As illustrated in **Figure 2.1**, when the structure of a given target of pharmaceutical interest is available, a structure-based design approach is often pursued. Amongst SBDD methods, virtual screening based on molecular docking, i.e., high-throughput docking (HTD), is the most popular choice and many different programs and protocols are in use and active development [33,34,402–406]. Another promising approach that has been increasingly explored with some success is that of structure-based pharmacophore modelling and searches, wherein structural data of the target receptor sites are used to derive 3D representations (pharmacophore hypothesis) of a set of features (and their spatial relationships) defining the interactions with ligands. Models thus generated can be used to screen chemical databases and retrieve compounds that fit them.

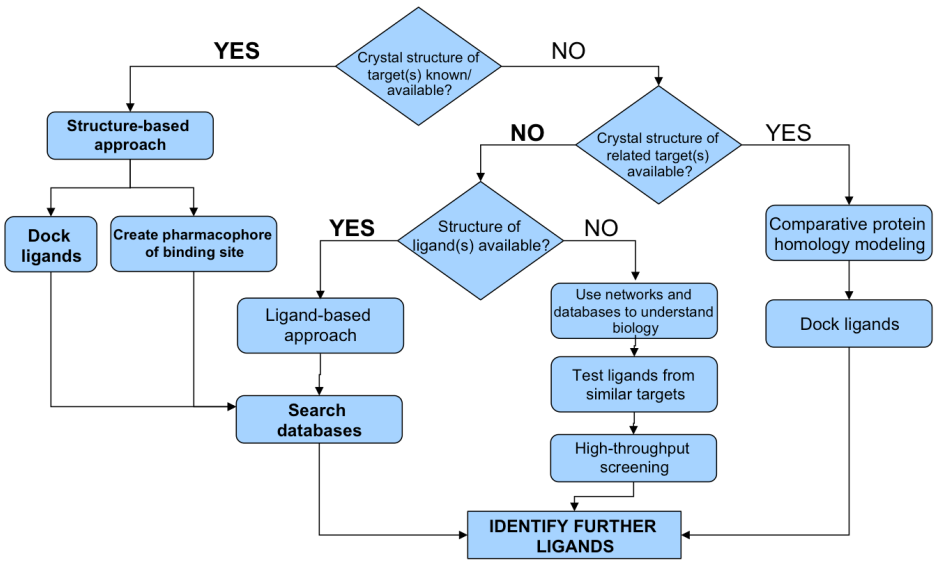


Figure 2.1. Rational in silico drug design workflow. In this Chapter, molecular docking is explored. In Chapter 3, receptor- and ligand-based pharmacophore approaches are considered and tested. This figure was adapted from reference [407].

Transthyretin (TTR) was extensively reviewed across the third section of the introductory chapter. This protein is implicated in the deposition of amyloid fibrils in the peripheral nerves and heart tissue. The formation of insoluble TTR fibrils in FAP and SSA involves an intermediate step whereby the native tetramer dissociates to monomers with

altered tertiary structure. This step is believed to be rate-limiting for fibril formation and can be modulated by the binding of thyroxine-like compounds to the two equivalent pockets of TTR. Over the last decade, several small molecules were screened for their ability to stabilise the protein and thereby prevent amyloid fibril formation, but the undesirable side effects associated with the discovered binders are still a major obstacle. In the present chapter, the application of molecular docking to the study of transthyretin-ligand interactions is explored.

The Protein Data Bank (PDB) is an important repository of biological macromolecular structures [408]. In 2000, Hornberg et al. compared all 23 TTR structures available in the PDB by then, including three wild-types, three non-amyloidogenic mutants, seven amyloidogenic mutants and nine complexes [409]. Besides putting forward a new high-quality TTR structure, this study demonstrated that the structural differences found amongst several X-ray structures of TTR variants, claimed to be of significance for amyloid formation, to be statistically non-significant. As of March 2008, 101 entries could be found for “Transthyretin" in the PDB, of which 42 corresponded to protein-ligand complexes. Thirty-five of these complexes have 2.0 Å resolution or better and, of these, 26 complexes correspond to the human form of the protein. The significant amount of structural data of TTR in complex with a number of ligands renders this protein an appealing target for structure-based design of small organic molecules to interfere with amyloid fibril formation. On the other hand, underneath this thrilling parade of structural information, several questions remain unanswered: Are there consistent structural differences between the two binding sites of TTR and across multiple TTR-ligand complexes? Does the available information account for the negative cooperativity phenomena observed for most binders?

In the work presented in this chapter, 30 high-quality X-ray structures of TTR in complex with ligands were selected upon structural quality evaluation and studied using a multiple structural alignment algorithm. The main goals of this effort were (*i*) to derive a logical annotation for the four TTR chains enclosing the two equivalent binding sites of each tetramer, based on their structural similarity, and (*ii*) to gain insights about the cooperativity effects observed for most TTR binder compounds. Perhaps more importantly, this study has also paved the way to the identification of the most appropriate docking programs to handle TTR, by providing a common spatial reference for the comparison of docked poses on each TTR binding site. Thus, each cognate ligand was docked back into its respective TTR complex and binding site using five docking programs (AutoDock 4, AutoDock Vina, eHiTS, FRED, and GOLD) and their respective scoring functions. Moreover, every X-ray ligand was docked into every X-ray receptor available, allowing the identification of a set of TTR structures that perform better against a variety of ligands.

The results presented in this chapter are also at the foundation of the work presented in Chapters 5, 6 and 7 of this thesis, where several docking-based virtual screening protocols are respectively evaluated against TTR, against multiple targets of pharmaceutical interest, and utilised for large-scale virtual screening campaigns.

1.1. Structural information on TTR

Hornberg et al. offered a remarkable analysis of 23 relevant TTR structures available as of mid-2000 [409]. More than ten years later, there are more than 110 PDB entries for TTR complexes (as of March 2011). The vast majority of these structures (more than 80) were resolved from orthorhombic crystals (mostly in the P 2_1 2_1 2 symmetry space group), where the asymmetric unit comprises only half of the full functional structure of the native TTR tetramer (i.e. a dimeric unit). Only a small portion of the complexes available in the PDB was determined from mono-clinic crystals (P 2, P 2_1 or C 2 space groups), and mostly at lower resolutions. Oddly enough, this aspect alone hinders the structural analysis of TTR complexes, and limits our understanding of the subtle structural rearrangements happening in the binding sites upon ligand binding.

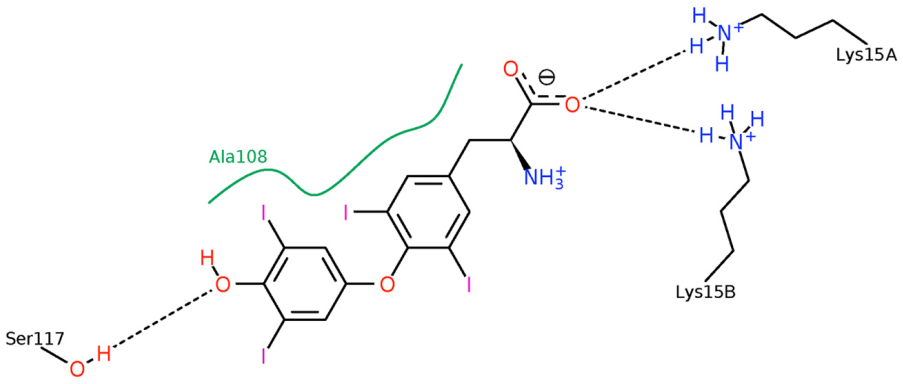


Figure 2.2. Two-dimensional structure-diagram for a TTR-T4 complex (PDB entry 2rox), showing hydrogen bonds and polar interactions in dashed black lines, and hydrophobic/apolar interactions as green contour lines. The diagram was created with PoseView**.**

Given the symmetry of the native homotetrameric protein, the two binding sites of TTR are identical and each site itself has two-fold symmetry [410]. Each site can be subdivided into subtle sub-pockets (see **Figure 2.2** and **Figure 2.3**). Thyroxine is one of the few TTR binders that can mediate interactions with both Lys-15 and Glu-54 residues at the entrance of the channel, ion-paring its carboxyl and amino groups, respectively. The hydroxyl group at the inner aryl ring is buried deep in the binding sites, possibly establishing a hydrogen-bond network with a water molecule conserved in one of the sub-pockets and the hydroxyl groups of Ser-117 and Thr-119. A specific water molecule is indeed ob-

served in TTR complexes where the ligand holds a hydroxyl group at the *para* position of the inner aryl ring. In the majority of TTR structures, the hydroxyl groups of Ser-117 residues are pointing away from the binding site and towards the neighbouring serines. However, in these complexes the Ser-117 hydroxyls adopt two alternative (yet equally favourable) rotamer conformations that permit hydrogen bonding with the phenol of the ligands through the conserved water molecule. Through its four iodine atoms, thyroxine can also occupy four out of the five remaining sub-pockets on each site: the inner ring iodines fit identical hydrophobic sub-pockets lined with the methyl groups of Leu-17, Ala-108, Val-121 and the polymethylene side chain of Lys-15; the outer ring iodines occupy sub-pockets formed by the side chains of Ala-108, Leu-17 and the hydroxyl groups of Thr-119 and Ser-117. Interaction between one of the iodines and the backbone carbonyl oxygen of Ala-109 was suggested as an important feature, as close intermolecular contacts between oxygen and iodine atoms were observed in several X-ray structures [411].

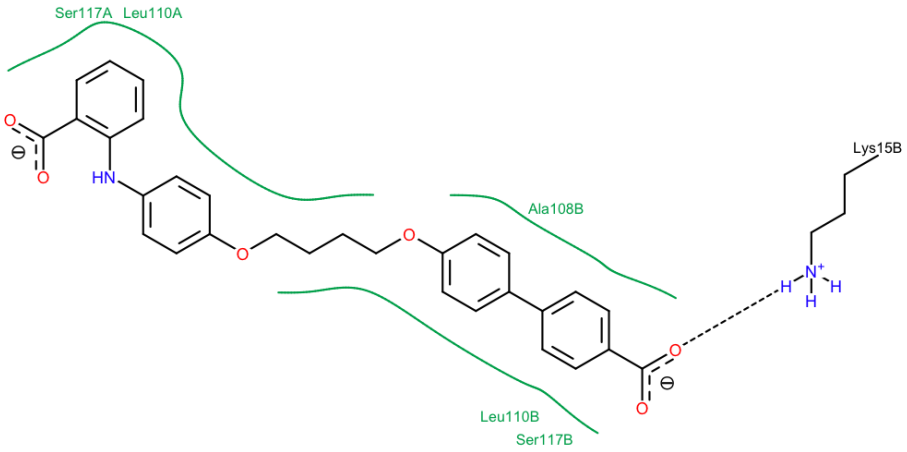


Figure 2.3. Two-dimensional structure-diagram for TTR in complex with a bivalent amyloid inhibitor. TTR's quaternary structure assembles around bivalent inhibitors when they interact with the protein before assembly occurs. Under this circumstance, strong inhibition of amyloid formation is produced. PDB accession code: 2fbr. The 2D structure diagram was generated using PoseView.

1.2. Structural alignment of TTR complexes using geometric hashing

As for sequence analysis in Bioinformatics, achieving a good alignment between two or more protein structures can be of critical importance for structural analysis in Structural Bioinformatics and Biology. For example, this procedure can allow a direct visualization of the impact of mutations on the structure of a protein. Furthermore, it can allow a de-

tailed comparison of the binding modes of dissimilar ligands to the same protein. In order to obtain a common spatial reference for the structural analyses carried out in this chapter and in Chapter 3, multiple structural alignment was performed with all 30 TTR complexes. As will be shown below, this procedure was also indispensable for the cross-docking studies reported in this chapter.

Brakoulias and Jackson developed a geometric matching method, known as GH8 [412], which constituted the basis for SitesBase (http://www.bioinformatics.leeds.ac.uk/sb), a database for structure-based protein-ligand binding site comparisons [413]. GH8 was extensively used throughout this project. It will be described in this subsection and mentioned further in Chapters 3 and 5. The algorithm works by detecting common atomic features through the identification of a maximum common clique, and producing an atom-to-atom similarity match for any two protein binding sites (see **Figure 2.4**). It proceeds by generating all possible atom triplets for each binding site, where each triplet consists of three atoms forming a triangle. A discretised 3D image of the reference binding site is created by storing atom identities at their nearest grid points on a 1 Å spaced grid. The atom triplet set of the reference binding site then acts as a template against which the atom triplet set of the query binding site can be compared. All possible atom triplets between the two structures are compared, and those with the same atom types at triangle vertices and similar interatomic distances are defined as a match. Corresponding triangle edges must differ by no more than 1.5 Å i.e. matched atoms must be within 0.5 Å of one another. On finding a match, a least-squares fitting routine is used to determine the translation and rotation matrix that maps the query atom triplet onto the reference atom triplet. The total number of atoms of the query binding site that are co-incident with the reference binding site, upon applying this transformation matrix, are counted ('atom-atom score'). This is then repeated for all possible atom triplets, and the transformation that results in the greatest number of co-incident atoms (i.e. the highest atom-atom score) is selected to perform the structural alignment of the query binding site onto the reference binding site.

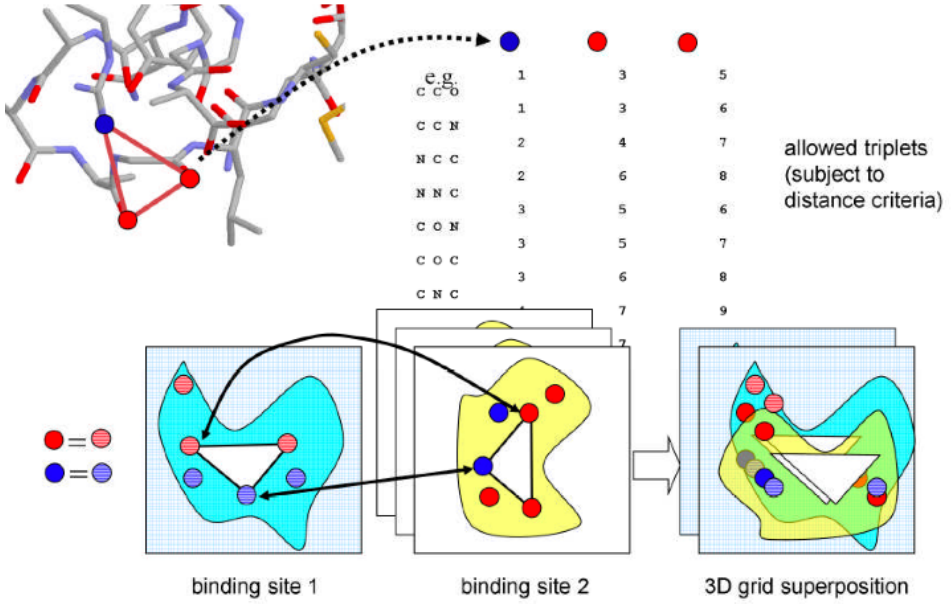


Figure 2.4. Schematic representation of the methodology used by GH8, a geometric hashing algorithm. A list of all atom triplets from each binding site is generated and each triplet from the query is rotated and translated for superimposition onto the reference. Matching triplets have the same atom types and inter-atomic distances. The maximal clique is determined by identifying multiple triplets that share the same transformation matrix.

1.3. Docking and Scoring

Generally, molecular docking uses the three-dimensional structure of the target protein to search for plausible binding modes for a putative ligand in a virtual protein-ligand complex (docking). It then compares different binding modes to predict the most favourable, and hence attempts to estimate the ligand's binding affinity (scoring). This procedure can be applied to rank each ligand in a sizeable compound library and thus filter out those that are less likely to binding to the protein and exert a beneficial function. The main advantage of docking-based VS methods over ligand-based methods is its search specificity, deriving from the explicit use of the structural information on the target receptor. This is a critical aspect for target selectivity, and is also associated with higher search flexibility, which is important for the elucidation of new scaffolds.

There are important hurdles and challenges in the docking field, not only to the methodology itself but also to the critical evaluation and application of different methods to different protein targets [414–421]. The vast majority of the docking programs use rigid receptor structures, performing conformational sampling on the ligands. In many cases, however, the dynamics of the protein receptor site is important to ligand binding. Over the last decade, a few docking methods have been developed to allow a certain degree of receptor flexibility, and thus predict induced fit effects, either by using ensembles of

structures instead of a single conformation or by explicitly modelling flexibility [422–428]. Alternative approaches to account receptor flexibility integrate the use of more complex conformational search techniques, such as geometry-based sampling [429], molecular dynamics simulations [430,431], and other mixed techniques [432]. However, flexible docking approaches are too computer intensive to be used in virtual screening, and the increase in docking accuracy resulting from their use does not seem to compensate the significant additional burden in terms of computing costs. Indeed, some of the faster rigid-receptor docking approaches perform rather well at mimicking the native/experimental ligand pose in protein-ligand complexes of many targets, even though discarding the natural flexibility of the receptor structure [433].

While the generation of relevant binding geometries (the "docking problem") is believed to be solved in the case of rigid receptors, the recognition of the correct ligand pose and the assignment of an accurate binding affinity (the "scoring problem") are still major challenges to structure-based VS. Basically, this is related to the limited ability of the scoring functions to discriminate active ligands from inactive ones across different protein targets, which is a fundamental requirement of VS. In particular, binding sites mainly comprised of non-polar residues and lacking an adequate number and/or placement of hydrogen bonding groups or other specific pharmacophore features are frequently associated with poor predictions of binding affinities by docking scoring functions [415]. Multiple approaches and combinations of docking and scoring have been attempted, but it has become clear that a universal solution is still far from sight [89,434].

1.4. Docking programs

In this work, five molecular docking packages were selected for evaluation against TTR. The selection was primarily based on availability, but preference was given to open source solutions and packages distributed through free licensing to academics. AutoDock 4 and AutoDock Vina are quite unique in this respect, because they are available to any user free-of-charge and as open-source codes. At the time these studies were conducted, a free academic licensing scheme was still available for eHiTS. Therefore, both eHiTS and FRED were obtained through free academic licenses. In fact, GOLD was the only software escaping this premise, being tested under a pre-existing protocol between the University of Leeds and the Cambridge Crystallographic Data Centre.

The five docking programs use quite dissimilar approaches to solve the docking problem. The details of each algorithm are described across the following subsections, but **Table 2.1** summarizes their basic differences.

Table 2.1. Summary description of the docking programs evaluated for pose fidelity against TTR.

Program	AutoDock 4	AD Vina	eHiTS	FRED	GOLD
Docking algorithm	Lamarckian genetic algorithm	Gradient-based iterated local search	Incremental construction	Shape fitting	Genetic algorithm
Scoring function	Free Energy Function	Vina Scoring Function	eHiTS scoring function	-ChemGauss2 -ChemGauss3 -ChemScore -PLP -ScreenScore	-GoldScore -ChemScore -ASP
Main features	-Successful applications reported, including blind docking -Widely used -Slow	-Relatively new tool -Inherits AD4 philosophy -Very fast	-Includes predocking filtering tools -Used to be academic-free (not anymore) -Fast	-Consensus scoring -Requires good shape complementarity -Very fast	-Many examples of accurate binding mode prediction -Medium to low throughput

Based on the assumption of no previous experience on the part of the user of the software under evaluation, all programs were tested employing their default settings, thus providing baseline performance.

1.4.1. AutoDock 4

AutoDock has been developed by the laboratory of Arthur Olson and was first presented to the scientific community in 1990 [435]. It stands today as one of the most popular docking programs worldwide [436]. The program had the original goal of providing an automated procedure for studying the interaction of ligands with macromolecular targets. To accomplish that goal, it combines conformational search methods with a grid-based method for energy evaluation. Rapid energy evaluation is achieved by precalculating ligand-protein pairwise interaction energies through atomic affinity potentials for each atom type present in the ligand on a grid (see **Figure 2.5**). This procedure is carried out by the AutoGrid program and follows the principles of the Molecular Interaction Fields first described by Peter Goodford [39] (introduced in Chapter 1). The energetics of a given ligand conformation is then computed by tri-linear interpolation of affinity values of eight grid points surrounding each ligand atom. Electrostatic interactions are assessed independently but equally by interpolating the values of the electrostatic potential and multiplying by the charge on the atom.

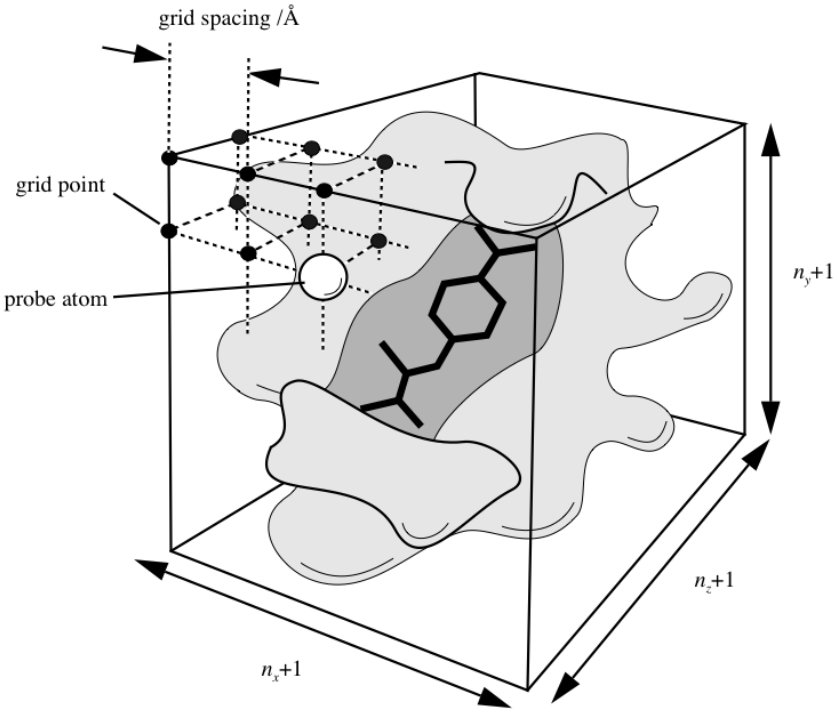


Figure 2.5. Schematic representation of the main features of a grid map. A grid map is composed of a 3D lattice of periodically spaced points centreed on a site of interest of the protein, encompassing it (partially or completely). Typically, the spacing between grid points varies between 0.2 and 1.0 Å, but the default value in AutoDock 4 is 0.375 Å (approximately a quarter of the length of a carbon-carbon single bond). The grid map points store the potential energy of *probe* atoms relative to all atoms in the protein.

The early versions of AutoDock made use of a Monte Carlo (MC) simulated annealing (SA) technique for ligand configuration space exploration, where global searches were performed at higher temperatures (allowing transitions over energy barriers) and local searches took place at lower temperatures (allowing a local optimisation on the current energy valley) [435,437]. With the protein held static during the simulation, the ligand performs a random walk in the space around the protein. At each MC iteration, a small random displacement is applied to all degrees of freedom of the ligand: translation of its centre of gravity; orientation; and rotation around each of its flexible internal dihedral angles. AutoDock 3.0 introduced the use of genetic algorithms (GAs) for global searching as an alternative to SA, combined with an *adaptive* local search (LS) method to perform energy minimization. The LS method can adapt the step size based upon the latest history of energies: a customizable number of successive failures (i.e. increases in energy) cause the step size to be doubled, whereas a customizable number of repeated successes (i.e. decreases in energy) cause the step size to be halved [34]. This hybrid and adaptive global-local search method performs better than SA and GA alone and is best known as the *Lamarckian genetic algorithm* (LGA).

The basic principles behind GAs were explained above in the introductory chapter. In AutoDock, the chromosomes are composed of strings storing the ligand's translation (in three Cartesian coordinates), orientation (in four variables defining a quaternion) and torsions (one real-value per torsion). The GA starts by creating a random population of individuals, whose size is controlled by the user. Each random individual is then given a random value between the minimum and the maximum x, y and z extents of the grid maps, producing a translation move; in turn, the four genes defining the orientation are given a random quaternion: a random unit vector and a random rotation angle between -180° and +180°. The random number generator is used in the LS, GA and LGA search engines. The creation of the random initial population is followed by a loop over generations, repeating until the maximum number of generations or the maximum number of energy evaluations is reached. A generation consists of five sequential stages:

1. Mapping and fitness evaluation – *mapping* decodes individual genotypes to their respective phenotypes, allowing *fitness* to be evaluated, which is the sum of the intermolecular interaction energy between the ligand and the protein, and the intramolecular interaction energy of the ligand;
2. Selection – individuals that have better-than-average fitness receive more offspring (proportionally);
3. Crossover – two-point crossover is applied, with breaks occurring only between genes, not within one gene;
4. Mutation – a random real number that follows a Cauchy distribution is added to the variables defining translational, orientational and torsional genes;
5. Elitist selection – determines how many of the top individuals automatically survive into the next generation (this is an optional, user-defined parameter).

The GA iterates over generations until one of the termination criteria is met. Most GAs imitate the key features of Darwinian evolution and employ Mendelian genetics. This is represented on the right side of the diagram shown in **Figure 2.6**, where a one-way transfer of information from the genotype to the phenotype is evident. By contrast, an inverse mapping function (genotypes deriving from phenotypes) would allow finishing a local search by replacing the individual with the result of the local search (left side of **Figure 2.6**). In the case of molecular docking the local search is carried out by constant conversion of the genotype into the phenotype, rendering the use of inverse mapping unnecessary. Still, the genotype of the parent is swapped with the resulting genotype in agreement with the Lamarckian principles.

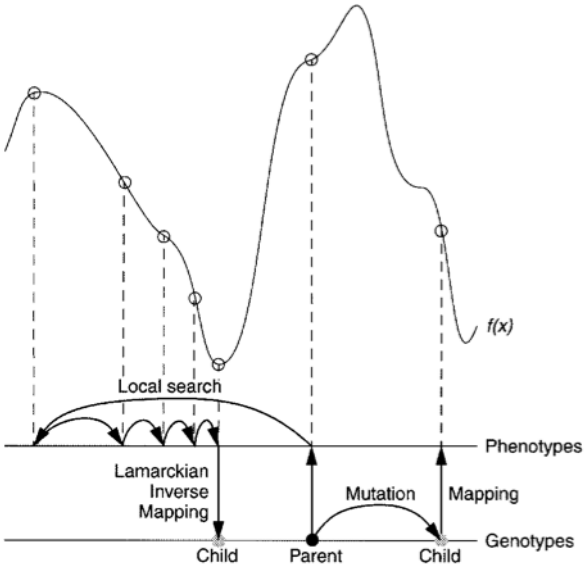


Figure 2.6. Illustration of the differences between Darwinian and Lamarckian searches, showing both the genotypic and phenotypic spaces. Genotypes are mapped to phenotypes by a developmental mapping function. *f*(x) represents the fitness function. On the right-hand side, the result of applying the genotypic mutation operator to the parent's genotype is shown, along with the respective phenotype. The left-hand side of the diagram illustrates a local search performed in phenotypic space employing knowledge on the fitness landscape. Extracted from reference [34].

AutoDock 3.0 and later versions use a five-term force field-based scoring function based on the AMBER force field and on a comprehensive thermodynamic model that allows incorporation of intramolecular energies into the predicted free energy of binding. This so-called Free Energy Function (FEF) comprises a Lennard-Jones 12-6 dispersion term, a directional 12-10 hydrogen term, a Coulombic electrostatic potential, an entropic term, and an intermolecular pairwise desolvation term (see Equation 2.1).

$$\begin{aligned}\Delta G = {} & \Delta G_{\mathrm{vdW}} \sum_{i,j} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right) \\ & + \Delta G_{\mathrm{hbond}} \sum_{i,j} E(t) \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} \right) \\ & + \Delta G_{\mathrm{elec}} \sum_{i,j} \frac{q_i q_j}{\varepsilon(r_{ij}) r_{ij}} \\ & + \Delta G_{\mathrm{tor}} N_{\mathrm{tor}} \\ & + \Delta G_{\mathrm{sol}} \sum_{i,j} (S_i V_j + S_j V_i) \mathrm{e}^{(-r_{ij}^2/2\sigma^2)}\end{aligned}$$

Equation 2.1

The scaling factor of each of the five terms in FEF was empirically calibrated using several protein-inhibitor complexes for which both structure and inhibition constants (K_i) were known. For a detailed description of how FEF was derived and parameterized, I refer the reader to reference [34].

The main advances in AutoDock 4 (AD4) in respect to its predecessor versions are (*i*) the ability to account for receptor-site flexibility and (*ii*) the introduction of a new method for covalent docking. Receptor flexibility is particularly important in many protein-ligand interactions, but handled in a rather limited way in AD4: the user selects specific side-chains in the protein that are then moved by rotation around torsional degrees of freedom, yet applying the exact same methods used for the sampling of flexible ligand's conformational space [438]. Although this new mechanism may be effective in some cases, it also poses important problems, such as the extra computational burden in the calculation of the receptor energy (a full pairwise energy evaluation must be carried out for the flexible regions of the protein) and a higher potential for yielding false positives (due to the larger conformational space).

1.4.2. AutoDock Vina

AutoDock Vina operates in a very similar way to AutoDock 4. However, and even though their search algorithms and scoring functions share similarities in overall form, their implementations are distinct.

In terms of the search algorithm, both programs employ hybrid global-local search; the key difference is in the local optimisation. While the local search in AD4 uses small random steps to seek more favourable conformations, Vina makes use of derivatives of the scoring function with respect to its arguments (position, orientation and torsions of the ligand), altogether composing a gradient, and performs optimisation accordingly. Interestingly, coupled with a more efficient use of CPU resources through multithreading, this difference renders Vina calculations faster than those performed by AD4 by orders of magnitude, particularly when handling large ligands holding several rotatable bonds. Moreover, although Vina does not offer a cluster analysis of all generated ligand poses (a helpful feature in AD4), the authors have reported superior performance over AD4 reproducing experimentally observed binding modes as the number of rotatable bonds increases [439].

The scoring functions of AD4 and Vina are both empirically weighted, enclosing similar terms for values such as hydrogen bonding and rotatable bond penalties. However, they have been calibrated differently. The Vina Scoring Function (Vina SF) attempts to merge

the advantages of knowledge-based potentials and empirical scoring functions, by mining the empirical information from both conformational preferences found in receptor-ligand complexes and experimental affinity data [439].

To sum up, **Figure 2.7** illustrates the key differences between AD4 and Vina.

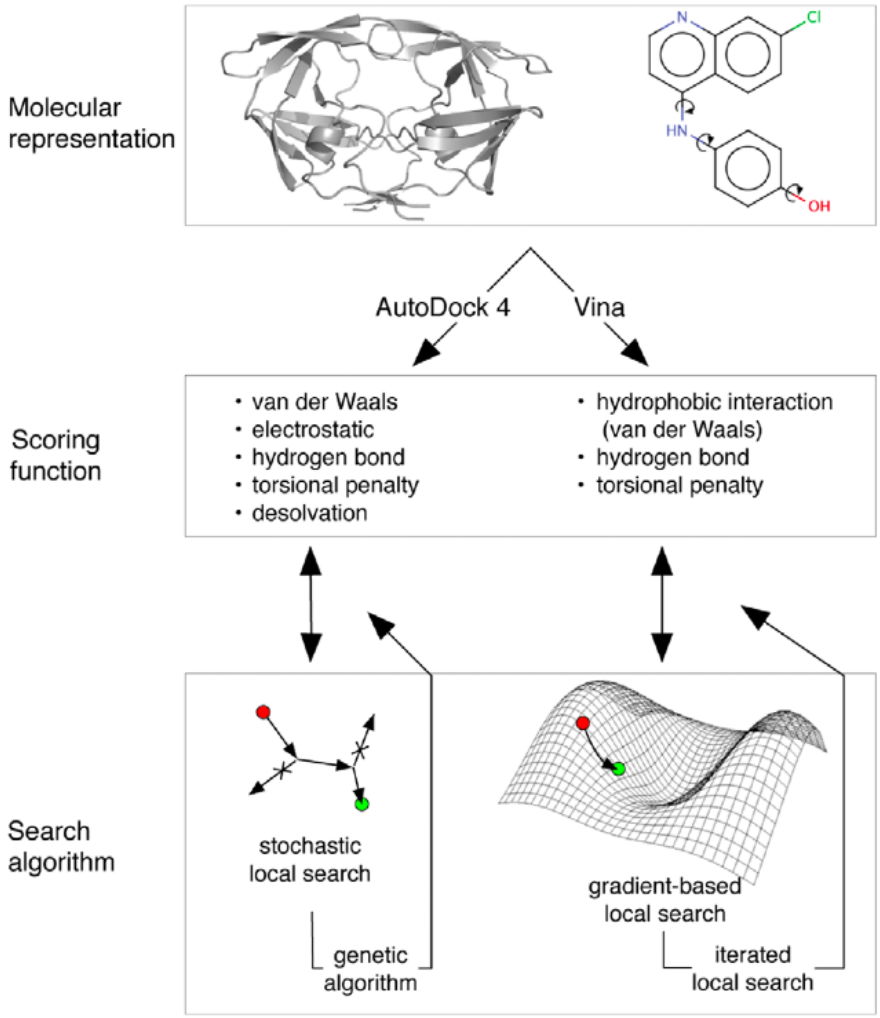


Figure 2.7. AutoDock 4 versus AutoDock Vina: a comparison of both methodologies. Extracted from reference [440].

1.4.3. eHiTS

Originally developed at the University of Leeds, eHiTS is an exhaustive docking program that employs the so-called *divide and conquer* approach to map the interaction of ligands within the binding sites of proteins. Firstly, a cavity description of the binding site is built, consisting of thousands of polyhedra. The ligand is then broken down into rigid fragments and connecting flexible linkers. Each rigid fragment is independently docked into

all possible places in the cavity. A clique detection algorithm is then used to carry out exhaustive matching of compatible rigid fragment pose-sets, resulting in anywhere from a few hundred to several million combinations of acceptable poses. The flexible linkers are then fitted to the specific rigid fragment poses that comprise a matching pose-set. The resulting reconstructed solutions represent approximate binding poses. Finally, a local energy minimisation, driven by the scoring function, is used to refine these solutions within the binding site. **Figure 2.8** illustrates the methodology employed by eHiTS.

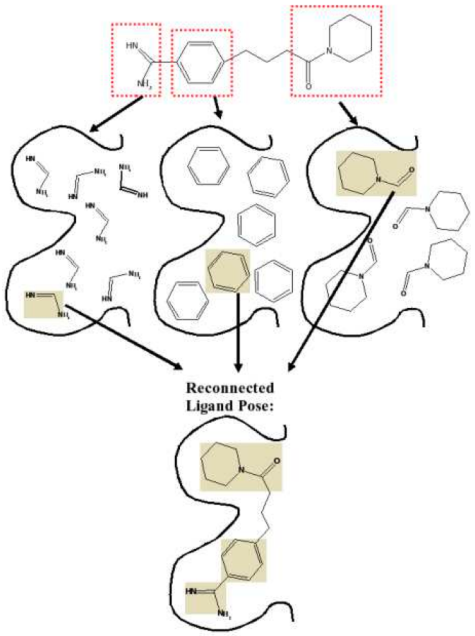


Figure 2.8. Schematic representation of the method used by eHiTS for ligand docking.

Three different scorings are used by eHiTS throughout the docking procedure. The initial scoring stage uses a simple surface map-based statistical function, whereby molecular surface contacts between the ligand and receptor are counted and scored based on the compatibility of chemical features (e.g. hydrogen bond donor, acceptor, hydrophilic, metal). During the optimisation phase of the eHiTS process, a more sensitive, empirical scoring function is employed. It consists of the following terms; hydrophobicity, aromatic π-stacking, electrostatic potential, van der Waals contact energy, metal ion interactions, and a penalty for incompatible contacts. The final binding poses are evaluated by a third, more time-consuming scoring function, which attempts to provide a more accurate estimation of the binding free energy for ranking the generated solutions. It is similar to the empirical scoring function except that it includes additional grid-based geometrical terms determined by the distance of the surface points from one another and the angle of the surface normal vectors, as well as the following additional terms: total surface contact area between the receptor and ligand, hydrophobic surface area of the receptor that is not buried by the ligand, exposed hydrophobic surface area of the ligand, and sum of

pocket depth values for all ligand atoms. The statistical score, all empirical terms, and the additional terms from the third scoring function are combined with an adjustable set of weights to produce the final energy score. The default weights provided by eHiTS have been calibrated using 133 receptor-ligand complexes for which both high-resolution X-ray crystal structures and corresponding experimental binding energy values were publicly available [402,441].

1.4.4. FRED

FRED (Fast Rigid Exhaustive Docking) is a shape-based protein-ligand docking program distributed by OpenEye Scientific Software, which takes a multi-conformer library/database – often generated by OpenEye's OMEGA [442] – and a receptor structure as input and outputs molecules of the input database most likely to bind to the receptor. A schematic of the FRED docking process is shown in **Figure 2.9**. FRED docks molecules using an exhaustive search algorithm throughout the following steps:

1. Enumeration of all possible poses by systematically searching rotations and translations of each conformer of the ligand within the active site at a specified resolution.
2. Filtering of the pose ensemble by discarding poses that do not fit within the larger of two volumes specified by a customisable receptor shape potential grid and a contour level.
3. Filtering of the pose ensemble by discarding poses that do not hold at least one heavy atom within the smaller of the two volumes specified by the receptor shape potential grid and a contour level.
4. Ranking of all surviving poses using one of the following scoring functions: Shapegauss, PLP, Chemgauss2, Chemgauss3 or CGO.

Following the exhaustive search, the 100 top scoring poses are subject to systematic solid body optimisation (a local exhaustive search at a finer resolution than the global exhaustive search). This step can be carried out using one of the abovementioned scoring functions and, additionally, Chemscore, OEChemscore, Screenscore or CGT. Finally, the best scoring pose is used to rank the ligand against other ligands in the screening database. This can be accomplished by one or several of the abovementioned scoring functions, by employing weighting schemes or consensus scoring. During the docking process the protein is held rigid, as are the conformers of the ligand. Ligand flexibility is implicitly included by docking the previously generated conformer ensemble of each molecule.

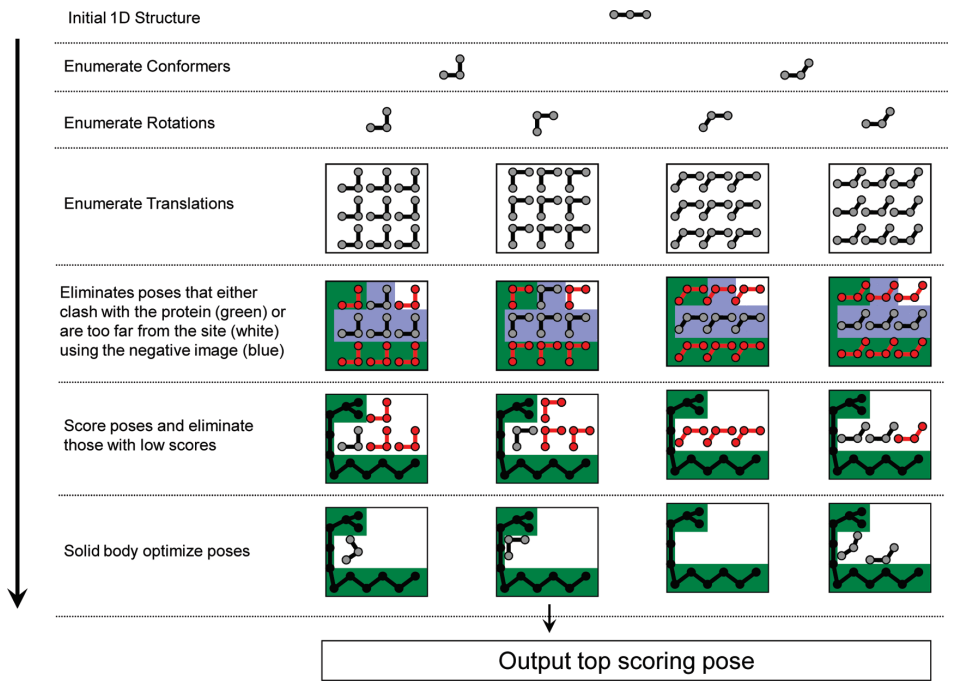


Figure 2.9. Schematic representation of the FRED docking process. Extracted from reference [443].

1.4.5. GOLD

GOLD (Genetic Optimisation for Ligand Docking) uses a genetic algorithm (GA) to explore full ligand flexibility, along with rotational flexibility for receptor polar hydrogens. GOLD features three scoring functions to evaluate and rank generated solutions: the force field based GoldScore, the regression-based ChemScore and ASP. In addition, there is an option that allows using a user-specific scoring function. The GOLD scoring function, GoldScore, uses a set of empirical parameters from a modifiable parameter file. Some correlation was found with experimental binding affinities [444], although it was originally designed for optimal pose-selection [445]. The function uses a 6-12 Lennard-Jones potential for the intramolecular (internal) vdW score and a "soft" 4-8 potential for the intermolecular (external) score. Equation 2.2 shows the components of GoldScore as output in GOLD.

$$\text{GoldScore} = 1.375 \cdot \text{vdW}_{\text{external}} + \text{HB}_{\text{external}} + \text{SE}_{\text{internal}}$$

Equation 2.2

$\text{vdW}_{\text{external}}$ is the intermolecular van der Waals score of the protein-ligand complex, $\text{HB}_{\text{external}}$ is the intermolecular hydrogen bonding score of the protein-ligand complex, $\text{SE}_{\text{internal}}$

is the ligand strain energy (vdW and torsional energy). The vdW term is multiplied by an empirical weight factor of 1.375 to promote hydrophobic protein-ligand contacts.

The GOLD implementation of Chemscore ($Chemscore_{GOLD}$) differs from the original scoring function [91], since a clash penalty (P_{clash}) and an internal torsion term ($P_{internal}$) are added to the final score to penalise bad contacts and poor conformations. Equation 2.3 shows the components of Chemscore as implemented in GOLD.

$$Chemscore_{GOLD} = Chemscore + P_{clash} + P_{internal}$$

Equation 2.3

1.5. Docking accuracy evaluation

Docking techniques are commonly evaluated using one or more of the following criteria: (*i*) their ability to reproduce the correct bound conformation of the ligand (often referred to as "pose-prediction" or "pose fidelity"), (*ii*) their ability to assign scores that correlate well with the binding affinities measured for known ligands, and (*iii*) their ability to discriminate active molecules seeded in a data set of decoys ("enrichment"). While existing methods tend to predict correct binding poses and often provide reasonable enrichment, docking scores notoriously show poor correlation with experimentally-determined binding affinities across compounds series. This limitation may be linked to the simplicity of the scoring functions, which are mostly optimised for high throughput analyses [80].

The majority of docking failures is due to the inability of the scoring function to reliably rank optimal "native-like" ligand conformations above "non-native" orientations. Thus, although the correct binding mode can be retrieved in most cases, assigning the lowest energy score to the correct binding pose has proven to be more challenging. Over the years a wide range of studies have been conducted to compare different docking tools, and from their outcome it is clear that there is not one docking tool that outperforms all others over a wide range of targets [415,416,446–449]. Indeed, the success of docking scoring functions has been shown to be highly target dependent [89]. Target dependency is an issue for docking because ligand binding can either be dominated by enthalpic or entropic contributions, the latter of which is generally poorly represented by scoring functions. The quest for a scoring function that is both highly accurate and generally applicable is ongoing.

In this chapter, the docking accuracy of selected docking programs is evaluated for pose fidelity and binding affinity prediction, through the study of their ability to reproduce the experimental ligand poses found in TTR X-ray complexes – a study often referred to as

"redocking" or "self-docking". This evaluation is critical to any structure-based design studies. The ability of docking scoring functions to discriminate actives from inactives is assessed later on (in Chapter 5) under the scope of the evaluation of programs for the virtual screening of new TTR amyloid inhibitors.

2. Computational methods

In the following subsections I provide details about the computational methods used to evaluate the structural quality of TTR complexes available in the PDB and to dock small molecules into TTR binding sites.

2.1. Structural quality analysis

Structures deposited in the Protein Data Bank (PDB) are the result of experimental work. Therefore, they are prone to experimental errors. Errors in PDB files span from violations of nomenclature, through slight inaccuracies in bond geometry, wrong side chain rotamers or badly modelled loops, all the way to gross errors that may render some of them inadequate for molecular modelling projects. In this work, we used both the WHAT_CHECK module [450] of WHAT IF [451] and PROCHECK [452] to evaluate the structural quality of 50 TTR high-resolution X-ray structures deposited in the PDB and thus select 30 complexes for docking evaluation (see **Table 2.2**). These programs provide indications of how likely certain residue geometries in a structure/model are compared to a set of high quality structures. They yield detailed reports on structural quality validation for each input molecule. The reports include Ramachandran plots and the secondary structure as described by DSSP, along with the description of problems that are detrimental to structural quality and to structural quality validation itself (e.g. space group problems, missing atoms, etc.).

The selection of TTR structures for redocking and crossdocking studies took the following three main criteria into consideration:

- Resolution: only structures with resolution ≤ 2.00 Å were included;
- R-factor: only structures with an R-value ≤ 0.25 were included;
- WHAT_CHECK structural quality indicators: only structures displaying appropriate structure Z-scores (first and second generation packing quality, Ramachandran plot appearance, chi-1/chi-2 rotamer quality and backbone conformation)

and RMS Z-scores (bond lengths and angles, omega angle restraints, side chain planarity, improper dihedral distribution and B-factor distribution).

Table 2.2 Thirty X-ray complexes of transthyretin used in redocking and cross-docking studies.

PDB code	Resolution (Å)	R-value	Description
1bm7	**2.00**	**0.189 (work)**	**Transthyretin in complex with flufenamic acid**
1dvt	1.90	0.197 (work)	Transthyretin in complex with flurbiprofen
1dvu	1.90	0.198 (work)	Transthyretin in complex with dibenzofuran-4,6-dicarboxylic acid
1dvy	1.90	0.202 (work)	Transthyretin in complex with phenoxazine-4,6-dicarboxylic acid
1dvz	1.90	0.192 (work)	Transthyretin in complex with o-trifluoromethylphenyl anthranilic acid
1e4h	1.80	0.193 (work)	Transthyretin in complex with pentabromophenol
1e5a	1.80	0.214 (obs.)	Transthyretin in complex with 2,4,6-tribromophenol
1eta	1.70	0.184 (obs.)	Transthyretin in complex with 3,5,3',5'-tetraiodo-l-thyronine (T4)
1etb	1.70	0.163 (obs.)	Transthyretin in complex with thyroxine
1f86	1.10	0.140 (obs.)	Transthyretin (t119m variant) in complex with 4-(2,6-diiodo-4-methylphenoxy)-2,6-diiodophenol
1tlm	1.90	0.173 (obs.)	Transthyretin in complex with milrinone
1tt6	1.90	0.188 (obs.)	Transthyretin in complex with diethylstilbestrol
1tz8[a]	1.85	0.198 (obs.)	Transthyretin in complex with diethylstilbestrol (monoclinic crystal)
1y1d	1.70	0.198 (obs.)	Transthyretin in complex with iododiflunisal
2b15	1.70	0.190 (obs.)	Transthyretin in complex with 2,4-dinitrophenol
2b16	1.75	0.191 (obs.)	Transthyretin (y78f variant) in complex with 2,4-dinitrophenol
2b9a	1.54	0.231 (obs.)	Transthyretin in complex with 3',5'-difluorobiphenyl-4-carboxylic acid (diflunisal analogue)
2f7i	1.60	0.234 (obs.)	Transthyretin in complex with 2',6'-difluorobiphenyl-4-carboxylic acid (diflunisal analogue)
2f8i	**1.54**	**0.217 (work)**	**Transthyretin in complex with a benzoxazole derivative**
2fbr[a]	1.46	0.211 (obs.)	Transthyretin in complex with a bivalent ligand
2g5u	**1.80**	**0.180 (work)**	**Transthyretin in complex with 3,5,3',5'-tetrachloro-biphenyl-4,4'-diol**
2g9k	1.85	0.216 (obs.)	Transthyretin in complex with 2',3,3',4',5-pentachloro-biphenyl-4-ol
2gab	1.85	0.211 (obs.)	Transthyretin in complex with 3,3',4',5-tetrachloro-biphenyl-4-ol
2rox	2.00	0.170 (work)	Transthyretin in complex with thyroxine
1dvx[b]	2.00	0.192 (work)	Transthyretin in complex with diclofenac
2b77[b]	1.70	0.210 (obs.)	Transthyretin in complex with a diflunisal analogue
2qgc[b]	1.30	0.154 (work)	Transthyretin in complex with 2-(3,5-dimethyl-4-hydroxyphenyl)benzoxazole
2qgd[b]	1.50	0.162 (obs.)	Transthyretin in complex with 2-(3,5-dibromo-4-hydroxyphenyl)benzoxazole
2qge[b]	1.45	0.166 (obs.)	Transthyretin in complex with 2-(3,5-dimethylphenyl)benzoxazole
3b56[b]	1.55	0.190 (obs.)	Transthyretin in complex with 3,5-diiodosalicylic acid

[a] Removed from the pool of structures in crossdocking studies. [b] Added to the pool of structures in crossdocking studies.

Because most TTR structures deposited in the PDB were determined in the P 2_1 2_1 2 space group and the asymmetric unit of all orthorhombic crystal lattices contains only the structure of one TTR dimer, the full biological units corresponding to the selected PDB entries were downloaded from the Protein Quaternary Structure (PQS) server [453].

2.2. Binding site and chain annotation via clustering

Clustering of similarity scores matrix resulting from multiple structural alignment of TTR binding sites was performed with CLUTO, a family of data clustering software tools [454,455]. CLUTO provides multiple classes of clustering methods such as partitioning, agglomerative and graph partitioning, and is useful for clustering low and high dimensional data sets. Amongst the clustering criterion functions available are single criterion functions and hybrid criterion functions. While the single criterion includes functions that tend to either maximise similarities within a cluster or minimize dissimilarities between clusters, the hybrid criterion simultaneously optimises multiple single criterions.

Here, both partitioning and agglomerative clustering algorithms were explored. The *direct* method is recommended over the *repeated bisection* method when the number of desired clusters is less than 10 [454]. Since clustering was performed to investigate the possibility of splitting the data set into two, corresponding to the two binding pockets of TTR, the *direct* partitioning method was used. When using the agglomerative algorithm, the recommended H2 clustering criterion was employed. The degree of agreement between the results from the two clustering algorithms was determined through the RAND index using the e1071 package in R [456].

2.3. Ligand energy calculations and conformational randomisation

Ligand structures of all complexes were verified visually, to ensure that no atom-atom clashes were present in the X-ray complexes. A calculation of conformational energies was performed – using the MMFF94 force-field [30] within SYBYL [457] – to ensure that no ligand structures had high strain energies.

Openeye's OMEGA version 2.2.1 [458] was used to randomise all ligand conformations prior to crossdocking studies (the lowest energy 3D conformations were retained). In OMEGA, molecular energies were calculated using the Merck Molecular Force-Field [30], and any conformations above a 10 kcal.mol^{-1} energy threshold (in relation to the lowest

energy state), or less than an 0.8 Å RMSD threshold (in relation to all other conformations), were discarded. This randomisation procedure removes any bias towards the initial positional and conformational information of the ligand. Therefore, it recreates the typical virtual screening setting more accurately, where the "bioactive" conformation of the molecules in the database is unknown.

2.4. Docking runs parameters

Each cognate ligand was individually redocked into its respective TTR binding site using AutoDock version 4.0.1 [34], AutoDock Vina version 1.1.2 [439], eHiTS version 6.2 [441], FRED version 2.2 [459] and GOLD version 3.2 [33]. With very few exceptions, the results reported over the following section were obtained using the *default* docking parameters, considered by the respective developers as an optimal setup for a good compromise between speed and accuracy. Given their convoluted structure, examples of AD4 and GOLD docking parameter files are given in Section A of the Appendix.

AutoGrid version 4.0.1 was used to generate atom-specific affinity maps, electrostatic and desolvation potential maps for AD4, employing the default grid spacing of 0.375 Å and a distance-dependent dielectric of -0.1465. AD4 calculations then were carried with the default Lamarckian genetic algorithm, involving populations of 150 individuals, a maximum number of 27,000 of generations and a maximum of 2,500,000 energy evaluations, over a total of 10 LGA-LS runs. As output, the top 10 best-ranked docking poses were retained.

Vina was run with the default exhaustiveness (for the global search) of 8, a value that is roughly proportional to the minimum time spent searching for a global minimum of the scoring function. The number of binding modes to generate was adjusted to 10, and the maximum energy difference between the best and the worst binding modes was kept at 3 $kcal.mol^{-1}$.

Like with AD4, all GOLD parameters affecting the performance of GA searches were kept as default (see an example in Section A of the Appendix). Automatic atom-type assignment was switched on for the ligands and the protein. Each ligand was docked ten times (default).

In eHiTS, the default *accuracy* parameter was changed from 3 to 4, in order to promote a more thorough search. The default *select* parameter was kept unchanged, resulting in the selection of a representative subset of 32 solution poses for each ligand, but the *toprank* parameter was set to 10, so that only the top 10 ranked solutions were reported.

In FRED, exhaustive scoring and solid body optimisation of the generated poses were carried out with Chemgauss3. The *num_poses* parameter, reflecting the number of poses to generate per ligand was kept as default (100), but the *num_alt_poses* parameter, was set to 10, corresponding to the number of alternate poses to be output. All FRED calculations were run after a preliminary step of ligand conformational search performed with OpenEye's OMEGA (generating up to 400 possible conformations per ligand).

Docking poses generated by AD4, Vina and eHiTS were scored using the scoring functions implemented as objective functions in the programs – FEF, Vina SF and eHiTS SF, respectively. Docking poses generated with GOLD were scored with GoldScore [33] and re-scored with ChemScore [91] and ASP [460]. Poses generated by FRED upon exhaustive scoring and optimisation with Chemgauss3 were scored with Chemgauss2 and Chemgauss3 [459], Chemscore [91], PLP and Screenscore [461].

Finally, it is worth noting that eHiTS and GOLD were the only programs allowing the explicit inclusion of crystallographic water molecules in the docking calculations in a straightforward manner. For the sake of a fair comparison, all results reported in the following section were obtained upon exclusion of water molecules. Docking calculations including crystallographic water molecules were performed with eHiTS and GOLD, but no visible gains in docking accuracy were detected, not even in complexes where water-mediated interactions appear to be relevant for ligand binding.

2.5. Quantitative comparison of docking accuracy

The accuracy of docking generated poses was inspected visually with VMD [462] and quantified by the root-mean-square deviation (RMSD) from experimentally determined ligand coordinates (Equation 2.4). All RMSD values were calculated with a script written in-house based on the *smart_rms* utility supplied with the GOLD package (CCDC). *Smart_rms* calculates the RMSD between two conformations of the same structure, taking into account ligand topology symmetry (to accommodate effects such as the flipping of a phenyl ring by 180 degrees). Using a graph isomorphism algorithm, an RMS score is calculated for each way of mapping the molecule onto itself.

$$RMSD = \sqrt{\frac{\sum_{i=1}^{n} d_i^2}{n}}$$

Equation 2.4

As explained above, the contents of the asymmetric unit of most TTR complexes require crystallographic symmetry operations to generate the complete macromolecule. How-

ever, these operations produce an unfortunate effect over the ligands, which sit along one of TTR's symmetry axes: for several ligands, principally the asymmetric ones, an alternative (and yet plausible) binding mode is inherited from the symmetry operations (see **Figure 2.10**). To handle this problem during RMSD calculations, the angles between the planes intersecting the ligands in their two binding modes were determined on each of the two TTR binding sites of every TTR complex downloaded from the PQS server. This was accomplished using the *Structure Measurements* panel of UCSF Chimera [463]. Based on the angle value distribution for ligands showing very similar symmetry-related binding modes (hence, non-alternative), such as the strongest TTR binder (PDB/PQS entry 2g5u), all binding modes showing a difference of 7 degrees or more between their respective intersecting planes were labelled as alternative. In these cases, two RMSD values were calculated against each of the alternative binding modes, and the RMSD value used for comparisons corresponds to the lower value found.

Exceptional cases requiring an even tighter control are the TTR complexes of flufenamic acid (PDB/PQS entry 1mb7), wherein 4 alternative binding modes had to be considered (2 symmetry-related binding modes times 2 alternative ligand conformations), and of a bivalent ligand (PDB/PQS entry 2fbr), where the 2 alternative binding modes resulting from symmetry operations apply to the 2 alternative binding modes across the entire TTR channel. In these cases, visual inspection of the docked poses became even more critical to the selection of RMSD values to use for comparison.

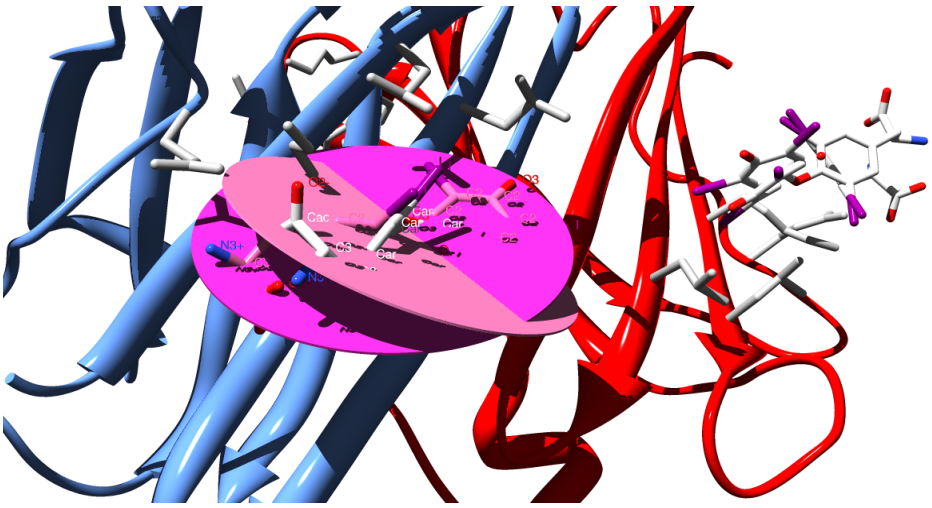


Figure 2.10. Crystal structure of the TTR:T_4 complex showing two alternative binding modes for T_4 resulting from the symmetry operations required to generate the complete TTR binding sites. The planes that intersect the ligand structures on each binding mode are displayed for one of the binding sites. In this case, these planes form an angle of about 24 degrees between each other. This figure was produced using UCSF Chimera [463].

3. Results and discussion

Structural quality analysis was followed by an exhaustive structural similarity analysis towards a logical systematization of TTR chain and binding site annotations. This was accomplished through the coupling of a geometric hashing algorithm with hierarchical clustering. Next, to identify a program capable of providing accurate predictions for the interactions between TTR and a multitude of chemical compounds, we tested five docking programs that employ one or several scoring functions.

3.1. Structural quality evaluation

All TTR complexes selected from the PDB were subjected to structural quality evaluation with WHAT_CHECK and PROCHECK. Sixteen complexes with near-zero or negative Z-scores for overall Ramachandran plot appearance were rejected; 4 complexes were rejected due to *poor (or worst)* chi1/chi2 rotamer normality of the binding site residues (as classified by WHAT_CHECK). Thirty complexes were retained. Overall, the quality of these 30 complexes is acceptable, with no more than 10 outliers, on average, populating their Ramachandran plots. Moreover, most of these outliers are located on the surface of the protein, thus not affecting binding site residues.

A more careful analysis of WHAT_CHECK reports was enforced for the best-performing complexes at docking accuracy evaluations (see the "Crossdocking studies" section below). The idea was to ensure that structures to be used for further computational studies (such as virtual screening) met most structural quality requirements that can be expectable. **Table 2.3** presents Z-scores reported by WHAT_CHECK for critical quality parameters, namely the Ramachandran plot appearance, chi1/chi2 rotamer normality, backbone conformation, side chain planarity, improper dihedral distribution and B-factor distribution. Of all parameters, the most upsetting results concern the chi1/chi2 rotamer normality, with all three best crossdocking performing structures (see subsection 3.4 of this chapter) scoring negative. However, it is worth mentioning that, although negative, the magnitude of the scores is small, meaning that, globally, the side chain conformations of structures differ only slightly from the global rotamer distribution within WHAT_CHECK database of known reliable structures. Results for the worst crossdocking performing structures are also given in **Table 2.3**, for comparison purposes. Interestingly, although the Ramachandran plot appearance parameter discloses higher quality for 1bm7 and 2g5u, overall, it was not possible to identify a quality indicator showing a clear correlation between structural quality results and docking accuracy results.

Table 2.3. Selected quality indicators from the WHAT_CHECK summary report for two TTR structures performing best and worst against a variety of ligands. WHAT_CHECK uses experimental evidence from known reliable structures for comparisons.

PDB entry	Structure Z-scores[a]			RMS Z-scores[b]		
	Ramachandran plot appearance	chi1/chi2 rotamer normality	Backbone conformation	Side chain planarity	Improper dihedral distribution	B-factor distribution
Crossdocking best-performing TTR structures						
1bm7	2.2	-0.5	1.3	0.882	1.257	0.888
2g5u	2.1	-0.2	1.8	0.779	0.946	0.521
Crossdocking worst-performing TTR structures						
1e4h	1.5	0.6	1.2	0.363 (tight)	0.910	0.588
1tlm	1.4	-1.0	0.2	1.534	1.809 (loose)	0.926

[a] Positive values mean better than average. [b] Values should be close to 1.0.

Figure A.1 of the Appendix reports the Ramachandran plots for the best performing TTR structures selected through the crossdocking analyses reported below.

3.2. Receptor site analysis: structural overlap and chain annotation

In order to derive logical annotations for the four TTR chains (and, thus, for the two binding sites they comprise), the 30 TTR complexes were first screened through an even tighter quality cut-off: only TTR structures holding resolution of 2.0 Å or better and R-factor of 0.2 or lower were retained. More importantly, redundancy was avoided through the selection of structures with highest quality from amongst a group of closely related structures, resulting in a total of 21 complexes. Multiple structural alignment of these 21 complexes was then performed with GH8. The similarity scores were calculated by comparison of all C_α atoms of the protein in an all-against-all basis. TTR in complex with 4-hydroxy-2',3,3',4',5-pentachlorobiphenyl (PDB entry 2g9k) was found to hold the highest similarity score amongst the data set (illustrated in **Figure A.2** of the Appendix). This complex was thus used as the reference template for the structural alignment of the 30 TTR complexes listed in **Table 2.2**. **Figure 2.11** illustrates the structural overlap resulting from this multiple structural alignment.

Chains [A,C] and [B,D] enclosing ligand binding sites of TTR, were segregated for the 21 selected complexes. Coordinates of TTR binding site residues (Glu 54, Lys 15, Leu 17, Thr 106, Ala 108, Ala 109, Leu 110, Ser 117, Thr 119 and Val 121) were extracted from their respective complexes and aligned with GH8. Since the differences between two binding pockets may be solely based on the arrangement of a few atoms in the structure, two dis-

tinct analyses were carried out for the extracted active site residues: (*i*) comprising only the C_α atoms and (*ii*) comprising all side chain atoms. Minor, yet systematic, structural differences between the two pocket domains were identified in the protein backbone across the complexes, using clustering (see **Figure A.4** of the Appendix). Conversely, no systematic (nor significant) differences were found for the residue side chains composing the binding sites. TTR binding site in complex with 2',6'-difluorobiphenyl-4-carboxylic acid (PDB entry 2f7i) was the highest scoring structure resulting from the all-atom side chain alignment, disclosing the highest overall similarity amongst all TTR binding site side chains (illustrated in **Figure A.3** of the Appendix). Visual analysis also revealed that the binding site of 2f7i presents well-defined side chain conformations for residues Ser-117 and Thr-119, with their side-chain hydroxyls pointing away from the binding site and possibly establishing hydrogen bonds with Ser-115 residues of the flanking monomer. In fact, this configuration is common to most TTR complexes, which can influence the outcomes of docking and scoring.

Guided by these results, the arrangements of chains A, B, C and D and their respective geometric centres in PDB entry 2g9k were inspected using VMD [462]. The above process was repeated one-by-one for the remaining 29 complexes and their arrangement was compared to that of 2g9k. Any complex that did not conform to the 2g9k chain layout had its chains appropriately renamed.

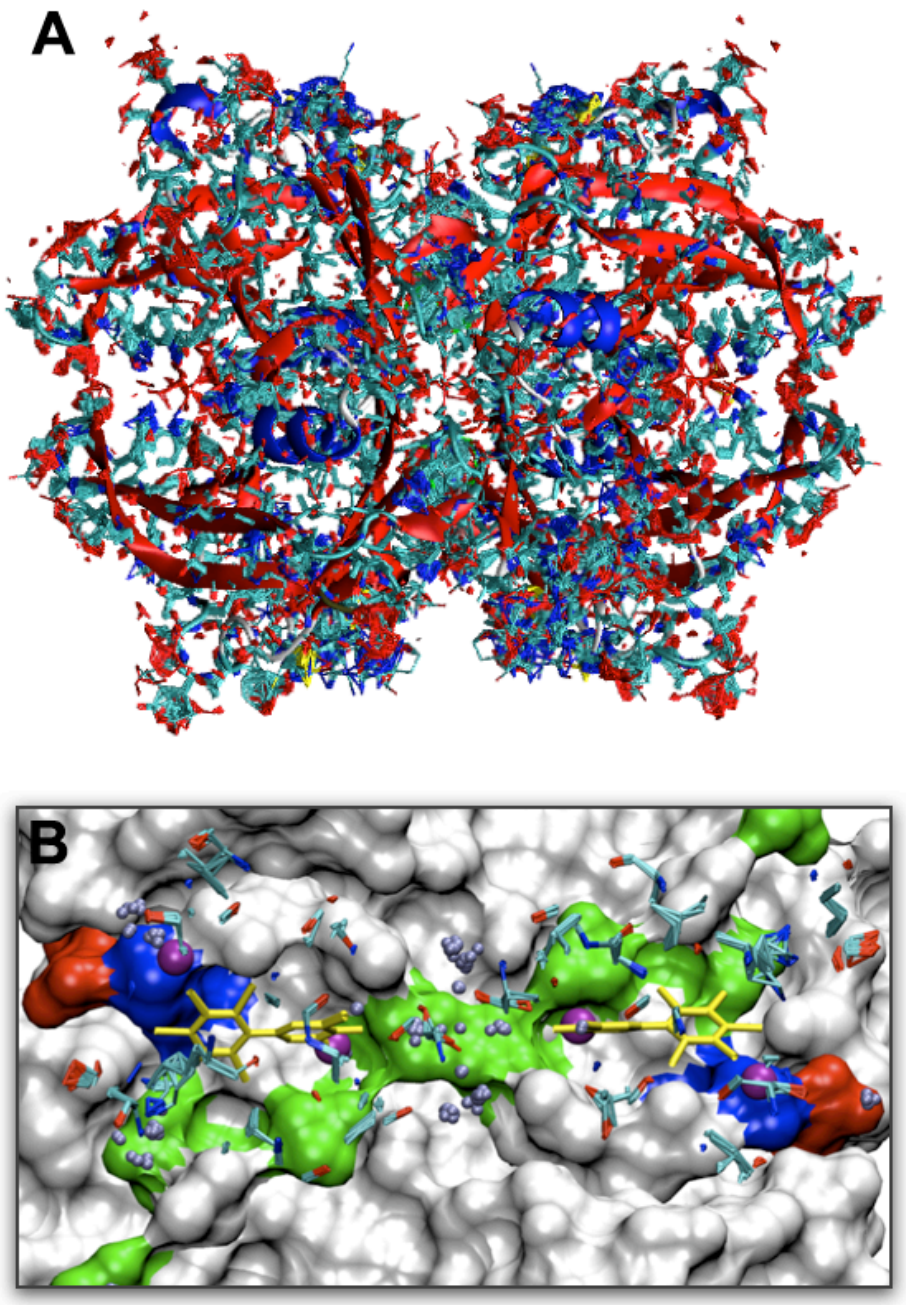


Figure 2.11. Structural overlap resulting from the multiple structural alignment of 30 TTR complexes performed with GH8. (A) Representation of the full tetramer assembly. (B) A view from TTR binding pockets, with one dimer at the bottom (far from the plane of the viewer's eye) represented by its molecular surface and the superimposed side chains at the top (closer to the plane of the viewer) represented in sticks. The molecular surface is coloured by residue type (white for apolar, green for polar, blue for positively charged and red for negatively charged residues). The sticks are coloured by atom type.

3.3. Redocking studies

The basic postulates of structure-guided molecular design emphasize the value of a thorough understanding of the structural aspects driving the association between two interacting molecules. Accordingly, the utility of a docking program in structure-based design is mainly determined by its ability to recognize ligand poses that are close to those revealed by experimentally determined complexes. PDB accession codes and details of the 24 TTR X-ray complexes used in the redocking studies can be found in **Table 2.2**. The native form of TTR is a tetramer comprised by four identical chains, whose assembly produces a central channel with two equivalent binding sites. As illustrated in **Figure 2.2** and **Figure 2.11**, these tunnel-shaped binding sites are mostly comprised of apolar residues, such as alanine, leucine and valine, featuring two lysine and two serine residues respectively at their external and internal ends. Besides its natural penchant for hydrophobic moieties and halogen bearing organic molecules, TTR displays cooperativity in ligand binding, i.e. the binding of a first molecule to one of the sites affects the affinity of binding for a second molecule. Even though the structural grounds of this effect remain elusive to date, given the full correspondence in amino acid composition between the two binding sites, it prompted us to study both sites individually. Thus, each cognate ligand was individually docked back into its respective TTR binding site using AD4, Vina, eHiTS, FRED and GOLD.

The results of redocking studies show that, in general, docking programs performed well at sampling near native poses, all of them being able to produce at least one ligand conformation whose root-mean-squared deviation (RMSD) is equal to or below 2.0 Å, amongst the top ten docking solutions. Programs using genetic algorithms for conformational sampling (namely, AD4/Vina and GOLD) performed marginally better than those using incremental construction (eHiTS) or shape fitting (FRED) algorithms (see an example in **Figure 2.12**). However, the true impact of the more efficient sampling of GAs was noticeable only in the case of a bivalent amyloid inhibitor (PDB code 2fbr), a palindromic ligand containing 12 rotatable bonds that fills the entire TTR channel, i.e. both binding sites simultaneously (see **Figure 2.3** and reference [400] for details). This suggests that, even though programs employing GAs tend to be slower than those using other approaches, especially when handling highly flexible ligands, the extra time cost (and CPU burden) seems to pay off and result in higher docking accuracy.

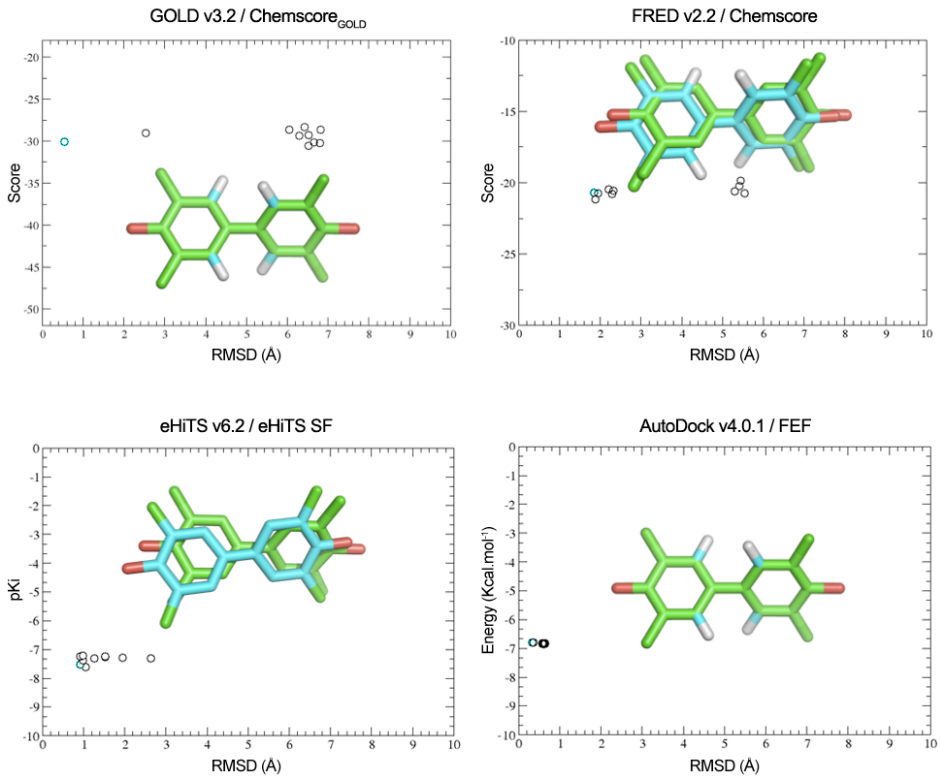


Figure 2.12. Score versus RMSD plots illustrating the pose fidelity of four docking programs while handling a polychlorinated biphenyl against TTR. Results obtained with GOLD v3.2, FRED v2.2, eHiTS v6.2 and AutoDock v4.0.1 are shown. For each program, the top 10 best-ranked poses are shown as circles in the plots. The pose holding the lowest RMSD is coloured in cyan and shown in licorice representation against the reference X-ray pose (coloured in green).

The real challenge to the docking programs was only uncovered when trying to identify native-like ligand poses through the adequate scoring of each predicted pose. As seen in **Figure 2.12**, little discrimination between correct (RMSD ≤ 2 Å) and wrong poses (RMSD > 2 Å) occurs in most cases. For example, while GOLD was able to sample a near native pose (RMSD = 0.55 Å, Score = -30.1), GOLD's implementation of Chemscore assigned an incorrect pose (RMSD = 6.53 Å) with the best score (Score = -30.6; see the top left pane). On the other hand, while some discrimination between correct and wrong poses is witnessed with FRED's implementation of Chemscore and the eHiTS scoring function, the difference between scores assigned to correct/wrong poses is small. Interestingly, in this case all top 10 ranked poses scored by AD4's FEF correspond to correct poses (with RMSDs < 1 Å). A more challenging example in terms of conformational sampling is provided in **Figure 2.14**.

Figure 2.13 summarizes the results of the 48 redocking operations performed with TTR complexes (24 complexes times 2 binding sites). Most scoring functions failed to identify the best poses from the pool of generated ligand conformations. More than 70% of the

highest ranked solutions had RMSD values above 2.0 Å in both TTR binding sites. Without doubt, the free energy function (FEF) implemented in AD4 outperformed all other docking and scoring schemes (with 63% accuracy), yet closely followed by Vina (56 % accuracy), revealing an acceptable discriminative power at pose prediction in both TTR binding sites (see **Figure 2.14** as well). Surprisingly, FRED and its built-in scoring functions presented the worst docking accuracy for TTR amongst all tested algorithms. Consensus scoring of the docked poses was attempted, employing several combinations of weighting/voting, but no visible improvements were witnessed. The results obtained with eHiTS and GOLD disclose a higher dependence on a correct placement of the ligands directed by hydrogen bonds and charge interactions, respectively.

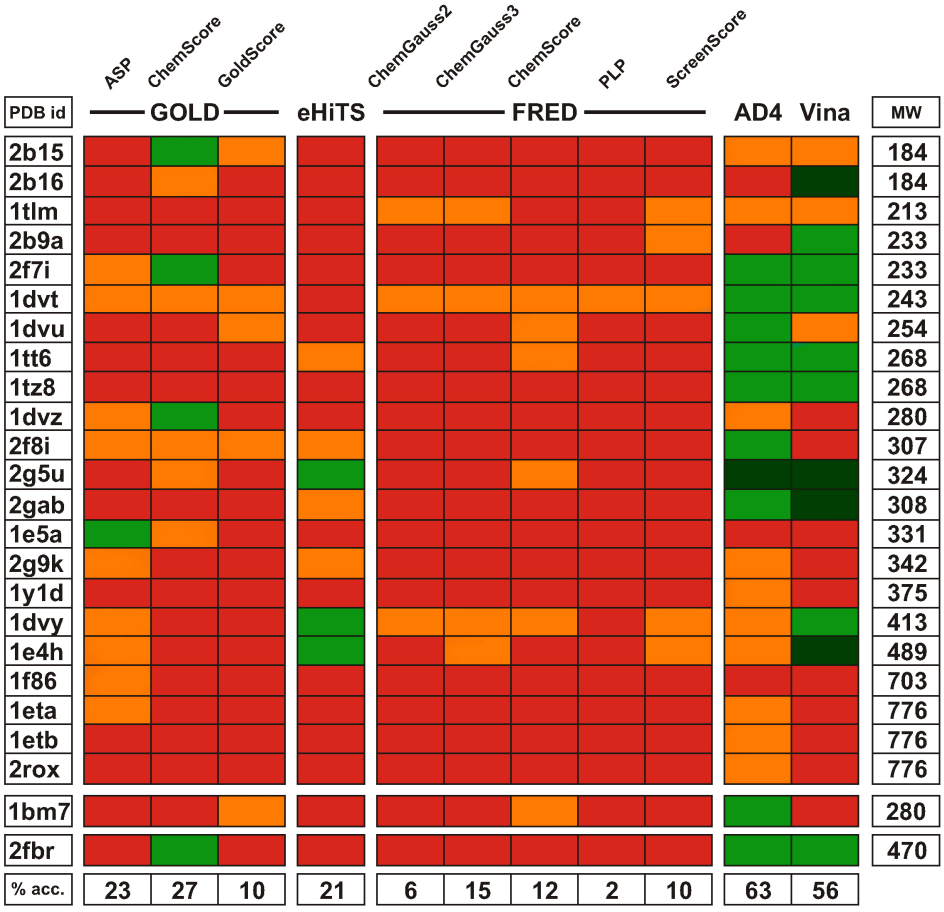


Figure 2.13. Pose prediction accuracy for five docking programs on 24 TTR X-ray complexes. Results of 48 redocking operations performed on the two binding sites of each TTR complex are shown. The rows are sorted according to the ligands' molecular weight (MW column), with the exception of entries 2fbr, a complex of a bivalent ligand, and 1bm7, a complex with two correct ligand poses per binding site (last two separate rows). The top-scoring docking solution provided by each docking/scoring scheme is compared with the native X-ray pose. The colours represent a correct prediction (defined as a top-scoring pose with ligand RMSD ≤ 2.0 Å) in both TTR binding sites (green cells), in one of the pockets

(orange cells) and in none of the pockets (red cells). Exceptional predictions (defined as a correct prediction in both sites with RMSD ≤ 0.5 Å) are shown in dark green.

In terms of ligand molecular weight, all programs seemed to perform better in the range between 230 and 310 Daltons. However, it should be noted that in the case of TTR binders this was most likely related to added difficulty in handling exotic species (particularly iodine) rather than in handling larger ligands. In fact, only AutoDock 4 was able to correctly predict the binding mode of thyroxine, TTR's natural endogenous ligand (containing 4 iodine atoms and 5 rotatable bonds), in one of the binding sites (PDB codes 1eta, 1etb and 2rox).

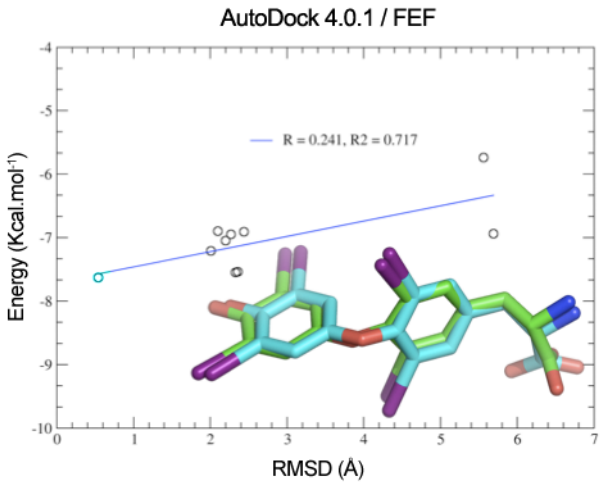


Figure 2.14. Score versus RMSD plot illustrating the pose fidelity of AutoDock 4 while handling a more flexible ligand (T_4) against TTR. The top 10 best-ranked poses are shown as circles in the plots. An acceptable correlation between score and RMSD is disclosed (R^2 = 0.72). The pose holding the lowest RMSD is coloured in cyan and shown in licorice representation against the native X-ray pose (coloured in green).

3.4. Crossdocking studies

Crossdocking is a valuable analysis to understand receptor and ligand specificities, allowing the identification of protein structural models that perform best against a wide variety of ligands. The procedure is very similar to that of redocking, but every X-ray ligand is docked not only to its respective receptor structure but also to every other receptor structure available. To do so, and to quantify the accuracy of the predicted ligand poses through the calculation of an RMSD, all protein-ligand complexes need to be structurally aligned before docking operations take place. As explained in subsection 1.2, in the case of TTR structures, the alignment was produced using a geometric hashing algorithm available in-house [412]. AutoDock 4 was chosen for the crossdocking study, since it was the program showing the highest docking accuracy in redocking tests. Six new complexes

were added to the pool, increasing redundancy (and thus evidence) at the receptor level, and two complexes were removed: 2fbr, due to the lack of two independent ligands occupying TTR's binding sites (in 2fbr, TTR's binding sites are occupied by a unique bivalent ligand), and 1tz8, due to problems during the alignment to the reference structure.

The results of the cross-docking experiments are shown in **Figure 2.15**. The docking accuracy of native ligand pose prediction on each of the 28 TTR receptors was dissected across nine RMSD intervals. It is interesting to notice that despite their structural similarity, the results obtained for the two binding sites differed considerably. As can be verified by contrasting the heat map's diagonal intersection (**Figure 2.15**) with the redocking results in **Figure 2.13**, randomisation of initial ligand coordinates did not produce important loss in docking accuracy. Overall, better predictions were obtained for the binding site designated as "AC" than for the "BD" binding site (designations were taken from PDB chain annotations and dictated by the resulting structural alignments). TTR receptor structures with PDB accession codes 1bm7, 2f8i and 2g5u had the best crossdocking success rates, all with 6 correct predictions (RMSD ≤ 2.0 Å) out of 28 structures of TTR "AC" binding sites. 1bm7, in particular, had the highest overall success (results of both TTR binding sites combined) with 12 correct predictions out of the 56 TTR binding site domains. 2g5u, in turn, corresponds to the complex of TTR bound to the strongest amyloid inhibitor known to date, a polychlorinated biphenyl, and one of the few TTR binders displaying significant positive cooperativity.

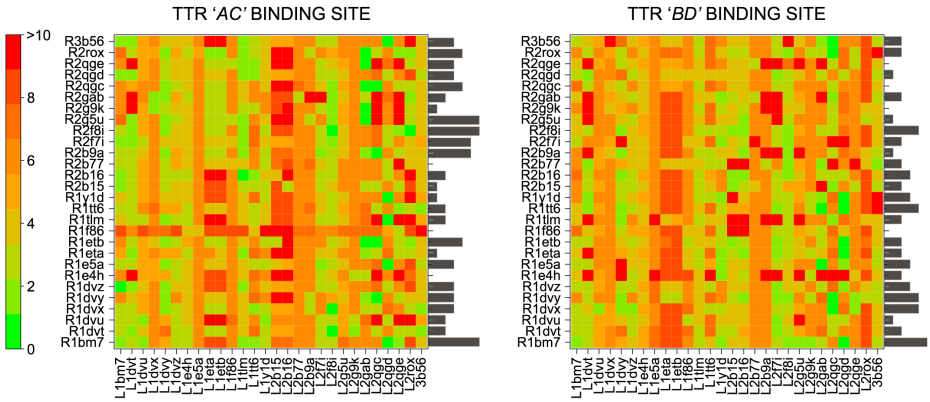


Figure 2.15. Heat maps for cross-docking results on each binding site over 28 TTR complexes. The pool of TTR complexes was refined to make it more diverse and appropriate for cross-docking studies. Six new entries were added in (PDB codes 1dvx, 2b77, 2qgc, 2qgd, 2qge and 3b56) and two entries were removed (1tz8 and 2fbr). The receptor structures are disposed vertically and ligands horizontally. The colour code corresponds to intervals of RMSD values (in Angstroms) determined for the top-ranked poses. For each receptor, the success rate is defined as the number of top-ranked poses with an RMSD ≤ 2.0 Å and is represented in the dark grey horizontal bars.

3.5. Binding affinity predictions

The ability to predict the affinity (or energy) of binding events involved in protein-ligand and protein-protein interactions is a long sought goal in the field of molecular modelling, particularly by the developers of docking programs. However, and as explained in the introductory section, docking scoring functions are mostly tailored for speed of calculation, which largely results from criteriously discarding or simplifying certain elements of reality – such as polarization effects, or any effects involving electrons for that matter. Still, while many scoring functions quantify the strength of P-L and P-P interactions using an undefined metric (e.g. GoldScore, Chemscore, ASP, etc.), others actually attempt to estimate known measures of affinity, such as inhibition constants (e.g. early versions of eHiTS), or the change in Gibbs free energy upon binding (e.g. AutoDock 4 and Vina).

The comparison of experimentally derived binding affinities with the calculated docking scores typically shows weak correlations [448]. Nevertheless, it would be interesting to verify whether binding affinities predicted by AutoDock 4 and Vina correlate with known experimental data of known TTR ligands. Unfortunately, the large amount of structural information on TTR-ligand complexes can only be contrasted with the lack of publically available binding affinity data for TTR ligands. Indeed, no more than 5 association constants (K_a's) could be retrieved from the literature and/or public resources, which greatly prevented meaningful comparisons to be made or significant conclusions to be drawn. Other experimental affinity measures were sporadically found, such as IC50s and pKi's, but these are known to vary significantly with the test assay conditions, limiting their use in this type of study. **Figure 2.16** provides a very brief illustration of an evaluation of AD4's and Vina's scoring function, based on the reported K_a's for T_4, genistein, flufenamic acid, phenox and PCB18 (a detailed description of these compounds was provided in subsection 3.4 of the introductory chapter). While based on very limited evidence, the plot suggests a systematic failure by AD4 and Vina to assign stronger binders (higher K_a's) with stronger scores (more negative scores) and thus to correctly guess experimental binding affinities. In fact, in both cases positive correlation coefficients were found showing that the FEF scoring function tends to favour ligands with higher molecular weight, over ligands showing stronger experimental binding affinities, such as the PCB18.

Even though this study is based on little evidence, the results provide a clear indication that binding affinity predictions of TTR-ligand interactions are unreliable and that great care should be taken when using them in drug design studies such as virtual screening. Alternative methods can be envisaged in order to withdraw more reliable predictions, but these methods are often very slow and unpractical for VS. A viable approach to tackle

this problem would be to identify an alternative scoring function to complement AD4's and Vina's good performance at pose prediction, by correctly scoring and ranking ligand poses. This approach will be pursuit in Chapter 5.

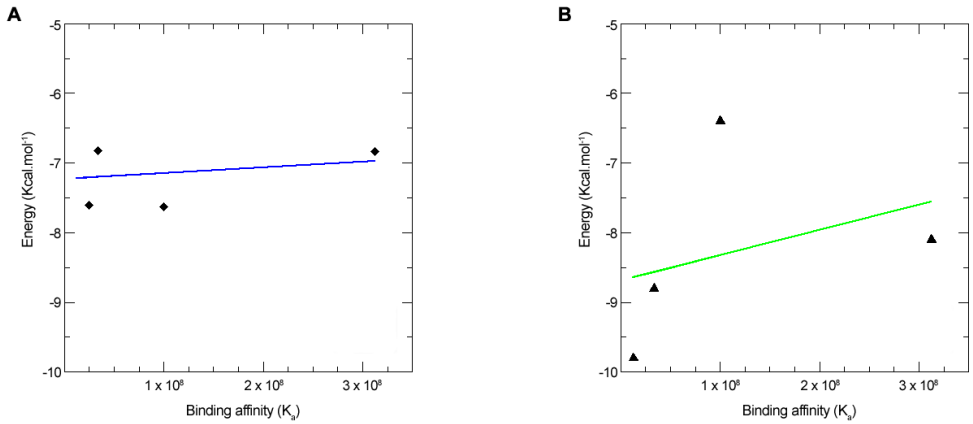


Figure 2.16. Comparison of binding affinities predicted by AutoDock with experimental binding affinities. The diamonds correspond to affinity values determined by AD4 (A), whereas the triangles correspond to affinities predicted by Vina (B). Correlation coefficients of 0.24 (blue regression line) and of 0.35 (green regression line) were found for AD4 and Vina, respectively. In both cases, the positive correlations suggest failure to assign stronger binders (higher K_a's) with stronger scores (lower, i.e. more negative scores).

4. Concluding remarks

In this chapter we conducted structural quality evaluation for a number of TTR X-ray complexes deposited in the PDB, in order to identify a set of high-quality structures for structure-based drug design approaches. The selected complexes were then studied using geometric matching and the complex with PDB code 2g9k was identified as the structure holding highest overall structural similarity amongst a set of 30 TTR complexes. This structure was thus used as reference template both for alignment and for chain renaming. The structural similarity scoring matrix resulting from the multiple structural alignments were clustered and minor, yet systematic, structural differences in Cα atom coordinates were detected and used as guidance to derive new chain annotations and, hence, logical binding site nomenclature.

Furthermore, five well-known docking programs were tested for pose fidelity against TTR. AutoDock 4 and its native free energy scoring function was the program providing the best overall predictions, despite the low speed of its algorithms. AutoDock 4 was further used to conduct crossdocking studies that allow the identification of a narrow set of TTR X-ray receptor structures performing better against a variety of ligands. In general, TTR binding sites annotated as *AC* seem to provide better results than binding sites *BD*.

However, given the more or less subjective nature of the binding site annotation, the results in both binding sites were considered. With this in mind, TTR structures with PDB codes 1bm7, 2f8i and 2g5u offered the most reliable crossdocking results, with 2g5u holding the best structural quality indicators.

In spite of the careful design of the analyses carried out in this chapter, the implications of some results should be taken *with a grain of salt*. My *self*-criticism mainly goes to the metrics used for quantification of docking accuracy. Although every redocked pose was visually inspected throughout the redocking studies, in order to ensure that the computed RMSD value mirrored the quality of the predictions, the same could not hold for the crossdocking analysis, wherein approximately 1500 docking calculations were performed. Often, RMSD values do not reflect the quality of a solution provided by a docking program. Most frequently, they can be unrealistically high, simply because a small part of the ligand adopted a wrong conformation, while most of the ligand structure was correctly docked. Lately, a few approaches have been proposed to tackle this problem, such as IBAC [464] and RSR [465]. While the real value of these metrics as practical solutions still needs to collect the consensus of experts in the field of docking, the RMSD still stands as the most widely used metric allowing a direct comparison with most results published in the literature.

No sensible correlations between experimental binding affinities and docking-based predictions could be withdrawn from this study. While the limited information on experimental binding affinities for TTR ligands does not permit broad conclusions to be drawn, the obtained results seem to corroborate an increasing view that docking scoring functions are poor predictors of small-molecule affinities for protein targets. Nevertheless, AutoDock's ability to correctly predict native ligand poses on TTR-ligand complexes was a major find of this work, which was taken advantage of in subsequent stages of my PhD project.

5. Acknowledgements

Acknowledgements are due to the Olson's lab for the free distribution of AutoDock 4 and Vina, to SimBioSys for the free/academic licensing of eHiTS, to OpenEye Scientific Software for the academic license of FRED and to the Cambridge Crystallographic Data Centre for support with GOLD. I also feel obliged to acknowledge Gert Vriend for the free licensing of WHAT IF, for showing me how useful a user-unfriendly program can be throughout the WHAT IF course, for the free-of-charge accommodation during the International Computational Drug Discovery course at the CMBI (Nijmegen, The Netherlands),

and for the inspiring (and spiteful) talk (and discussions thereafter) delivered at the Astbury Centre, University of Leeds. As he himself said, “structural quality is very important but do not worry too much: docking does not work anyway.” I also thank Elmar Krieger for his sympathy and support with YASARA.

Chapter 3

Receptor- and ligand-based pharmacophore models of TTR: questions of symmetry

"The man who sees both sides of a question is a man who sees absolutely nothing."

[Oscar Wilde]

1. Introduction and theory

Several approaches exist to search small molecule structure databases for novel lead compounds, depending on the presence or absence of an experimental protein structure [466]. Pharmacophore modelling and searches are widely utilised to identify common functional groups between a set of known active molecules aligned in 3D space and to screen large databases for molecules that fit the models thus obtained [99]. The *pharmacophore* concept was originally defined by Paul Ehrlich, more than one hundred years ago, as "a molecular framework that carries (*phoros*) the essential features responsible for a drug's (*pharmacon*) biological activity" [467]. The modern definition by IUPAC states that a pharmacophore is "the ensemble of steric and electronic features that is necessary to ensure optimal supramolecular interactions with a specific biological target structure and to trigger (or to block) its biological response" [95]. Indeed, the pharmacophore entails the features reflecting one molecule's ability to participate in interactions such as electrostatic and hydrophobic interactions, charge transfer and hydrogen bonding. These features may then be translated into a query for database searching and to identify new active compounds [96,466,468–470].

Receptor-based pharmacophore models, in particular, provide an excellent alternative to docking-based virtual screening, while still representing specific protein-ligand interactions. As mentioned in the introductory chapter, the exploitation of available experimental protein structures is a relatively new feature in the long history of the development and application of the pharmacophore concept. Nonetheless, a number of studies show enhanced performance of combined receptor- and ligand-based approaches in pharmacophore modelling: I point the reader to references [100,101]. Several programs have been recently developed for the automatic construction and visualisation of pharmacophore models derived from protein structures. An example is the LigandScout software, developed in Thierry Langer's laboratory [103]. LigandScout automates the process of protein and ligand perception and interpretation (hybridization states, bonds), followed by deduction of the pharmacophore. Pharmacophore models thus produced can be exported for VS with other programs, often proprietary software, such as Catalyst (developed by Accelrys), MOE (developed by Chemical Computing Group) and Phase (developed by Schrodinger).

In fact, and in spite of the increasing popularity of pharmacophore modelling and searches, access to this type of methodologies has been seriously hindered to the academia by the outrageous prices of the license fees charged by most software providers. Some but not all are open to negotiate academic licenses at reduced rates, but too often

these are still above the budget of most academic researchers. Only very recently have two novel approaches for 3D pharmacophore elucidation and searching been proposed as freely accessible to the academic community [471–473], and even implemented as web resources [473,474]. While these tools may represent a turning point in the thus far restricted utilisation of pharmacophore-based virtual screening at the academia, examples of their experimental validation and of success cases are still missing.

In the previous chapter I have underlined the limitations known to docking and scoring programs in common use. In the current chapter, a first alternative strategy for the rational design of new TTR amyloid inhibitors is envisaged, relying on a combination of both receptor-based and ligand-based virtual screening. This strategy can be developed with the support of the vast information on the structure of TTR and on strong TTR amyloid inhibitors. To the best of our knowledge, to date no study has reported on the development or use of pharmacophore-based approaches towards the virtual screening of new TTR amyloid inhibitors. Here, we attempt to identify systematic differences in ligand binding modes across the two equivalent TTR binding sites leading to the definition of five training sets for pharmacophore modelling. Molecular Interaction Field (MIF) analyses were conducted in order to identify *hot spots* of strong interaction with several probe atoms within TTR binding sites. Pharmacophore perception resulted from the combination of all conducted analyses and was translated into five pharmacophore models: three models for TTR binding site AC and two models for binding site BD. Each model was evaluated for selectivity against a benchmark set comprised of known TTR stabilisers and inactive (decoy) molecules.

1.1. Pharmacophore modelling techniques

Figure 3.1 illustrates a general workflow for the generation of pharmacophores from multiple ligands. The creation of correct compound structures to use as input for analysis is the first key step to any pharmacophore modelling project, in which several elements, such as atomic valences, correct bond orders and appropriately defined aromaticity need to be accounted cautiously. Moreover, the correct stereochemical flags have to be inserted for a correct treatment of stereochemistry. Most pharmacophore generation programs in common use include compound builders, but the modellers can also import the ligands from external sources using common file formats (e.g. SMILES, MOL, SDF, MOL2, etc.).

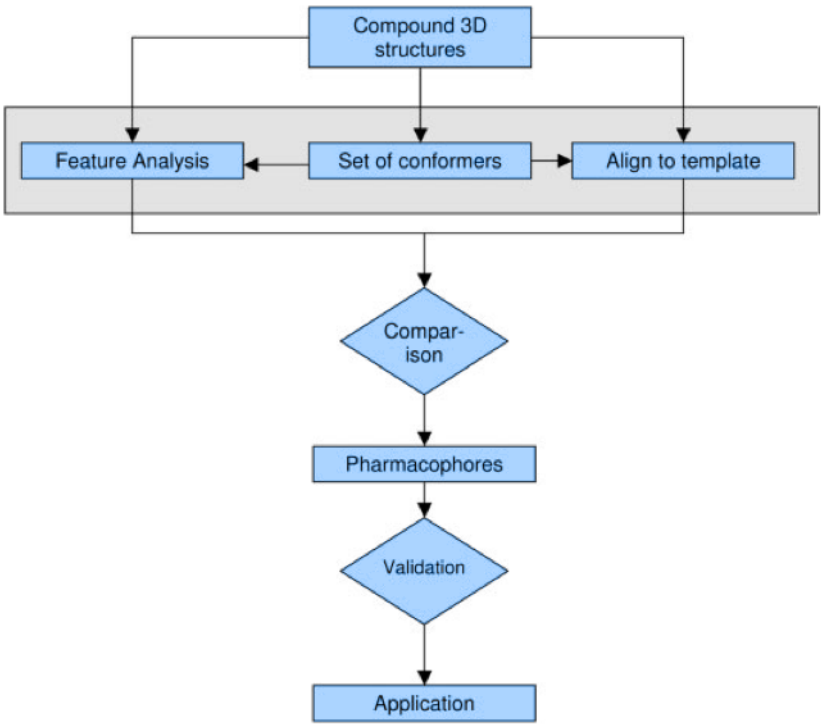


Figure 3.1. A workflow describing the most common procedures used for pharmacophore modelling. Adapted from reference [475].

The representation of pharmacophore models varies significantly from one program to another. However, most representations comprise the nature of the pharmacophore points (fragments, chemical features) and the geometric constraints connecting these points (distances, torsions, 3D coordinates constraints).

Feature Analysis (see **Figure 3.1**,), i.e. the interpretation of the chemical structures of the molecules, can be accomplished at two main levels:

1. Substructural, where molecules are dissected into multiple fragments, each fragment holding certain specifications (e.g. basic nitrogen or aromatic ring).
2. Functional, wherein each molecular fragment of the compounds is translated into the general property it conveys. In this stage, the properties mapped on the fragments are chemical properties, e.g. hydrophobic or ionic interactions or hydrogen bonding features.

The description of the chemical properties of compounds requires their functions to be accessible for the interaction with the binding partner (receptor, enzyme or nucleic acid), so in case the bioactive conformation of the ligand is not known, a *conformational expansion* analysis is a necessary step in order to identify a conformation that prepares those functions for interaction with the target.

The consideration of ligand flexibility is probably the most demanding task in pharmacophore modelling. Methods to account for this problem are generally subdivided into four classes: systematic search in torsional space, optionally followed by clustering, stochastic methods (e.g. Monte Carlo), sampling (e.g. Poling [476]) and molecular dynamics. The goal is not only to have the most representative coverage of the conformational space of the ligands, but also to have either the bioactive conformation as part of the set of generated conformations or at least a cluster of conformations that are close enough to the bioactive conformation. Back in the mid-1980s, Marshall et al. presented the so-called Active Analog Approach [52] that still forms the basis of some of the most prominent automated pharmacophore modelling techniques: the conformational space of flexible molecules is constrained to the geometry of a reference molecule (generally active and as rigid as possible) and pharmacophore models are then derived from the set of resulting alignments. At this point I refer the reader to reference [475] for a comprehensive review of the methods for automated molecular/pharmacophore alignment and generation.

The comparison step shown in **Figure 3.1** corresponds to the actual pharmacophore creation process itself. Most pharmacophore generation programs generate qualitative pharmacophores that do not account for the potency of the ligands. Thus, equipotent compounds should be used. The majority of methods are based on minimizing the RMS superposition error between conformations of several molecules, while attempting to improve the 3D overlay of pharmacophores. Typically, the result is a small set of pharmacophore solutions, ranked according to different metrics depending on the program used.

1.2. Combinations of pharmacophore models

Strategies based on the combination of multiple pharmacophore models are often pursuit with two principal (yet very distinct) objectives: (*i*) to improve the specificity and the selectivity of the models for one single target, by combining their strengths and diluting their weaknesses; (*ii*) to discover new ligands that bind to multiple targets. While the idea behind objective (*i*) seems easy to grasp and will be explored in this chapter, the motivation underlying objective (*ii*) may look more surprising. Without any doubt, many highly specific mono-target drugs have proven the value of mono-target medicine. However, clinicians have been increasingly convinced that modulating an array of targets can be advantageous in treating a range of pathologies. Nowadays, it is already broadly recognized that high specificity for a single target may not deliver the necessary effectiveness versus adverse-effects ratio and, in many cases, a balanced activity at multiple targets may yield better results. Morphy et al. weighed the opportunity and the advantages associated to the design of drugs acting on two (or more) targets, as opposed to seren-

dipitous promiscuous ligands [477]. A growing number of publications reflect an awakening of interest in the rational design of "promiscuous" drugs and may imply an ongoing re-assessment of the "one target, one drug" concept that has dominated thinking in the pharmaceutical industry over the last decades [478–480].

Certainly, computer-aided drug design strategies combining multiple pharmacophore can lead to the identification of novel chemical entities merging in one molecule the fundamental elements of multiple partners.

1.3. Validation of pharmacophore models

Following analysis on a training set of compounds, multiple pharmacophore hypotheses are often generated and the careful selection of model(s) with biological and/or statistical relevance becomes a critical necessity prior to any further research endeavours.

Typically, the validation methods can be gathered around three groups:

1. Statistical significance analysis, randomisation tests.
2. Enrichment measurements. These measurements quantify the recovery of active molecules from a benchmark data set in which a small number of known actives is seeded amongst selected decoys. The utilisation of receiver operating characteristic (ROC) curves fits in this group and will be extensively discussed in Chapter 5.
3. Experimental evaluation of a selection of hits.

In the work presented in this chapter, database searches were conducted with Tripos data mining utility UNITY 3D, which retrieves/discards database compounds on a pass/fail basis. Since no scoring or ranking of the compounds takes place, the performance of the pharmacophore models can only be assessed by an overall measure of enrichment, which will be explained in the "Results and discussion" section of this chapter.

The validation of methods and models is indeed a critical part of pharmacophore modelling and searches, in particular, and of virtual screening, in general. A review of the performance metrics used for validations applicable to the field of pharmacophore searches and virtual screening is given in Chapter 5.

2. Computational methods

In the following subsections, I will present and discuss the computational methods employed throughout this chapter to study, at the atomic level, the available structural information on the TTR receptor and known active ligands, which will serve as input to the construction of pharmacophore models of TTR.

2.1. Receptor-based pharmacophore perception

Once again driven by the significant amount of structural information available in the PDB for TTR, we decided to incorporate receptor knowledge in the construction of a set of pharmacophore models. Besides using structural data alone, MIF analyses were carried out with two main objectives: (*i*) the detection of *hot spots* in TTR binding sites and (*ii*) the detection of consistent differences between the two equivalent TTR binding sites. The idea was to map and make use of energy minima corresponding to highly favourable points of interaction of certain probe types within TTR sites during the construction of pharmacophore models. To this end, the X-ray complexes of four reference TTR stabilisers were studied in great detail: thyroxine (T_4), the endogenous thyroid hormone carried by TTR in the blood plasma and the cerebrospinal fluid, bearing four iodine atoms; diethylstilbestrol (DES), a synthetic non-steroidal estrogen with affinity for TTR and devoid of halogens; flufenamic acid (FLU), a non-steroidal anti-inflammatory drug holding three fluorine atoms and one of the first compounds found to potently inhibit amyloid fibril formation by TTR; and one particular polychlorinated biphenyl (PCB), PCB18, holding four chlorine atoms and the strongest TTR amyloid inhibitory activity known to date (**Figure 3.2**). The respective coordinates and structure factors were downloaded from the PDB and evaluated for structural quality. When relevant, the corresponding re-refined and conservatively optimised structure was obtained from the PDB_REDO [481]. PDB accession numbers are as follows: TTR:T_4 complex, 2rox; TTR:DES complex, 1tt6 and 1tz8; TTR:FLU complex, 1bm7; TTR:PCB complex, 2g5u.

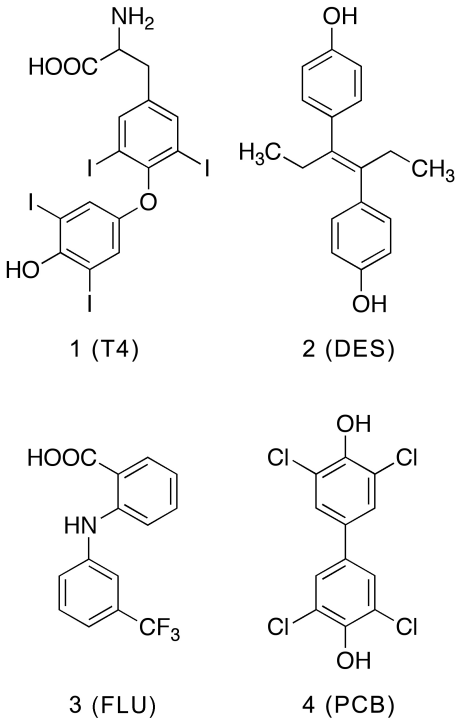


Figure 3.2. Chemical formulae of the TTR stabilisers bound in the complexes studied in greater detail throughout this chapter. Ligand 1, thyroxine (T_4), the endogenous thyroid hormone carried by TTR; ligand 2, diethylstilbestrol (DES), a synthetic non-steroidal estrogen with affinity for TTR; ligand 3, flufenamic acid (FLU), one of the first non-steroidal anti-inflammatory found to potently inhibit TTR amyloid; and ligand 4, a polychlorinated biphenyl (PCB), the strongest inhibitor discovered to date.

2.1.1. Receptor site analysis using MD simulations

Two 30 nanoseconds(ns)-long equilibrium Molecular Dynamics (MD) simulations of the TTR tetramer were performed in explicit water with two goals: (*i*) to ensure that no important conformational changes occur within TTR binding sites in the equilibrium state, motivated by possible structural clashes caused by the symmetry operations required to assemble the full tetramer from the X-ray coordinates of the dimer; (*ii*) to contrast the MD analysis of receptor site flexibility with the picture provided by the X-ray B-factors of the crystalline state. The TTR structure holding the highest degree of overall structural similarity amongst the pool of TTR complexes presented in **Table 2.2** of Chapter 2, was selected as input for the construction of the model system (PDB entry 2g9k). The initial coordinates of the full quaternary assembly of TTR were obtained from the corresponding entry in the Protein Quaternary Structure (PQS) server [453]. The missing N- and C-termini were added and subjected to simulated annealing with the rest of the protein held fixed. All minimization, equilibration and MD production steps were carried out with NAMD [482], version 2.6, using the all-atom CHARMM 27 force field [483]. All calcu-

lations were run on a cluster of 130 Sun Fire X4100 nodes with a total of 520 Opteron 275 CPU cores and using up to 64 cores (*Milipeia* at the University of Coimbra).

Briefly, a strict protocol was followed to set up the model system. Besides preserving crystallographic water molecules that looked meaningful, additional water molecules were inserted in hydrophilic cavities of the tetramer with the program Dowser [484], after stripping out the bound ligands. Hydrogen atoms were added to the protein structure using the *psfgen* plug-in in VMD [462]. The tetramer was then embedded in a large box of TIP3P water molecules [485] with the *solvate* plug-in, ensuring that a layer of at least 12 Å was comprised between the protein and the edges of the box. The simulation box was forced to assume a "perfect" cubic shape, thus ensuring that the rotation of the protein about itself would not compromise the minimum distance between the solute and the box boundaries. Sodium and chloride ions were included to bring the net electric charge of the system to zero and an average ionic concentration of 145 nM was reached in order to better reproduce the physiological environment.

The system was then subjected to three cycles of 2000 energy minimization steps: a first cycle applied to all hydrogen atoms of the system, followed by a second cycle applied to the solvent and the protein side chains, and a third cycle applied to the whole system. The system was equilibrated up to the target temperature of 310 K using an integration time step of 1 femtosecond (fs) and under Langevin dynamics (with a Langevin damping coefficient of 5). During the initial 50 picoseconds (ps), the protein backbone was held fixed. The equilibration proceeded for another 50 ps to bring the system to the target pressure of 1 atm using the Langevin piston method [486]. The equilibration continued at constant pressure, for 20 ps, with strong harmonic restraints placed on the C_α carbon atoms, and another 20 ps with weak restrains on the same atoms. Finally, a short 10 ps run of unrestrained classical MD was performed to test the stability of the system, completing an overall equilibration of 150 ps. To avoid biases taking place throughout the equilibration stage, each production run was independently equilibrated.

The two independent 30-ns long MD simulations were performed at 310 K, in the isothermal-isobaric ensemble (normal pressure and temperature conditions), and using periodic boundary conditions (PBC). The particle mesh Ewald summation method [487] was used for the quantification of long-range electrostatic interactions at every step. A 12 Å cut-off with a switching function applied at 10 Å was used for the treatment of the van der Waals interactions, with the pair list distances being evaluated every 10 steps. Hydrogen-heavy atom bond lengths were constrained with the SHAKE algorithm [488]. An integration time step of 2 fs was used during the production runs, and the coordinates of the system were recorded every 500 steps (i.e. at each ps of simulation).

2.1.2. Detection of *hot spots* with programs GRID, MINIM and FILMAP

The GRID software (version 22c) [39] was used for the detection of regions of favourable interaction (*hot spots*) within TTR binding sites, using the four model complexes presented above. After removal of ligand and solvent atoms, a grid of dimensions 40×14×14 Å and 4 grid points per Angstrom (i.e. a resolution of 0.25 Å) was constructed, encompassing the entire tunnel formed by both TTR binding sites. The effective dielectric constant of the macro-molecular matrix was kept at 4 and the dielectric constant of the environment surrounding TTR was kept at 80. Several grid maps were produced for each complex by computing the energies of interaction between 13 types of probes, placed at each grid point, and the protein (see **Table 3.1**). In order to study the presence of disfavoured regions of interaction with certain probe/atom types in TTR binding sites by competition with water molecules, the so-called Reverse Mode of GRID was explored by setting the LEAU directive to 3 and re-computing all grid maps for every probe type and each TTR complex.

The GRID module MINIM was used to identify energy minima in the grid maps, which are grid points surrounded by points of higher (more positive) energy. Different energy thresholds (measured in kcal.mol^{-1}) were used for 6 main probe types: -1.5 for DRY; -3 for C3; -9 for OH; -3 for F; -6 for Cl; -7 for I. The module FILMAP was then utilised to populate the energy minima in the grid maps, by annealing successive probes of each type into the list of minima produced by MINIM. This procedure refines the map of energy minima by discarding minima of probe types that repel each other and retaining minima of probe types that interact favourably.

Table 3.1. List of probes used for MIF analyses with the program GRID.

Probe Notation	Probe Name	Probe Type
F	Fluoride	Halogen
Cl	Chloride	Halogen
BR	Bromide	Halogen
I	Iodide	Halogen
N1	Amide	Hydrogen Bond Donor
O	Carbonyl	Hydrogen Bond Acceptor
O::	Carboxyl	Hydrogen Bond Acceptor
$N1^{+}$	Cationic amide	Positive Charge Centre
O^{-}	Anionic phenolate	Negative Charge Centre
OH	Hydroxyl	Hydrogen bond donor/acceptor
C3	Methyl	Non-polar/hydrophobic
DRY	Hydrophobic	Hydrophobic
OH2	Water	Water binding

2.2. Ligand-based pharmacophore perception

In order to carry out ligand-based pharmacophore perception, the 30 high-quality TTR complexes selected for docking studies (see **Table 2.2** of Chapter 2) were screened through tight quality cut-off and ligand redundancy was removed through selection of structures with highest quality from amongst a group of closely related structures. A set of 21 aligned protein-ligand complexes resulted (**Table 3.2**) and was used to define training sets for the modelling of 3D pharmacophore models.

Table 3.2. Twenty-one X-ray complexes of transthyretin used for ligand-based pharmacophore perception.

PDB code	Resolution (Å)	R-value	Description
1bm7	2.00	0.189 (work)	Transthyretin in complex with flufenamic acid
1dvs [a]	2.00	0.199 (work)	Transthyretin in complex with resveratrol
1dvt	1.90	0.197 (work)	Transthyretin in complex with flurbiprofen
1dvu	1.90	0.198 (work)	Transthyretin in complex with dibenzofuran-4,6-dicarboxylic acid
1dvy	1.90	0.202 (work)	Transthyretin in complex with phenoxazine-4,6-dicarboxylic acid
1dvz	1.90	0.192 (work)	Transthyretin in complex with o-trifluoromethylphenyl anthranilic acid
1e4h	1.80	0.193 (work)	Transthyretin in complex with pentabromophenol
1tt6	1.90	0.188 (obs.)	Transthyretin in complex with diethylstilbestrol
1y1d	1.70	0.198 (obs.)	Transthyretin in complex with iododiflunisal
2b15	1.70	0.190 (obs.)	Transthyretin in complex with 2,4-dinitrophenol
2f7i	1.60	0.234 (obs.)	Transthyretin in complex with 2',6'-difluorobiphenyl-4-carboxylic acid (diflunisal analogue)
2f8i	1.54	0.217 (work)	Transthyretin in complex with a benzoxazole derivative
2g5u	1.80	0.180 (work)	Transthyretin in complex with 3,5,3',5'-tetrachloro-biphenyl-4,4'-diol
2g9k	1.85	0.216 (obs.)	Transthyretin in complex with 2',3,3',4',5-pentachloro-biphenyl-4-ol
2gab	1.85	0.211 (obs.)	Transthyretin in complex with 3,3',4',5-tetrachloro-biphenyl-4-ol
2rox	2.00	0.170 (work)	Transthyretin in complex with thyroxine
1dvx	2.00	0.192 (work)	Transthyretin in complex with diclofenac
2b77	1.70	0.210 (obs.)	Transthyretin in complex with a diflunisal analogue
2qgd	1.50	0.162 (obs.)	Transthyretin in complex with 2-(3,5-dibromo-4-hydroxyphenyl)benzoxazole
2qge	1.45	0.166 (obs.)	Transthyretin in complex with 2-(3,5-dimethylphenyl)benzoxazole
3b56	1.55	0.190 (obs.)	Transthyretin in complex with 3,5-diiodosalicylic acid

[a] Even though this entry was excluded from the original selection for structure-based docking studies, it was selected for ligand-based pharmacophore perception given resveratrol's significance as an important TTR stabiliser.

2.2.1. Ligand mapping using a Java script

The X-ray TTR ligands were subjected to an all-atom comparison using the coordinates resulting from the superimposition of the 21 complexes. This comparison was accomplished using a simple Java script written in-house: coordinates found to be identical (or lying within 1 Å distance) in two bound ligands were considered a "hit"; the total number of hits was computed and used to calculate pair wise Tanimoto Indexes (T.I.) for all ligands on each of the two TTR pockets, according to the formula:

$$T.I. = N_{AB} / (N_A + N_B - N_{AB})$$

Equation 3.1

N_A corresponds the total number of ligand atom coordinates found in the first complex; N_B corresponds to the total number of atom coordinates in the second complex, and N_{AB} expresses the number of atom coordinates found to be identical in both complexes (the so-called "hits"). A T.I. was calculated for every ligand bound to the AC and BD binding sites of TTR and combined to produce a similarity score matrix for each pocket. Then, each similarity score matrix was clustered using the program CLUTO [454].

2.2.2. Ligand preparation with QuacPac

The appropriate preparation of ligands is a critical ingredient to any pharmacophore modelling project. The calculation of partial charges and the enumeration of tautomers (followed by canonicalisation) affect the way ligand features (such as H-bond donors and acceptors including tautomers; anions and cations, including resonance forms and hydrophobic and aromatic regions) are encoded into the resulting pharmacophore models. For each of the 21 ligands of the selected TTR complexes, the lowest energy tautomer was enumerated with the *tautomers* program, part of OpenEye's QuacPac suite [489], just after the calculation of AM1-BCC partial charges using the *molcharge* program. The AM1-BCC method first determines partial charges derived from the AM1 wave-function. Then, bond-charge corrections (BCC) are applied to the partial charges on every atom to create the final partial charges [490].

2.3. Training and validation sets

The pharmacophore models devised in this chapter were calibrated and tuned with the help of 5 separate *training sets*. Twenty-one known TTR actives were split across each set according to their spatial occupation of TTR binding sites (as will be explained below)

and seeded amongst 20 known TTR inactive molecules. Thus, each training set is comprised of 4 to 7 actives (presented in subsections 3.2 and 3.3) and 20 inactives. All TTR actives were compounds experimentally proven to bind and stabilise TTR, whereas the inactives were compounds exhibiting low or null (experimental) binding affinity for TTR.

For the purpose of validation, the generated pharmacophore models were run against a *validation* test set comprised of 22 potent and structurally dissimilar TTR actives, seeded amongst 738 carefully-selected decoy molecules, sharing similar physicochemical properties with the TTR actives, yet dissimilar chemical topologies. Details about the construction of this validation test set, particularly in what concerns the selection of decoys, are given in Chapter 4.

2.4. Pharmacophore model generation using UNITY and SYBYL

SYBYL 8.0, a molecular modelling and computational informatics platform provided by Tripos, was used to develop pharmacophore models for TTR binding sites. SYBYL allows user-friendly visualization and definition of pharmacophore features for a set of ligands, and the models thus developed are referred to as *queries*. SYBYL's UNITY module allows the addition of features and constraints to a query, which may then be screened against a virtual library of compounds. UNITY allows 3D searching using queries not only from pharmacophores but also using surface and excluded volumes, queries derived from receptor binding sites and 2D substructure and similarity searching. The 3D flexible searching is based on the Directed Tweak method [491]. Conformational flexibility during searching with UNITY is handled on-the-fly.

Ligands grouped in five training sets were imported to SYBYL in the MOL2 format. Since the alignment of the ligands was obtained as a result of the multiple structural alignment of TTR complexes, no automated features were used to generate ligand alignment underlying the pharmacophore query. Moreover, since all ligand structures were determined by X-ray crystallography in complex with TTR, it was assumed that their pose corresponded to the bioactive conformation. Thus, ligand coordinates were taken *as is* and no structure optimisation was performed.

Amongst the constraint options available in UNITY, the most relevant are distance constraints, angle constraints, spatial constraints and bond path constraints. Distance and angle constraints define the allowed distance separating two features and the angle between the features respectively. Spatial constraints in the form of spheres, limit the geometrical position of a feature to within a sphere. Bond path constraints define the maxi-

mum and minimum number of bonds between two features. All generated pharmacophore queries include a spatial constraint of 1.80 Å on aromatic features covering all ring atoms. A higher spatial constraint, when added, was either to accommodate additional ligand atoms or to account for greater flexibility of receptor site residues. Receptor site information was used to define exclusion spheres with volumes directly proportional to a normalized scale of residue flexibility. This scale was empirically parameterized using insights extracted from B-factor analysis and MD simulations.

3. Results and discussion

In this section, the results of receptor- and ligand-based analyses leading to the construction of five pharmacophore models for TTR are presented, along with a detailed description of the generated models and respective selectivity analysis.

3.1. Receptor-based pharmacophore modelling

Here, we present the results of a set of studies centred on the structure of TTR, allowing for a full characterisation of the target receptor site and the modelling of receptor-based pharmacophore features (such as receptor site volume spheres).

3.1.1. Receptor site flexibility

Receptor site flexibility is a critical aspect in drug design, given the dynamic nature of biomolecules in solution. Additionally, the explicit consideration of receptor flexibility can allow the construction of *dynamic pharmacophore models* instead of static ones. This is of the utmost importance when the receptor undergoes significant structural changes upon ligand binding (e.g. induced fit effects). Receptor flexibility can be accounted in several different ways, from the use of experimental data to the application of simulation methods, such as conformational searches and MD simulations. In this work, we made use of the large amounts of structural data on TTR deposited in the PDB and carried out statistical analysis of the temperature factors (also referred to as B-factors) associated with each atom of the receptor side chains and bound ligands. The B-factors are linearly related to the mean square displacement of an atom and provide an indication of structural flexibility in the crystalline state. It has been suggested that, in general, active site residues predominantly occur in regions of low B-factors, whereas non-active site residues at the protein surface have a tendency to exist in the high B-factor regions [492].

Visual comparison of the B-factor values for all entities involved in ligand binding to TTR is facilitated using the box plots from **Figure 3.3** to **Figure 3.5**. As expected, residues lining the outer end of the binding sites possess the highest median values, namely Lys 15, Glu 54 and, to a lesser extent, Val 121. While it could be expected that, by contrast, atoms of the inner residues had lower median B-factor values, the analyses show that, except for the TTR:FLU complex, Ser-117 and Thr-119 residues have slightly higher flexibility than residues comprising the centre channel of TTR binding sites (Ala 108 and Leu 110). Interestingly, of the 4 complexes studied, FLU is the only compound who does not hold an hydroxyl group at the *para* position of the inner aromatic ring, thus not being able to interact with Ser 117 or Thr 119, via hydrogen bonding, either directly or mediated by water molecules.

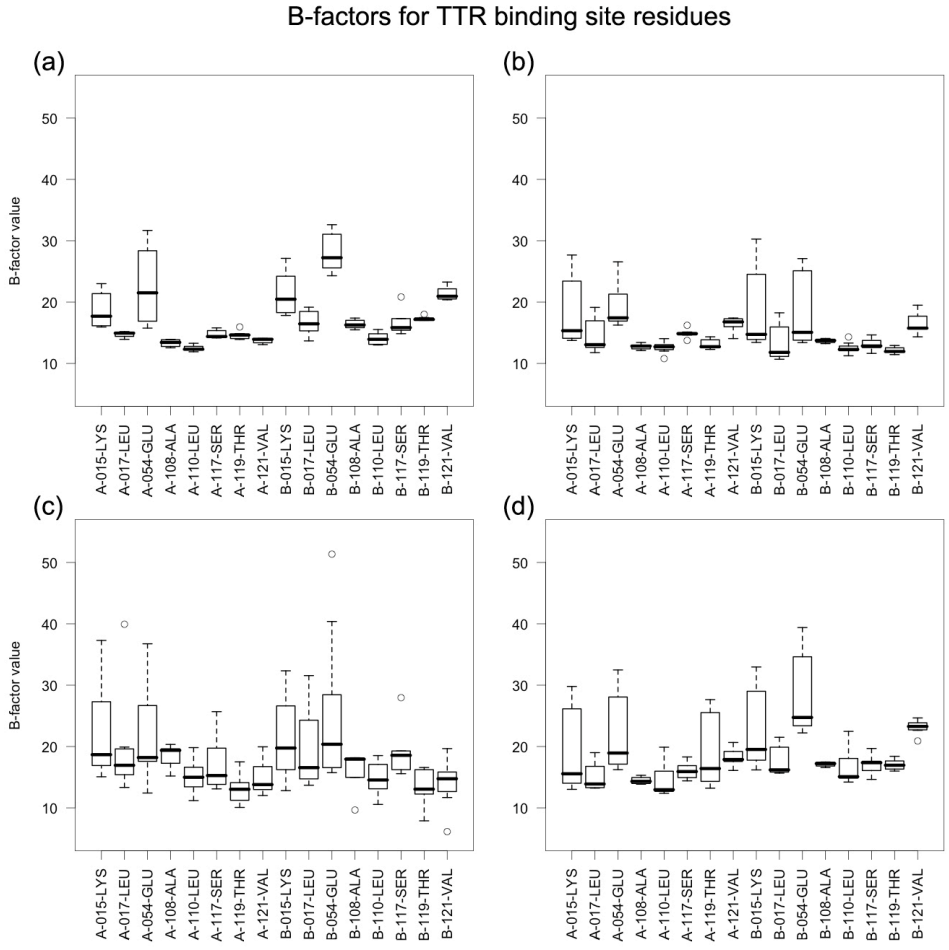


Figure 3.3. Box plots of the B-factor values for all heavy atoms of each TTR binding site residue of monomers A and B of the crystallographic dimer (box region providing a graphical view of the median and quartiles of the distribution; dotted lines extending to maximum and minimum values of the data set). The B-factor values of the protein atoms composing the data set were extracted from the structure of TTR complexed with (a) T_4, resolved at 2.00 Å (PDB entry 2rox), (b) DES, determined at 1.80 Å (PDB entry 1tt6), (c) FLU, solved at 2.00 Å (PDB entry 1bm7), and (d) PCB18, determined at 1.80 Å (PDB entry 2g5u).

Furthermore, the median B-factor values of the ligands correlates reasonably well with their molecular weight, which may in turn be linked to their ability to diffuse along TTR binding sites. As depicted in **Figure 3.5**, PCB18 holds the highest B-factors values compared to any other binder under study.

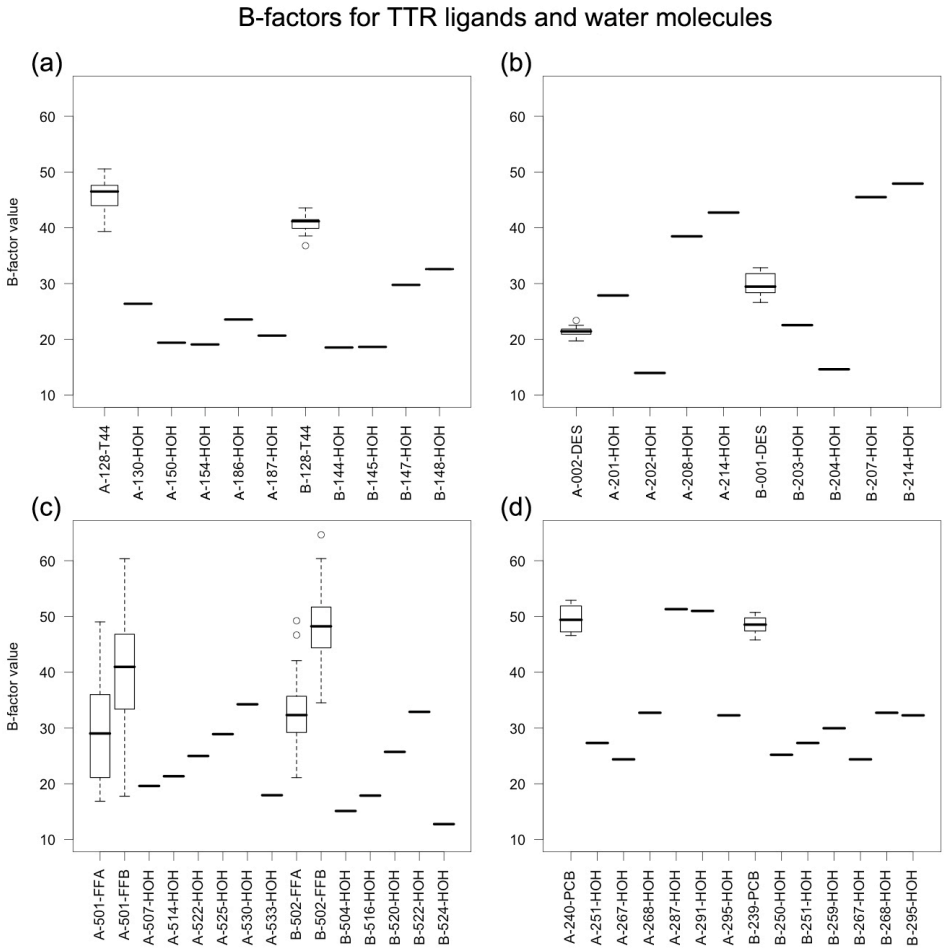


Figure 3.4. Box plots of the B-factor values for each reference ligand bound to each TTR pocket and for water molecules found within 7 Å of any ligand atom (box region providing a graphical view of the median and quartiles of the distribution; dotted lines extending to maximum and minimum values of the data set)**.** The B-factor values of the ligand atoms composing the data set were extracted from the respective TTR complex: (a) T_4, resolved at 2.00 Å (PDB entry 2rox); (b) DES, determined at 1.80 Å (PDB entry 1tt6); (c) FLU, solved at 2.00 Å (PDB entry 1bm7); and (d) PCB18, determined at 1.80 Å (PDB entry 2g5u).

Water molecules seem to play a role in the interaction of PCB18 with TTR. Unlike in other complexes, two water molecules appear "strategically" placed at the inner end of the pocket, mediating interactions between the compound and the hydroxyl side chains of Ser 117. In fact, this is the only structure where these side chains adopt three alternate positions with equal values of occupancy (0.33). In most other structures, the hydroxyl groups of Ser-117 residues are either hiding from any interaction site or just exchanging

between the site of interaction with ligands and site of interaction with an adjacent Ser 117 of the neighbour TTR monomer.

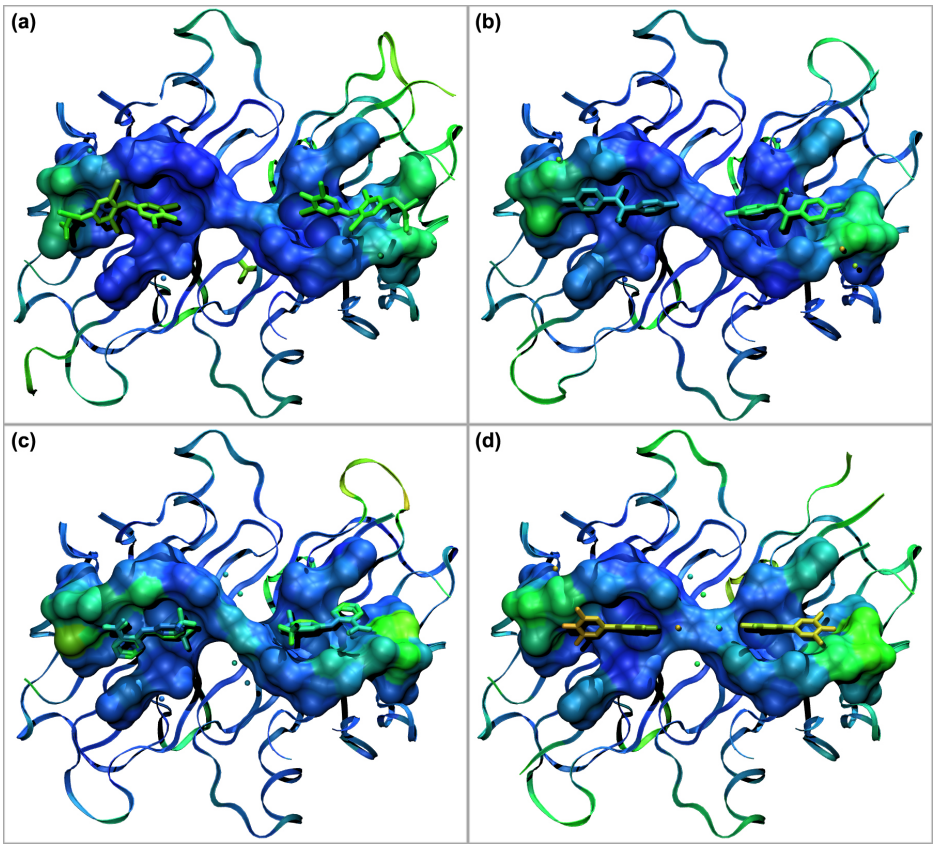


Figure 3.5. Three-dimensional (3D) representations of TTR-ligand complexes determined by X-ray crystallography and coloured according with their B-factors. The asymmetric unit dimers are drawn as ribbons; the binding site residues are represented as molecular surfaces; the ligands and their neighbouring water molecules (within 7 Å of any ligand atom) are shown in licorice representations. The colour scale goes from blue to green and from green to yellow as the B-factor values increase. (a) T_4, resolved at 2.00 Å (PDB entry 2rox); (b) DES, determined at 1.80 Å (PDB entry 1tt6); (c) FLU, solved at 2.00 Å (PDB entry 1bm7); and (d) PCB, determined at 1.80 Å (PDB entry 2g5u).

Remarkable cases are known where medicinal chemists have been deceived by incomplete experimentally-determined structures and/or by "deleterious" modelling operations. One good example relates to the design HIV integrase inhibitors [493], where the authors used the so-called *Relaxed Complex method* based on MD simulations [494] to take protein flexibility into account during the docking procedure. Eventually, they found out that the original structures had been strongly affected by crystal packing effects [495].

Because all TTR complexes used in project were obtained (indirectly) via symmetry operations to assemble the full biological unit, equilibrium MD simulations of TTR tetram-

ers were carried out both to detect possible problems with X-ray derived models (for instance, due to crystal packing effects) and to study the structure and flexibility of TTR tetramers. The main focus of our analyses was put on the structure of TTR binding sites throughout the simulations, to ensure that no major conformational rearrangements occurred. Visual inspection of the MD trajectories did not reveal any important conformational changes in TTR, no significant rearrangements in the binding sites were witnessed and no consistent differences, neither in the structure nor in the dynamics, could be detected across the two binding sites. These results suggest that the static biological unit of TTR – obtained through symmetry operations – represents well the structure of the equilibrium state.

Root mean square fluctuations (RMSF) were determined over 50 ns of the MD simulations (see **Figure 3.6**) and contrasted with the B-factor distributions. The first 5 ns of each of the two trajectories were discarded. Small variations were observed for the C_α atoms of the residues composing TTR's binding pockets (the first pane of **Figure 3.6** should be interpreted with caution, due to the reduced scale of the *y*-axis). The RMSF values suggest that the most flexible side chains in the binding sites are those belonging to Lys 15 and Glu 54 (Arg 104 is a neighbouring residue, but does not participate in protein-ligand interactions). Amongst the residues composing the inner portion of TTR binding sites, Ser-117 residues are the most flexible. By contrast, Ala 108 and Leu 110 show the lowest RMSFs. These results are in good agreement with our B-factor analyses.

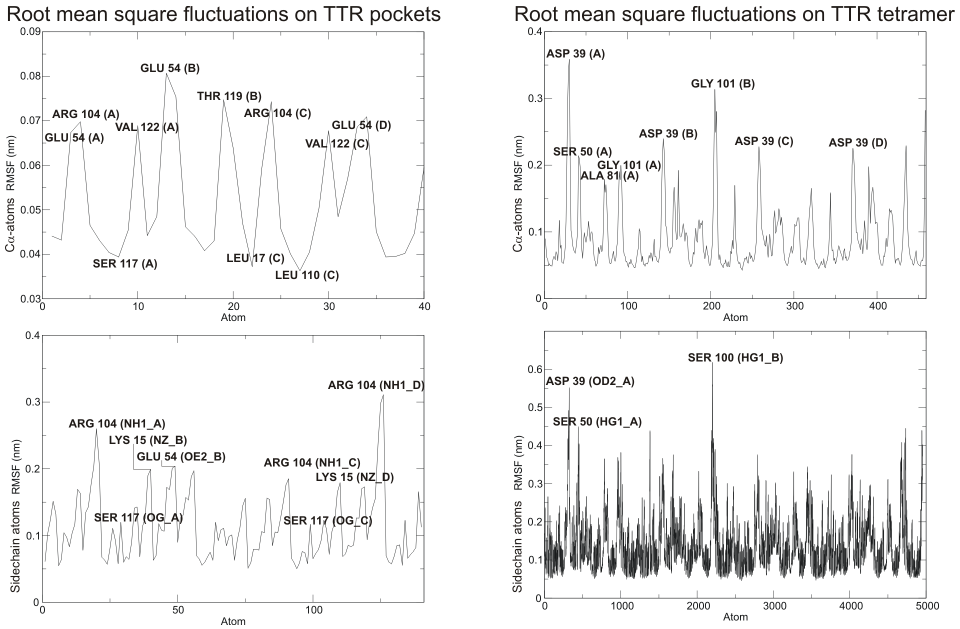


Figure 3.6. RMSF plots derived from 50 ns of MD simulations of an apo structure of TTR. The top panels refer to C_α atom fluctuations whereas the lower panels refer to all side-chain atoms. The left panels refer to residues comprised in TTR binding sites and the right panels refer to all protein residues.

In Section B of the Appendix, the C_α-RMSD variation from the crystal structure throughout one MD trajectory provides a picture of the consistency of the simulations (**Figure B.1**). The RMSD values vary mostly below the 2 Å threshold, disclosing an adequate stability. Further analyses and insights could have been extracted from the trajectories. However, because of the intricate nature of MD simulations, their extensive analysis falls outside the scope of this thesis. The results of these studies will be reported elsewhere.

3.1.2. Detection of *hot spots* within TTR binding pockets

Besides identifying favourable regions of interaction within a protein binding site, it is also important to know if specifically-unfavourable places exist on the target, because they can influence the affinity and selectivity of drug-receptor interactions, and the orientation of a ligand at its receptor site. Often, at a later stage of the research when the X-ray structure of the protein-inhibitor complex has been determined, it sometimes turns out that the ligand does not bind as predicted but has found an alternative binding mode. This can happen because there are positions, which may be called *disfavoured sites* in some receptor clefts from which a specific chemical group seems to be excluded because water would interact better. **Figure 3.7** shows the isocontours for five probes and at different energy thresholds (adjusted to each probe), determined on the X-ray structure of each of the four model complexes. The solid surfaces represent the most favourable regions of interaction, whereas the dots disclose the regions were water molecules could displace a probe atom of the respective type. The FILMAP program takes these disfavoured sites into consideration while selecting only the relevant minima for interaction (**Figure 3.8**).

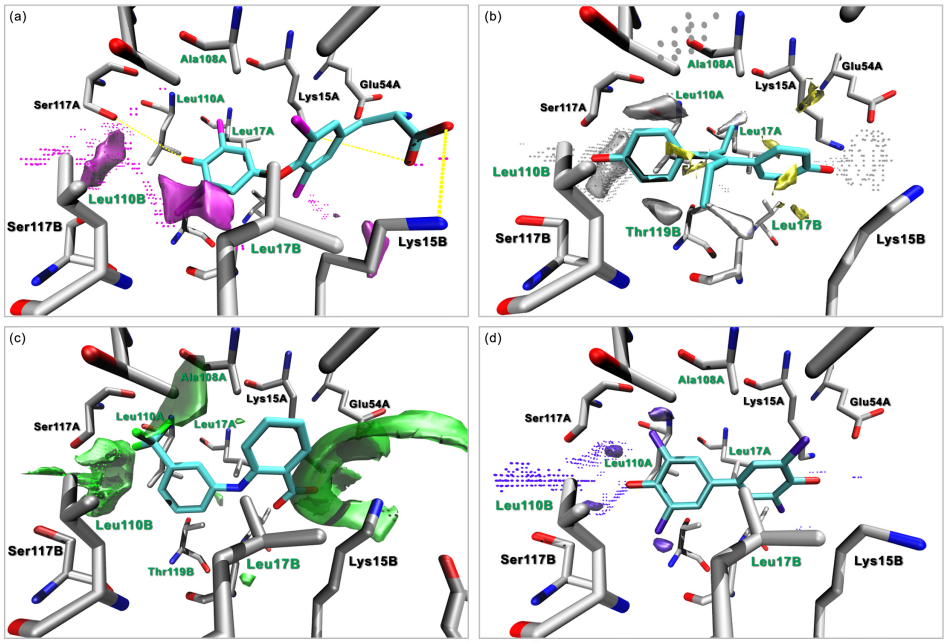


Figure 3.7. MIF analysis of TTR binding sites in four model complexes with several probes computed by the program GRID. (a) TTR-T_4 complex versus the iodine probe: while strongly favourable interaction fields are ubiquitous across the sites, the strongest region for interaction (countered as solid surfaces in magenta at an interaction energy of -14 kcal.mol^{-1}) map well onto an inner halogen binding pocket that is occupied by one of the iodine atoms of T_4. Hydrogen bonds and polar interactions are indicated as yellow broken lines. (b) TTR-DES complex versus the methyl (C3) and aromatic (DRY) probes: while the strongest sites of interaction with DRY (countered as yellow solid surfaces at an interaction energy of -1.8 kcal.mol^{-1}) map well onto the regions where the aromatic rings of DES sit, the most favourable sites of interaction with C3 (countered as silver solid surfaces at an interaction energy of -4 kcal.mol^{-1}) highlight regions which are only dimly explored by the ethyl groups of DES. (c) TTR-FLU complex versus the fluorine probe: favourable interaction fields (countered as solid green surfaces at an interaction energy of -3.5 kcal.mol^{-1}) are found at the inner and the outer ends of TTR pockets, where the Ser-117 residues and the Lys-15 residues respectively sit. While the most prominent surface maps well onto one carboxyl oxygen of FLU near the Lys-15 residues, the inner fields highlight the lack of a hydrogen bond acceptor to interact with the hydroxyl side-chains of Ser-117 residues. (d) TTR-PCB complex versus the chlorine probe: favourable interaction fields (countered as solid violet surfaces at an interaction energy of -7 kcal.mol^{-1}) are found near the regions where the two chlorine atoms attached to the PCB's inner ring sit. In all panes, the small dots represent disfavoured sites of interaction, from which a probe atom would be excluded because water would interact better. This figure was produced using the perspective view mode of VMD to provide better notion of depth in 3D.

Figure 3.8 provides a brief summary of the most relevant energy minima detected within TTR binding sites, contrasting the positioning of the four model ligands along the different panes. For a detailed list of all minima and respective energy values, see **Table B.1** of the Appendix.

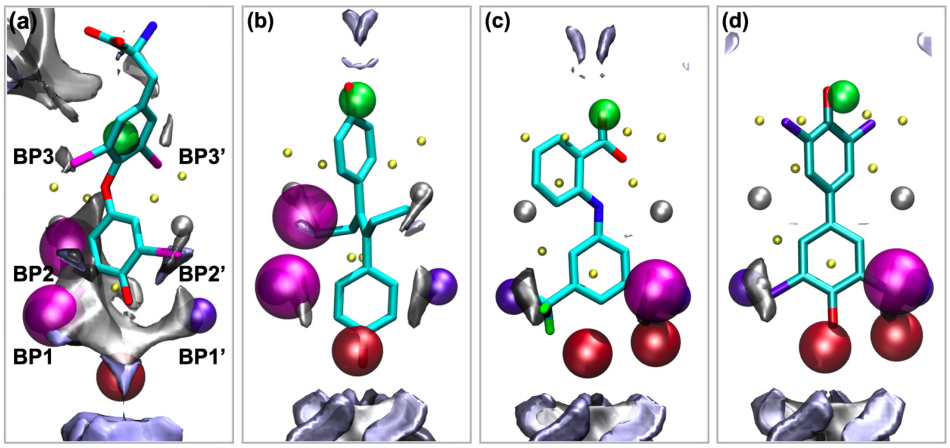


Figure 3.8. Three-dimensional representation of the affinity grids' energy minima for various probe types (yellow for aromatic, silver for methyl, red for phenolic hydroxyl, green for fluorine, violet for chlorine and magenta for iodine). The minima were populated by annealing successive probes on each grid map, and only those that are energetically appropriate for interaction with a probe of its own type were retained. The sphere radii are proportional to the respective binding energies, which are reported in **Table B.1**. The grid maps and their minima were computed for each X-ray structure (after striping the ligand from the structure): (a) T_4, resolved at 2.00 Å (PDB entry 2rox); (b) DES, determined at 1.80 Å (PDB entry 1tt6); (c) FLU, solved at 2.00 Å (PDB entry 1bm7); and (d) PCB, determined at 1.80 Å (PDB entry 2g5u). Grid isocontours for the water (polar) probe and the methyl (sticky) probe are displayed in ice blue and silver solid surfaces, respectively.

The knowledge (and the exact positioning) of the predicted minima was proven useful during the modelling stage of the pharmacophore queries presented below. As illustrated in **Figure 3.9**, the fact that all GRID calculations were performed on structurally aligned complexes, using a common reference structure (PDB entry 2g9k), has enabled the visual mapping of the energy minima grid points inside TTR binding sites, thus leveraging the combination of receptor-based pharmacophore perception with ligand-based perception.

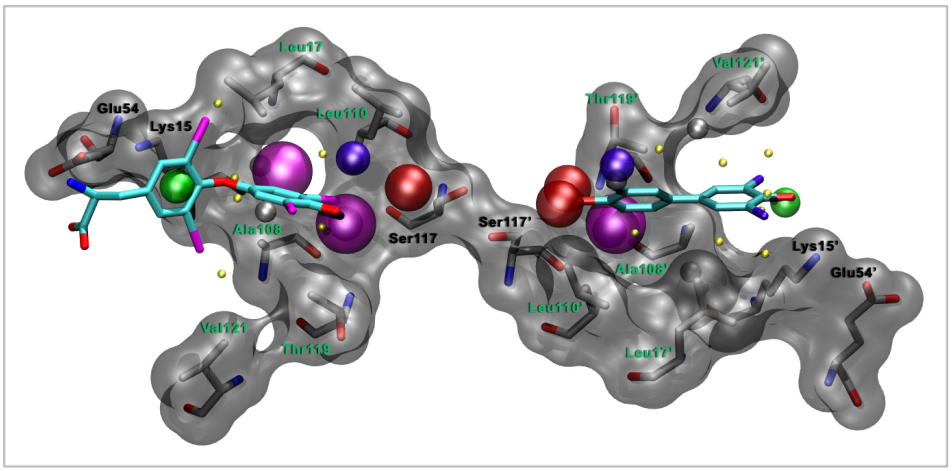


Figure 3.9. Representation of TTR binding sites along with their predicted *hot spots* for interactions with fluorine (green), chlorine (purple), iodine (magenta), hydroxyl

groups (red), aromatic groups (yellow) and methyl (silver). The size of the spheres corresponds to the magnitude of the interaction, according to the values computed by the programs GRID, MINIM and FILMAP for the energy minima. On the left binding site, the structure of thyroxine is shown in sticks; on the right binding site, the structure of PCB18, the strongest TTR binder known to date, is represented.

3.2. Ligand-based pharmacophore modelling

In this subsection, we present the results of ligand-centric analyses carried out to characterise in detail the interaction between TTR and the known stabiliser compounds.

3.2.1. X-ray ligand mapping and clustering

Ligands extracted from the 21 high-quality TTR complexes were subjected to an all-atom comparison: atom coordinates of any two specified ligands (ligand A and ligand B) were compared. Atom coordinates found to be identical or lying within 1 Å difference in the two ligands were considered a "hit" and the total number of hits was used to calculate Tanimoto Indexes (T.I.) to quantify the spatial relationship between the two ligands. The tanimoto indexes for all ligands bound to TTR binding site *AC* and binding site *BD* were calculated and combined to produce similarity score matrices for each pocket. **Figure 3.10** illustrates a similarity score matrix produced for TTR binding site *AC*.

3.2.2. Cluster analysis

The similarity matrices produced for each TTR binding site were then clustered to identify underlying groups of ligands that occupied similar geometrical positions in the binding sites. This was accomplished using the CLUTO tool [454], described in Chapter 2. The rational behind this approach is that ligands with similar binding site location are able to establish similar receptor site interactions. **Figure 3.10** shows the clustering dendrogram obtained for the similarity score matrix of ligands bound to TTR binding site *AC*. Three main clusters were identified on this basis: cluster A (PDB entries 1e4h, 1dvu, 1dvy, 2rox, 1dvt and 1tt6), cluster B (1dvx, 1bm7, 1dvz, 1y1d, 2f8i and 2f7i) and cluster C (2qgd, 2gab, 2b77, 2g9k, 2g5u, 1dvs and 2qge). Given the fragment-like nature of ligands corresponding to the last two rows (3b56 and 2b15), their structures were considered outliers of this analysis. The exact same analysis was applied to ligands bound to TTR binding site *BD*, yielding the following clusters: 2qgd, 2f8i, 2f7i, 1tt6 and 1dvs (cluster D) and 2qge, 2g9k, 2g5u, 2gab, 1dvz and 1bm7 (cluster E).

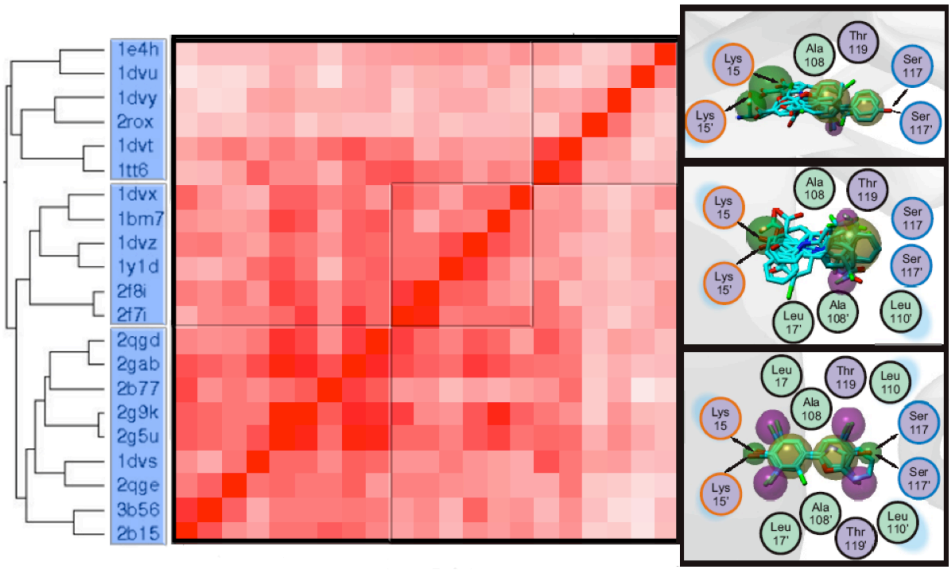


Figure 3.10. Illustration of a similarity score matrix for the spatial occupation of ligands bound to TTR binding site AC, along with the cluster analysis. The coloured map provides an indication of the degree of similarity between the ligands in terms of their positioning within TTR binding site: from white to red with increasing similarity. The dendrogram represented on the left of the matrix shows the result of the clustering procedure, disclosing three distinct clusters of ligands exploring slightly different regions of site *AC*.

Each of the five clusters of ligands identified in this subsection was used to devise five pharmacophore models for TTR, as explained in the following subsection.

3.3. Selectivity analysis: evaluation of multiple pharmacophore queries

Five different pharmacophore queries (QRY) were devised for TTR binding sites by combining all data resulting from the analysis of (*i*) structural information on TTR receptor sites, (*ii*) MD-derived information on TTR receptor flexibility, (*iii*) *hot spots* within TTR binding sites, and (*iv*) ligand mapping information. Three queries were generated for the *AC* binding site using information from each cluster identified amongst the 21 known TTR active compounds (training sets). **Figure 3.11**, **Figure 3.12** and **Figure 3.13** show the three final models developed for the *AC* binding site. Likewise, two queries were created for the *BD* binding site using information from two clusters (**Figure 3.14** and **Figure 3.15**). A detailed description of all models and their global performance, including information of applied constraints, is provided in **Table 3.3**.

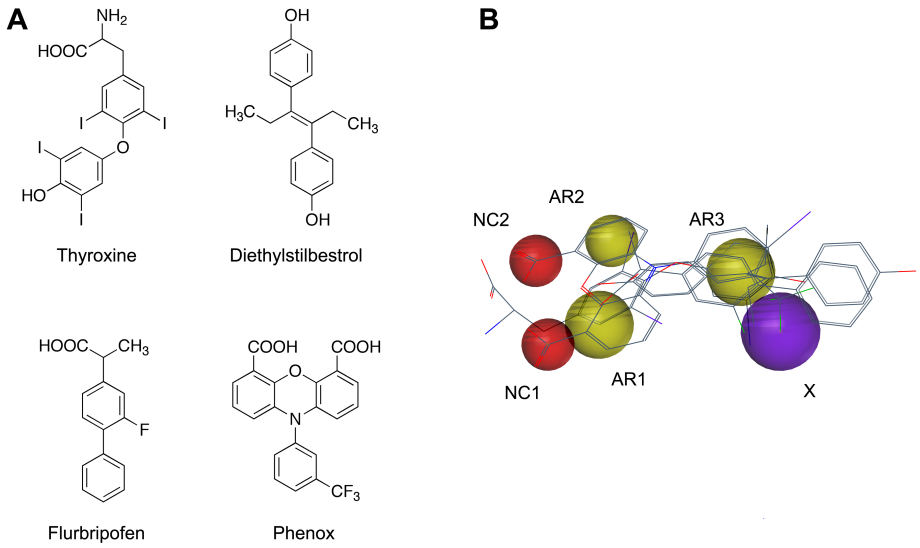


Figure 3.11. UNITY 3D pharmacophore query QRY AC1 generated with SYBYL for TTR pocket AC. (A) Training set of TTR amyloid inhibitors comprising cluster A. (B) 3D model of the common pharmacophore features. The alignment of the ligands resulted from the structural alignment of the respective TTR complexes with GH8. The colour of the spheres represents the type of pharmacophore feature, whereas their radius corresponds to the applied spatial tolerance: aromatic (AR, yellow), halogen atom (X, violet) and negative charge centre (NC, red). For the sake of clarity and ease of visualisation, receptor site and exclusion volume spheres are not displayed.

By visual analysis of the models and respective underlying ligand alignments, it can be recognized that the mapping of overlapping pharmacophoric features was a demanding task. In fact, except for cluster C, none of the automated routines within SYBYL was able to suggest queries that could reasonably capture the specificities of the alignments. Accounting for this is not only the high structural diversity found amongst the known inhibitors, but also the fact that the superimpositions of the ligands resulted from the structural alignment of the TTR complexes (and not from an automated procedure exploring different ligand conformations to bring similar functional groups into proximal locations, such as in the case of the GALAHAD module of SYBYL). Consequently, the different pharmacophoric features and constraints were manually assigned onto the alignments and calibrated in order to obtain the best possible enrichments with the test set. As can be seen in **Figure 3.12** and **Figure 3.13**, the applied spatial tolerances (equivalent to the sphere radii) are higher than commonly employed in this type of 3D models. It is also worth noting that, despite sharing features that are known to be essential for binding at the TTR receptor, the AC and BD models present distinctive pharmacophore arrangements reflecting the alternative binding modes observed for the two pockets in the X-ray complexes.

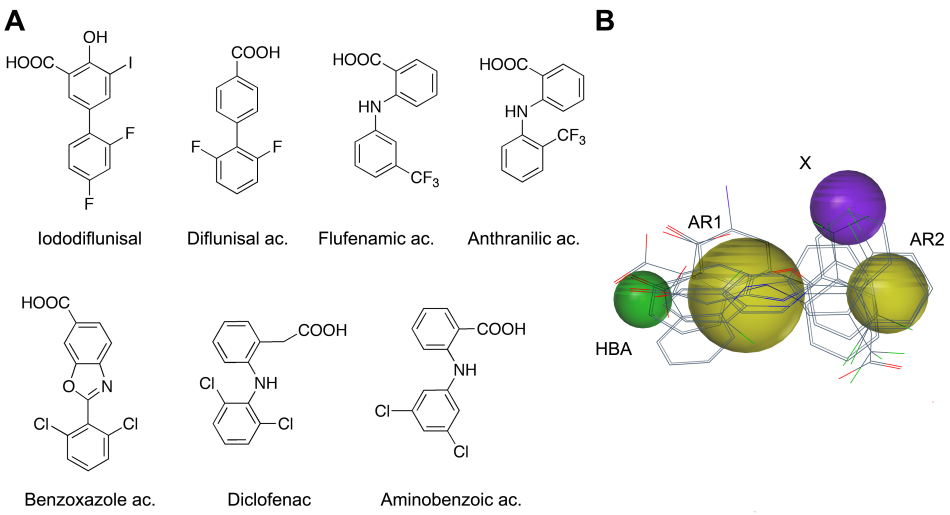


Figure 3.12. UNITY 3D pharmacophore query QRY AC2 generated with SYBYL for TTR pocket AC. (A) Training set of TTR amyloid inhibitors comprising cluster B. (B) 3D model of the common pharmacophore features. The colour of the spheres represents the type of pharmacophore feature, whereas their radius corresponds to the applied spatial tolerance: aromatic (AR, yellow), halogen atom (X, violet) and hydrogen-bond acceptor (HBA, green). For the sake of clarity, receptor site and exclusion volume spheres are not displayed.

As mentioned in the introductory section, since UNITY 3D searches retrieve/discard ligands in the database on a pass/fail basis (without scoring or ranking), the virtual screening performance of these pharmacophore queries was assessed by an overall measure of enrichment. Although this hinders the direct performance comparison with other VS methods that perform ranking of compound databases, the enrichment values provide an indication of the specificity of each query. A detailed explanation about ligand enrichment is provided in Chapter 5. Here, we used a *global* ROC Enrichment metric to quantify the specificity of the models (**Table 3.3**), which expresses the fraction of active molecules retrieved by one given model as a proportion of the fraction of decoy molecules retrieved by that model from the entire test data set (containing both actives and decoys). It has been shown that, to hold a reasonable chance of identifying true hits, VS methods/models should provide at least a 10-fold increase in enrichment [496].

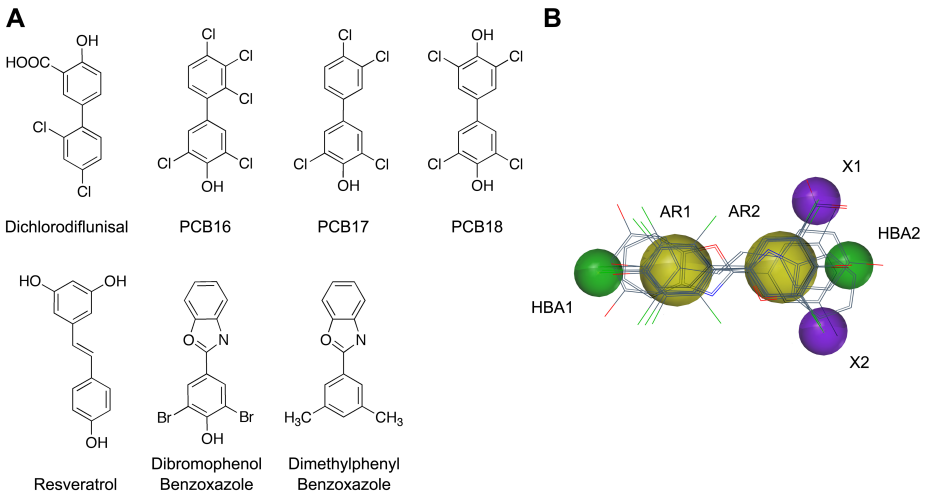


Figure 3.13. UNITY 3D pharmacophore query QRY AC3 generated with SYBYL for TTR pocket AC. (A) Training set of TTR amyloid inhibitors comprising cluster C. (B) 3D model of the common pharmacophore features. The colour of the spheres represents the type of pharmacophore feature, whereas their radius corresponds to the applied spatial tolerance: aromatic (AR, yellow), halogen atom (X, violet) and hydrogen-bond acceptor (HBA, green). For the sake of clarity, receptor site and exclusion volume spheres are not displayed.

In most cases, the models are able to retrieve all of the actives contained in their respective training set. However, it should be noted that only distinct chemotypes were integrated into the test set (to avoid redundancy as possible), causing a slight decrease in the enrichment values. Queries AC2 and AC3 offer a fair representation of the global pharmacophore of ligand binding to TTR, characterized by two aromatic rings, one hydrogen bond acceptor feature at the outer ring (to pair with one or two side chains of Lys-15 residues), and at least one halogen substituent at the inner ring. QRY AC3 includes a hydrogen bond acceptor feature at the inner ring, accounting for ligands that can interact with Ser 117. Both queries retrieve a high number of false positives, disclosing a low specificity of the models (enrichment values of 7.1 and 3.8 for QRY AC2 and QRY AC3, respectively).

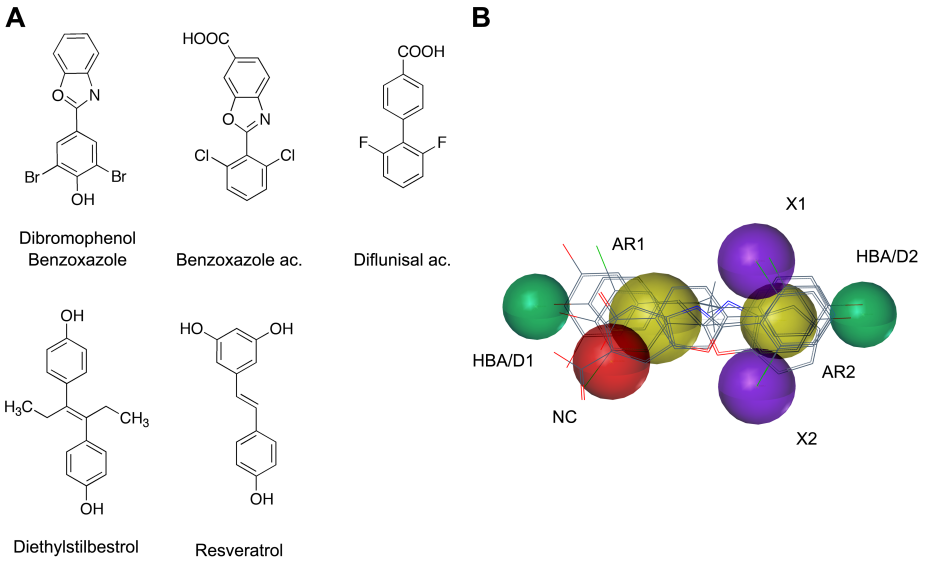


Figure 3.14. UNITY 3D pharmacophore query QRY BD1 generated with SYBYL for TTR pocket BD. (A) Training set of TTR amyloid inhibitors comprising cluster D. (B) 3D model of the common pharmacophore features. The colour of the spheres represents the type of pharmacophore feature, whereas their radius corresponds to the applied spatial tolerance: aromatic (AR, yellow), halogen atom (X, violet), hydrogen-bond acceptor and/or donor (HBA/D, green), and negative charge centre (NC, red). For the sake of clarity, receptor site and exclusion volume spheres are not displayed.

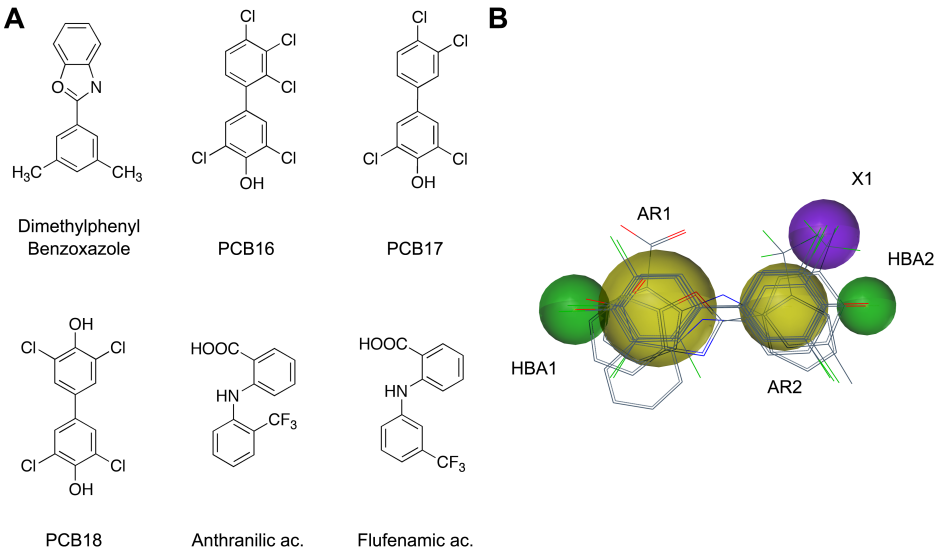


Figure 3.15. UNITY 3D pharmacophore query QRY BD2 generated with SYBYL for TTR pocket BD. (A) Training set of TTR amyloid inhibitors comprising cluster E. (B) 3D model of the common pharmacophore features. The colour of the spheres represents the type of pharmacophore feature, whereas their radius corresponds to the applied spatial tolerance: aromatic (AR, yellow), halogen atom (X, violet) and hydrogen-bond acceptor (HBA, green). For the sake of clarity, receptor site and exclusion volume spheres are not displayed.

Except in QRY AC1, which contains three aromatic features, the two aromatic rings stands as the core feature in all models. As mentioned early in the introductory chapter of this thesis, the biphenyl motif is a central determinant in ligand binding to TTR. A statistical analysis of NMR-derived binding data on 11 protein targets has revealed that the biphenyl moiety is a preferred substructure for protein binding [497]. We believe that the low specificity of our models can be linked not only to an inherent low specificity of the (highly symmetric) TTR receptor sites but also to the lack of an appropriate number of specific pharmacophoric features around the biphenyl motif (to disrupt its symmetry). QRY AC1, however, offers a better compromise between sensitivity and specificity, achieving the highest enrichment (33.1) and retrieving some of the strongest TTR amyloid inhibitors: phenox, all dibenzofuran derivatives, flufenamic acid and a diflunisal analogue.

Table 3.3. Description of 5 pharmacophore queries based on 3 clusters of ligands bound to pocket AC of TTR and 2 clusters of ligands bound to pocket BD. A *global* ROC enrichment is reported for each query.

	Template ligands (PDB IDs)	Hypothesis description[(a)]	# feats.	Constraints	# hits	True hits	ROCE
Query				***TTR pocket AC***			
QRY AC1 **Figure 3.11**	2rox, 1tt6, 1dvt, 1dvy (cluster A)	NC1-NC2-AR1-AR2-AR3-X Exclusion spheres Receptor spheres	6	Distance constraints: AR1-AR2 (5.14 +/- 1.00 Å) AR1-AR3 (5.42 +/- 1.00 Å) AR2-NC2 (2.95 +/- 1.00 Å) Partial match constraint: NC1/NC2 and AR1/AR2 (min. 2 and max. 4 matches)	12	6	33.1
QRY AC2 **Figure 3.12**	1y1d, 2f7i, 1bm7, 1dvz, 2f8i, 1dvx, 1u21 (cluster B)	HBA-AR-AR-X Exclusion spheres Receptor spheres	4	Distance constraint: AR1-AR2 (1.81 +/- 1.00 Å)	34	6	7.1
QRY AC3 **Figure 3.13**	1dvs, 2qgd, 2qge, 2b77, 2gab, 2g9k, 2g5u (cluster C)	HBA-AR1-AR2-X1-X2-HBA Exclusion spheres Receptor spheres	6	Distance constraint: AR1-AR2 (4.30 +/- 1.50 Å) Partial match constraint: X1/X2 (min. 1 and max. 2 matches) Bond path constraint: AR1-AR2 (min. of 1 and max. of 2)	49	5	3.8
				Combination of the 3 AC queries	77	13	6.7
Query				***TTR pocket BD***			
QRY BD1 **Figure 3.14**	2qgd, 2f8i, 2f7i, 1tt6, 1dvs (cluster D)	HBD-NC-AR1-AR2-X1-X2-HBD Exclusion spheres Receptor spheres	7	Distance constraint: AR-AR (4.78 +/- 1.00 Å) Partial match constraint: HBD1-NC-X1-X2-HBD2 (min. 2 and max. 5 matches)	30	9	14.2

	Template ligands (PDB IDs)	Hypothesis description[a]	# feats.	Constraints	# hits	True hits	ROCE
QRY BD2 **Figure 3.15**	2qge, 2g9k, 2g5u, 2gab, 1dvz, 1bm7 (cluster E)	HBA-AR1-AR2-X-HBA Exclusion spheres Receptor spheres	5	Distance constraint: AR-AR (4.25 +/- 1.00 Å)	19	8	24.1
				Combination of the 2 BD queries	37	10	12.3
Winning combinations							
QRY AC1 / QRY BD1					42	14	16.5
QRY AC1 / QRY BD2					30	12	22.1

[a] The features are described in sequence from the external region to the internal region of TTR pockets: aromatic (AR), halogen atom (X), hydrogen-bond acceptor (HBA), hydrogen-bond donor (HBD) and negative charge centre (NC).

Summing up, three out of the five pharmacophore models derived for TTR binding sites offer an acceptable (or even good) VS performance, with more than 10-fold increase in enrichment. The combination of queries devised for pocket AC and for pocket BD is clearly a desirable approach to follow, given our interest in identifying compounds to target both TTR binding sites. We have tested all possible combinations between AC and BD queries and identified two winning combinations, retrieving more than half of the actives in the test set and mostly corresponding to the strongest inhibitors. As will be described in Chapter 7, along with the individual queries, these combinations can be used as a post-screening tool to assess the compliance of VS hits (identified by several VS protocols) with the postulated pharmacophoric features for TTR binding.

4. Concluding remarks

In the work presented in this chapter, several analyses were performed towards an exhaustive characterisation of TTR binding sites, in an attempt to identify consistent structural and/or physicochemical differences between its two binding sites. A careful analysis of the X-ray B-factors deposited in the PDB for several TTR complexes offered insights about receptor site flexibility based on experimental evidence, which were complemented by the results of MD simulations of the TTR tetramer. Receptor-based pharmacophore perception was further aided by the use of algorithms for the detection of energy minima (*hot spots*) of favourable interaction with multiple probes within TTR binding sites. While the attempt to attain a tangible differentiation of the sites was mostly unsuccessful at the receptor level, ligand mapping followed by clustering analyses has evidenced systematic differences in ligand binding modes across the two TTR binding sites. The differential positioning (and alignment) of ligands in the two sites were captured into an ensemble of five pharmacophore models characterising the determinants of ligand binding to TTR. The models incorporate knowledge on the shape of TTR binding

sites in the form of exclusion volume spheres, at the cost of a high computational burden during the searches.

Limited chemical space coverage, due to the small size of the training sets used to derive the models, is considered today an important problem of ligand-based approaches. In this chapter we attempted to address this problem by combining multiple models generated from multiple training subsets comprising several TTR stabilisers with differential binding modes on TTR's binding sites. Selectivity analysis revealed poor selectivity for two pharmacophore models, but three models and two winning combinations (including models derived on both TTR binding sites) offered promising enrichment in the strongest TTR stabilisers. Further validation of the models will be carried out in Chapter 7. While the use of our receptor- and ligand-based pharmacophores in the screening of very large libraries of compounds may be hindered by the computing time costs associated with UNITY 3D searches, we regard our models as a promising post-screening tool for prioritising VS hits for biological evaluation.

In this chapter, my self-criticism goes to what I regard as an abridged use of the magnificent amount of information provided by MD simulations. A careful analysis via clustering, focused on TTR binding sites throughout the MD trajectories, either in the apo or bound forms of the protein, could have helped putting forward a set of previously-undetected conformations and provided a more dynamic picture of the TTR pharmacophore. Unfortunately, time is always a strong opponent and priorities need to be set. Without doubt, however, the experimental evidence provided by the X-ray complexes of TTR clearly suggests a more or less inert/static structure of the funnel-shaped binding sites, with minor conformational changes taking place on the side chain groups of no more than three residues. It is therefore reasonable to believe that, albeit interesting from an academic viewpoint, such analysis would not have a direct impact in the modelling of our pharmacophore queries of TTR.

5. Acknowledgements

My first acknowledgement goes to Trishna Mukherjee for her extraordinary dedication throughout the development of the pharmacophore models for TTR, but also for her resilience before the difficulties encountered throughout. Trishna was the main responsible for the Java script for ligand mapping. Acknowledgements are also due to Dr. Sean Killen for the installation of Tripos software. I thank Dr. Gunther Stahl and Dr. Ulrike Uhrig for technical and scientific support throughout this project, and Dr. Yohan Boursereau for his sympathy during the process of software acquisition and training in Paris.

Chapter 4

Molecular similarity and similarity search methods: surfing the chemical space

"(…) But I'm an optimist. We will reach out to the stars."

[Stephen Hawkins, in *The Daily Telegraph*, October 2001]

1. Introduction and theory

Fundamentally, every small molecule-based strategy to either identifying or designing a novel active compound focuses on the exploration of *molecular similarity*, albeit often from different viewpoints. For example, pharmacophore analysis and QSARs concentrate on *local* similarities when studying the molecular determinants of biological activity, such as functional groups and their specific geometric arrangements and/or resulting chemical properties. The modern understanding of molecular-similarity analysis stems from the *Similar Property Principle*, which states that molecules that are generally similar should display similar biologic activity. Even though the concept of similarity and comparison is ordinary to humans, a formal definition with unambiguous criteria and constraints is required to allow automatic decision-making. Equally, only through a rigourous description of molecules can meaningful comparisons be made.

Molecular descriptors are typically grouped in the following fundamental categories:

- 1D-representation: physicochemical properties, molecular codes, linear descriptors, etc.
- 2D-representation: connection tables, topological descriptors, (sub)graphs, (sub)structure descriptors, digital imaging processing, etc.
- 3D-representation: molecular shapes, quantized surface complementarity diversity (QSCD), molecular quantum similarity (MQS) and, in general, molecular surfaces and volumes.
- Merging representations: fingerprints, pharmacophores, and other merging approaches.

Chemical diversity increases with the growing number of molecules daily discovered or synthesized. Therefore, determining an underlying order and defining classes are demanding tasks. Several methods exist today to classify molecules, descriptors and even databases. They are generally grouped into two main categories:

- Supervised classification methods: neural networks, maximum likelihood classification (MLC), support vector machines (SVM), k-nearest neighbours (k-NN) classifiers, etc.
- Unsupervised classification methods: clustering, partitioning, self-organising maps (SOMs), principal component analysis (PCA), etc.

The analysis of molecular databases involves the calculation of (dis)similarities between all molecule pairs comprised in a data set. For practical reasons, it is sometimes useful to consider a reduced group of molecules, representatives of the whole population. Generally, the measures of similarity and/or diversity involve three main components: the descriptors (exemplified above), the coefficients and the weighting schemes. The similarity coefficients (or indexes) are functions that transform pairs of compatible molecular representations into real numbers (typically lying on the unit interval), while the weighting schemes are used to assign different levels of importance to the various components of the molecular representations.

In this chapter, the reader is introduced to the concepts of *chemical space* and molecular-properties (and descriptors) space, and to ways of mapping them and restricting them to biologically relevant subspaces. One- and two-dimensional similarity search methods are explained as tools to *surfing* the chemical space in pursuit of molecules with lead- and drug-like properties, safe ADMET profiles, and both ability and inability to bind TTR. It will be explained how the knowledge extracted from the known TTR stabilisers was used to select a set of decoy molecules (i.e. molecules with high likelihood of being inactive) and thus build a *benchmarking set* for the evaluation of VS methods against TTR amyloid. It will also be shown how the physicochemical properties of the active TTR amyloid inhibitors were integrated as filtering criteria leading to the assembly of a tailored *screening set* of approximately 2.3 million compounds for virtual screening against TTR amyloid.

1.1. Chemical libraries

Chemical space can be compared to cosmic space, where chemical compounds, instead of celestial objects, reside in space. The number of carbon-based compounds that populate chemical space has been estimated to exceed 10^{60} [498]. However, most of this "space" is biologically irrelevant. Restricting the number of compounds that are searched to biologically relevant and synthetically accessible molecules is a field of active research. Rarely, and often by serendipity rather than design, "stars" that can modulate biological processes have been found.

Virtual screening circumvents the problem of broad searches in chemical space by confining itself to libraries of specific, accessible (purchasable) compounds. There were times wherein an important barrier to entry into structure-based virtual screening was the lack of a suitable, easy to access database of purchasable compounds. Yet, nowadays, several resources have emerged that ease the access to compounds in the shelves of chemical suppliers around the world.

Several databases of already existing compounds have been released over the last decade, such as PubChem [136], ACX [499] and the NCI Open Database [500]. In 2005, Irwin et al. prepared a database of commercially available molecules, each with an associated 3D structure, using catalogues of compounds from multiple vendors [137]. By that time, the size of the library was no larger than 727,842 molecules. Nowadays, ZINC has become a truly established database that contains over 13 million purchasable compounds. All molecules have been assigned biologically relevant protonation states and are annotated with properties such as molecular weight, calculated LogP, and number of rotatable bonds. Moreover, each molecule in the library contains vendor and purchasing information, and is ready for docking using several popular docking programs.

Making use of its database facilities and compound annotations, ZINC also offers access by several subsets, assembled by following various criteria. Subsets can be defined as catalogs, with each subset corresponding the compound collection of one particular vendor. Perhaps more importantly, subsets can also be defined by physical properties. Predefined subsets available for download include fragment-like, lead-like and drug-like compounds, loosely following the filtering criteria proposed by Teague et al. [501], Carr et al. [128] and Lipinski [502], respectively. As explained by Irwin et al., these criteria have been deliberately relaxed to include a number of molecules at the periphery of what many investigators might consider desirable [137]. By doing so and allowing a small number of violations of Lipinski's rules, they accommodate for the uncertainties in the calculated property values.

The ZINC database is, as of October 2011, on its eleventh version and is available for free download in several common file formats including SMILES, MOL2, 3D SDF, and DOCK flexibase format (http://zinc.docking.org).

1.2. Library filtering: *surfing* the physicochemical-properties space

Filters may be applied to ensure that the chemical libraries used for screening meet some standards of biological relevance or drug-likeness. At this stage, molecular similarity is mostly sought at the level of one-dimensional descriptors (also known as bulk descriptors) that commonly associate a compound to a single value (describing a physicochemical property like its molecular weight, dipole moment, etc.). The filtering and assembly of a chemical library is indeed a very important step in any screening campaign, including the ones carried out "in silico". It dictates, amongst other things, what will be the extent of chemical space covered during the screenings and also how much (computing) time

will take to screen an entire library. Even though often overlooked, the choice of criteria during the filtering procedure is critical to ensure that no important chemical entities are missed out during the database filtering. Different types of filters are suitable under distinct situations.

In the early stages of a project, when little or no SAR is accessible, stringent drug-like filters are often employed. This precludes spending of chemistry resources pursuing intricate compounds that may not be customizable by adding desirable properties. Nevertheless, when considering compounds to acquire for HTS, different filters can be applied. As denoted by Oprea, the best molecules for early HTS campaigns are smaller and less functionalised than drugs, yet retaining some activity [503]. Thus, strict lead-like filters are used to guarantee that the HTS hits have enough *room* for expansion into larger and more functionalised leads. On the other hand, when SAR implies that certain compounds (or series) may produce valuable information, the filtering criteria are usually loosened, because the secondary screens (e.g. QSAR models, similarity to known actives) that are being applied are effective in detecting useful compounds.

Library filtering is only possible when appropriate descriptions of the molecular features of database compounds are available. Commonly used chemoinformatics packages like the Chemistry Development Kit (CDK) [504], CODESSA [505], Dragon [506], MOLCONN [507], Molecular Operating Environment (MOE) [508], Pipeline Pilot [509], and SYBYL [457] provide a multitude of molecular descriptors that span several descriptor classes. However, although molecular descriptors comprise a wide range of 1D, 2D and 3D features, from straightforward heavy atom counts to more complex fingerprints and shape representations, the notion behind library filtering is usually tied to simple (and computationally inexpensive) one-dimensional descriptions of molecules.

Amongst the most relevant one-dimensional descriptors are the molecular weight (MW), the number of hydrogen bond acceptors and donors, the octanol-water partition coefficient (logP), and the number of rotatable bonds. Albeit rudimentary in nature, a number of examples have shown how these physicochemical descriptors alone can be extremely useful in drug design and development. In 1997, Christopher Lipinski famously derived a set of principles based on the observation that most orally active drugs are small and lipophilic molecules [510]. These principles describe the molecular properties that are critical for an adequate pharmacokinetic profile in the human body (namely absorption, distribution, metabolism and excretion). Lipinski defined a rule that states that, in general, an orally active drug holds no more than a single violation of the following criteria:

- ≤ 10 hydrogen bond acceptors (nitrogen or oxygen atoms);

- ≤ 5 hydrogen bond donors (nitrogen or oxygen atoms with one or more hydrogen atoms);
- MW < 500 Da;
- logP ≤ 5.

This *rule of thumb* soon became known as Lipinski's Rule-of-Five, given the fact that all numbers are multiples of five. Even though there are many exceptions to the rule, it still represents one of the most well-known and applied indicators of *drug-likeness*. Over the last decade, further investigations have spawned many extensions to the original rule [502,511,512].

In this chapter, the program FILTER [513] was used to compute several molecular descriptors within two large libraries of virtual compounds. First, the program aided the selection of molecules holding physicochemical properties matching those of known TTR stabilisers during the construction of a benchmarking set. Then, FILTER was used to trim down a library of more than 11 million compounds, based on a combination of (*i*) a set of empirical rules for drug-likeness and bioavailability (pharmacokinetics), (*ii*) knowledge of toxic and reactive functional groups, dyes and aggregator molecules, and (*iii*) knowledge of the physicochemical properties of known TTR amyloid inhibitors. A detailed explanation of the molecular descriptors and the similarity search methods employed in the subsequent chapters of this thesis is provided in this chapter.

1.3. 2D searches in chemical space

While one-dimensional descriptors may be adequate to evaluate the drug-likeness of a molecular database, they are intrinsically related to the molecular structure and do not depict single specific (sub)structural components. Therefore, they are not suitable to capture the essence of molecular functions required for biological activity. Conversely, and as explained in the introductory chapter, two-dimensional (2D) descriptors encode topological information between the atoms of a molecular graph as binary vectors, often referred to as fingerprints (see **Figure 1.8**). Most 2D descriptors can be calculated rapidly, allowing hundreds of thousands of structures to be processed in an hour.

The fundamental idea underlying similarity-based VS is simple and was originally enunciated by Johnson and Maggiora based on their (aforementioned) Similar Property Principle [104]. If this principle holds, database compounds that are structurally similar to a reference molecule but have not been tested for biological activity are more likely to be active. The ranking of a database in order of decreasing similarity was first reported in

two papers published in the 1980s [514,515], both centred on the use of 2D fingerprints for the measurement of inter-molecular structural similarity. This strategy was quickly taken up and is now supported by a large body of evidence [515–522].

Over the past decades, several 2D fingerprint technologies have been developed and successful used, including hashed connectivity pathways [108], structural dictionary-based [109] and layered atom environment fingerprints [110]. Daylight fingerprints, in particular, have been used as a standard for benchmarking [108]. Indeed, compared to more sophisticated descriptors (like 3D pharmacophore patterns), simple 2D fingerprints persist as representations of choice for similarity-based VS. This is due not only to its high computational efficiency, but also to its demonstrated effectiveness against more complex approaches to similarity searches [120,516,517,523–528].

In this subsection, fingerprints methods that have been exploited throughout this PhD project will be described comprehensively.

1.3.1. 2D chemical fingerprints

A Chemical Hashed Fingerprint (CHF) is a bit string encoding the topological structure of a small molecule [529]. Atoms and bonds (linear patterns) and rings (cyclic patterns) are detected for each ligand and assigned to strings of a given length that can be used to characterize the molecule for full structure, substructure and similarity searches. The length of the strings (and thus their resolution) is defined by the user. The first step of the descriptor generation corresponds to the creation of the atom pair descriptor and the so-called start-end-short-path (SESP) vector descriptor – wherein all paths of a given length are counted using the shortest path algorithm [514,530]. The atom types are defined by the elements, and the possible walks between all atoms on the molecular graph are calculated and encoded as a binary vector. Thus, each possible walk is represented by a bit pattern that is added to the descriptor with a logical OR operation. By doing so, overlaps between different bit patterns are possible (recall **Figure 1.8**).

Similarly to CHF, 2D UNITY Fingerprints (or UNITY 2D-FP) are binary representations of the common features and groups in a structure (such as all fragments of length 2-7, phenyl rings, etc.). These fingerprints are specified in simple ASCII files, by following a set of rules pre-defined in a screen definition file within the program SYBYL [457]. Two types of definitions are included in a UNITY 2D screen definition file – BY_SLN and BY_LENGTH – dictating the class of fingerprint to be used – Fragment Based Fingerprints (FBFP) and/or Fragment Length Based Fingerprints (FLBFP), respectively. On a nutshell, FBFP simply characterise a molecule based on the particular fragment types it contains,

whereas FLBFP take into consideration aspects that help optimising the rules for the length of the typical fragments that are used in queries. This is particularly important for screening purposes.

1.3.2. 2D pharmacophore fingerprints

By contrast to CHF and UNITY 2D-FP, 2D-Pharmacophore Fingerprints (2D-PF) take into account only the chemical features related with binding and believed to be responsible for a particular pharmacological action, which are defined as follows: hydrogen bond donor (d); hydrogen bond acceptor (a); positive charge (+); negative charge (–); hydrophobic (h), and aromatic (r) [531]. Moreover, instead of a one-dimensional construct where no spatial relations are included (like in a bit string description), the relative position of pharmacophore points is represented by a topological distance (**Figure 4.1**). This distance is defined by the length of the shortest path between two nodes (atoms) of the chemical graph, i.e. the smallest number of graph edges (bonds) connecting the two atoms. As for CHF and UNITY 2D-FP, similarity searches using 2D-PF do not require the computation of 3D coordinates and were shown to be very fast and efficient for the virtual screening of large databases.

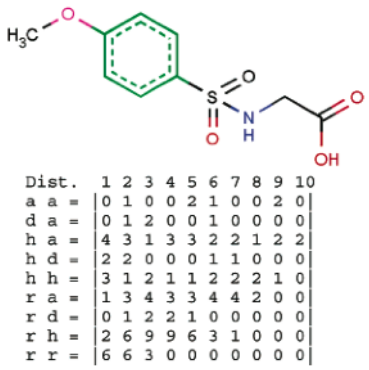


Figure 4.1. Example of a 2D-PF descriptor histogram. Dist.: Topological distance, a: hydrogen bond acceptor, d: hydrogen bond donor, h: hydrophobic, r: aromatic. Adapted from reference [531].

1.3.3. Similarity search metrics

At the heart of any system for similarity-based VS is the metric that is used to quantify the degree of resemblance between the reference molecule and each of the structures in the database of compounds to be screened.

Focusing more on the similarity coefficients that are employed for comparing fingerprints, the Tanimoto coefficient (Tc) is the metric of choice in both in-house and commercial software for chemical information management. However, the Tc is also associated with a few problems. One of its shortcomings is that it is known to yield low similarity values when the reference molecule has just a few bits set in its fingerprint [532]. Moreover, this has been linked to an inherent bias towards specific similarity values [533].

Several alternative metrics have been explored in order to address the limitations of the Tc, and Holliday et al. conducted a comparison of 22 coefficients in searches of the NCI AIDS database using 2D UNITY fingerprints [534]. **Table 4.1** provides examples of some of the coefficients that were tested in the study.

Table 4.1. Examples of some similarity/diversity coefficients commonly used with 2D fingerprints.

Coefficient	Expression
Tanimoto	$\frac{c}{a+b-c}$
Squared Euclidean	$\frac{a+b-2c}{n}$
Cosine	$\frac{c}{\sqrt{ab}}$
Hamming	$a+b-2c$
Russel-Rao	$\frac{c}{m}$
Forbes	$\frac{cm}{ab}$
Two compounds, A and B, are encoded by binary fingerprints with a total of *m* bits. The fingerprint for compound A (or B) has *a* (or *b*) bits set to one, and there are *c* bits set to one in both fingerprints (i.e. the logical *and* of the two bit-strings). The Hamming coefficient is in fact a distance, so that large values represent dissimilar pairs of molecules, and not similar pairs of molecules as with the other coefficients.	

The results obtained by Holliday et al. [534] and Salim et al. [535] imply that some of the available coefficients quantify different types of structural resemblance. This realization has inspired the use of multiple similarity coefficients and of *data fusion* methodologies for combining the results of database searches. While initially developed for use in signal processing (by combining the input of multiple sensors), data fusion is today applied to a broad spectrum of domains. Its application in VS has been relevant, allowing the combination of results employing different metrics but also different reference (template) molecules.

1.3.4. 2D similarity searches using multiple templates

The vast majority of similarity methods reported to date makes use of only a single template ligand during one search run. However, by focusing on only a single molecule, it is likely that molecules holding equivalent chemical and/or biological properties and activity, yet with dissimilar structure, will be overlooked, leading to a high false negative rate [536]. With the increasing awareness of the chemistry and availability of crystal structures of multiple, structurally diverse, active compounds for a given purpose or target, there has been recent interest in the use of multiple template ligands, especially in the field of VS [115,537]. Indeed, fingerprint searching is often more effective when multiple template ligands are used due to an increase in the amount of information available for the calculations [116,538]. Accordingly, several different approaches have been applied to multiple template-based fingerprint searching, including consensus fingerprints [539], centroid fingerprints [114,116,538], bit-scaling techniques [113] and nearest-neighbour methods [113,116,537,538,540,541]. A number of studies comparing these different search strategies have shown that the nearest-neighbour methods and the centroid methods often give the best performance [114,537]. Consensus fingerprints, centroid fingerprints, and bit-scaling techniques emphasise bit positions that are shared by all template ligands and are, therefore, likely to account for their observed bioactivity [116]. Although they are powerful methods, they may be of less use when applied to the structurally diverse molecules that are typical of high-throughput screening experiments [115,537]. Indeed, they are more likely to rediscover the common features of known bioactive molecules rather than discover new chemotypes, which is a desirable feature of scaffold-hopping approaches. Conversely, the nearest-neighbour method individually calculates the similarity of each compound in the database to all of the template ligands, and then either the largest observed similarity score is selected to represent each database compound (MAX fusion rule) or the mean of the similarity scores is taken for the k nearest neighbours (SUM fusion rule) [116]. Both Hert et al. [115] and Whittle et al. [540] showed that the best recovery of active compounds was achieved when ranking the database molecules according to the MAX fusion rule.

2. Computational methods

In this section, I describe the computational methods and the parameters that have been used during the assembly of (*i*) a benchmarking set for the evaluation of VS methods against TTR (explored in Chapter 5) and (*ii*) a tailored screening set for VS against TTR amyloid (explored in Chapter 7).

2.1. Assembly of a benchmarking set for TTR

The ultimate goal of this PhD project is to evaluate the application of a large battery of virtual screening methodologies to the identification of novel, functional and safe TTR amyloid inhibitors. To evaluate the power of the different methods to discriminate between active TTR binders/stabilisers and inactive compounds, we compiled a benchmarking set composed of active and decoy compounds. All collected data on known TTR binders reported in the literature and deposited in the PDB were re-analysed [206,214–220,222,397]. Twenty-two diverse compounds were selected as "actives" making use of the available binding affinity data and clustering by maximum common substructure (MCS) [542] to ensure maximal diversity and therefore avoid the inclusion of compounds coming from the same series (see ligands marked with a † in **Figure 4.2**).

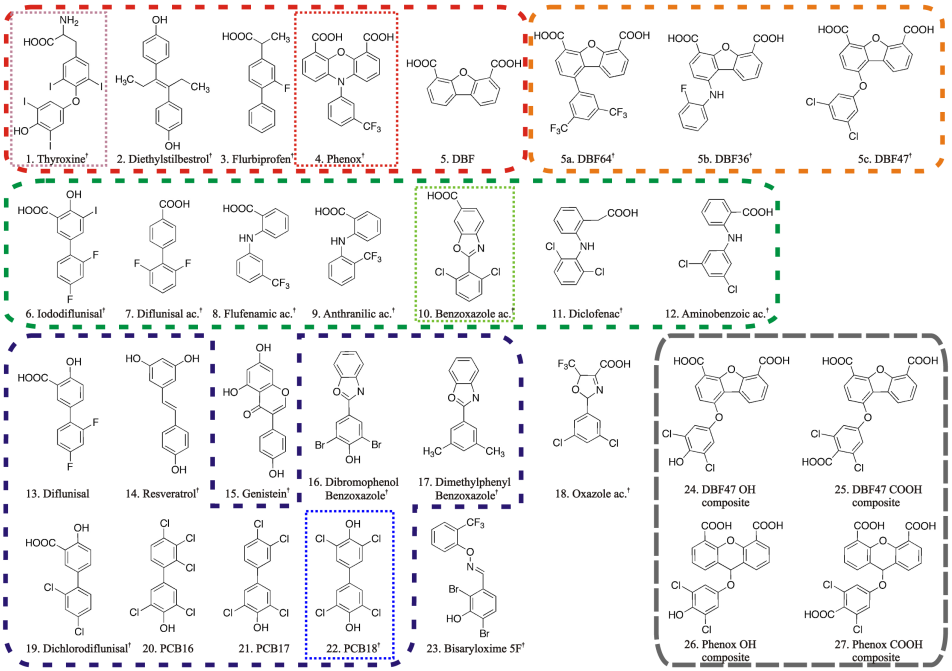


Figure 4.2. TTR binders reported in the literature and/or deposited in the PDB, along with the modelled composite ligands. Ligand mapping following an overlay of multiple TTR complexes revealed three clusters of ligands occupying similar geometrical positions within TTR binding pocket AC: red (cluster A), green (cluster B) and blue (cluster C) dashed boxes. The orange dashed box comprises molecules from three classes of dibenzofuran-based inhibitors: compound **5c** has the highest efficiency inhibiting fibril formation, compound **5b** shows the highest plasma binding stoichiometry, and compound **5a** represents a compromise between inhibition of fibril formation and binding stoichiometry. Compounds highlighted by the dotted boxes were used as template queries for 2D and 3D similarity searches in this work. The grey dashed box comprises four composite ligands that were modelled using compounds **4** and **5c** as reference. Ligands marked with a † were used as actives in the benchmarking set.

To select the decoys we followed a strategy similar to the one used to build the Directory of Useful Decoys (DUD) [543], in three main steps (see **Figure 4.3**). The 22 TTR actives were seeded amongst a library of 5.5 million Lipinski-compliant molecules deposited in the ZINC database [137]. Two-dimensional dissimilarity analyses were performed using UNITY fingerprints [457] and by computing Tanimoto similarity coefficients between each active and every compound in the library. Compounds holding a Tanimoto coefficient equal or greater than a threshold of 0.5 to any of the annotated actives were excluded and the initial library was thus reduced to approximately 4.5 million compounds topologically dissimilar to the TTR binders. Secondly, for each active ligand we extracted from the library approximately 100 ZINC compounds possessing similar physical properties. This was carried out using the program FILTER [513] and by tightly defining property ranges for the number of heavy atoms, number of hydrogen-bond acceptors, number of hydrogen-bond donors, sum of formal charges, octanol-water partition coefficient (XlogP) and number of rotatable bonds. Finally, we performed diversity analysis using clustering by maximum common substructure (MCS) and by computing Tanimoto coefficients within each group of 100 ZINC compounds. Only cluster representatives holding highest dissimilarity were selected amongst each group of 100 compounds (in equivalent proportions). This way we ensured that, as far as possible, each decoy is representative of a unique chemical series when considered by the VS algorithms.

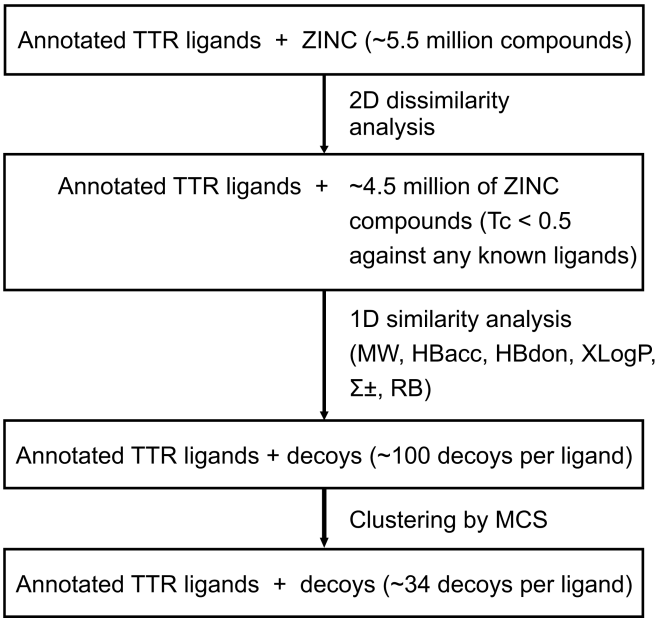


Figure 4.3. Schematic representation of the procedure followed to assemble a benchmarking set tailored for the evaluation of VS methods against TTR. Molecular weight (MW), number of hydrogen bond acceptors (HBacc), number of hydrogen bond donors (HBdon), octanol-water partition coefficient (XLogP), sum of formal charges (∑±), number of rotatable bonds (RB). The Tanimoto coefficients (Tc) were computed with

UNITY, using UNITY 2D fingerprints. This figure was inspired by a similar scheme presented in reference [543].

After removing all redundant entries from an initial set of 786 compounds, a final set of 738 decoys was obtained (approximately 34 decoys per active). The 22 annotated actives were then seeded amongst these 738 decoys to constitute a final benchmarking set comprised of 760 compounds and tailored for the evaluation of VS methods against TTR.

2.2. Analysis of molecular diversity and properties

The molecular diversity and properties of the active and decoy molecules selected for the benchmarking set have been analysed (and visualised) using multiple data mining techniques, such as clustering and principle component analysis (PCA). In this subsection, these techniques are briefly described within the context of the software packages used to carry out the analyses.

2.2.1. Clustering via Maximum Common Substructure – LibraryMCS

LibraryMCS is a program provided by ChemAxon that clusters a set of chemical structures on a structural basis [544]. It clusters together structures sharing a common substructure. The clustering engine recognizes the so-called Maximum Common Substructure (MCS), and it is always the largest one amongst all substructures found in the structure set. No predefined fragments are used while searching for the MCS.

The clustering technique employed by LibraryMCS is hierarchical, meaning that the clusters of input structures are grouped into second level clusters, then these second level clusters are grouped again and so forth, until a termination condition is reached (e.g. there is only one cluster left). These higher-level clusters are also built on a structural basis by identification of the MCS of the constituting clusters.

2.2.2. Molecular descriptors – The Chemistry Development Kit (CDK)

In post-filtering stages, the Chemical Development Kit (CDK) [504] was used as an alternative resource to analyse a number of basic descriptors for TTR stabilisers selected as actives and the decoy molecules comprised in the benchmarking set. Amongst these are: the molecular weight, the number of heavy atoms, the number of hydrogen bond acceptors and donors, the number of rotatable bonds, atomic polarizabilities, the calculated octanol-water partition coefficient, and the topological polar surface area. Besides repre-

senting an open source solution, the CDK has been successfully integrated in widely-used workflow systems, such as the Konstanz Information Miner (KNIME) [545,546] and Taverna [547], thus allowing a straightforward (and visual) pipelining of data and the automation of molecular descriptor analyses.

2.2.3. Analysis workflow – The Konstanz Information Miner (KNIME)

The Konstanz Information Miner (KNIME) is an open-source workflow software developed at the University of Konstanz [545,546]. KNIME provides an excellent data-mining platform for chemoinformatics and drug discovery, and was used multiple times throughout this PhD project to support the interoperability of its different stages. The default implementation of KNIME can be complemented by several extensions dedicated to chemistry and the life sciences. Important examples are the CDK-based and Open Babel-based nodes, and other nodes deposited at the KNIME's "Community Updates" site such as the RDKit and Indigo nodes. KNIME can thus be used to transform molecular structures, compute QSAR descriptors and fingerprints, and implement machine-learning algorithms (Support Vector Machines, Regression and Bayesian Modelling, PCA, etc.).

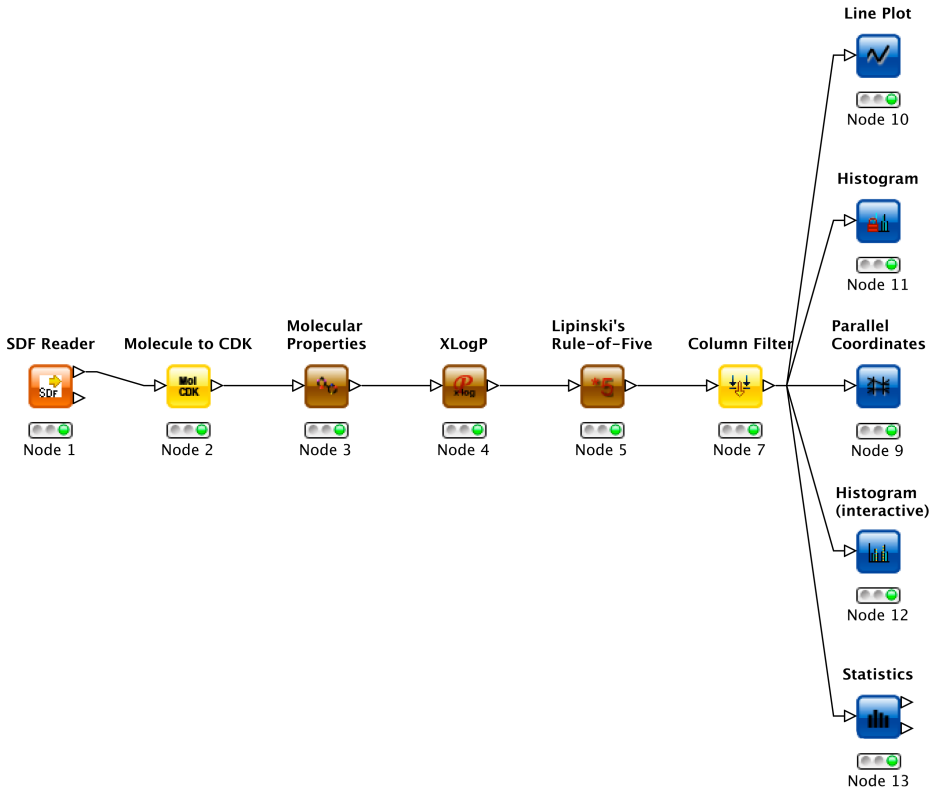


Figure 4.4. A simple KNIME workflow for the analysis of several molecular properties throughout this chapter. This workflow makes use of CDK extensions designed for KNIME

and was used to analyse the properties of active and decoy molecules comprised in the TTR benchmarking set.

2.2.4. Principal component analysis – ChemGPS

Principal component analysis (PCA) is a multivariate analysis technique widely used to study the relationships amongst a set of correlated variables. These variables are compressed into a smaller number of new uncorrelated variables called the principal components (PCs), while retaining as much information as possible. Thus, PCA filters out the noise, provides a reduction of data multidimensionality (for instance, the properties of the input molecules), offers a global overview of the data, and detects trends, groupings and outliers.

In this work, the ChemGPS system was used to position chemical structures (namely, the active and decoy molecules) in chemical space via PCA score predictions [548,549]. The ChemGPS space map coordinates are *t*-scores from PCA based on a subset of 35 descriptors that describe the size, shape, lipophilicity, polarity, polarizability, flexibility, rigidity, and hydrogen bond capacity of molecules [548,549].

2.3. Library filtering for the assembly of a screening set

Filtering of chemical libraries was performed with two distinct objectives: (*i*) to select compounds holding physicochemical properties similar to those of known TTR stabilisers (yet topologically dissimilar) to be integrated in a benchmarking set (as explained above), and (*ii*) to narrow down an initial library of approximately 11 million compounds downloaded from the ZINC database to an amenable size for virtual screening, thus yielding the screening set tailored for TTR.

2.3.1. FILTER – Program overview

FILTER is an extremely fast program provided by OpenEye Scientific Software that eliminates unsuitable or undesirable compounds from large libraries for virtual screening. FILTER's high speed of calculation, which is critical at the database preparation stage, results mostly from the strict use of two-dimensional and graph-based algorithms. The program's criteria for passing or failing a given molecule fall into three main categories: (*i*) physical properties, (*ii*) atomic and functional-group content, and (*iii*) molecular graph topology. Examples of the simple physical properties calculated by FILTER include mo-

lecular weight (MW), topological polar-surface area (TPSA) [550], XLogP [551], and aqueous solubility. The filters also include absolute and relative content of heteroatoms, as well as limits on the number of a very wide variety of functional groups. The graph topology filters address issues concerning the number and size of ring systems, the flexibility of the molecule and the size and shape of non-ring chains. Moreover, FILTER includes a *default* filter that encapsulates many of the *standard* filtering principles, such as the removal of unstable, reactive and toxic moieties.

Filter definitions are passed on to the FILTER program through a parameter file that stores four types of statements: physical property limits, rules, new rules and selections. For example, physical property limits include minima and maxima of simple properties: MW, number of heavy atoms, carbon count, number of chiral atoms, number of hydrogen bond acceptors and donors, sum of formal charges, halide fraction, TPSA and XLogP. Solubility is also a filter, but different categories, rather than a quantitative cut-off, are used. The six possible categories are: insoluble, poorly, moderately, soluble, very and highly. Secondary filters impose pharmacokinetic constraints and are based on reported combinations of the abovementioned simple properties.

2.3.2. Filter definitions

The choice of filter definitions was a paramount aspect in this project, considering our goal of identifying small molecules that, besides being capable of inhibiting amyloid formation, could represent promising leads towards the development of functional and safe new drugs. The filtering of the 10,962,930 molecules downloaded from ZINC (version 8) consisted of a combination of pre-defined FILTER rules for drug-likeness and bioavailability with knowledge of the physicochemical properties of the known TTR stabilisers. The pre-defined rules included Lipinski's Rule-of-Five [510], allowing up to one violation and defining hydrogen bond donors and acceptors as outlined in the work of Mills and Dean [552]. Other pre-defined rules included:

- Veber's/GSK's measure of bioavailability [553]: only compounds with (*i*) 10 or fewer rotatable bonds or (*ii*) polar surface area equal or less than 140 $Å^2$ were retained;
- Yvonne Martin's/Abbott's Bioavailability Score: this score assigns the probability (ABS) that a compound will have permeability and bioavailability (*F*) higher than 10% in the rat [554]; only compounds with ABS equal to or greater than 0.5 were retained;

- Egan's/Pharmacopeia's ("Egan egg") measure of bioavailability [555]: only compounds with (*i*) a calculated LogP equal or below 5.88 or (*ii*) a PSA equal or below 131.6 Å^2 were retained.

Aggregators and predicted aggregators, i.e. small molecules known or predicted to aggregate and sequester protein in solution [556,557] (thus interfering with assay results), were also filtered out. Molecules predicted to be insoluble or poorly soluble in water were discarded as well.

Furthermore, based on our analysis of various physicochemical properties of the known TTR actives and their distributions, we customized filter definitions to ensure that no redundant molecules were included and that no potential binders were missed at an early stage of the virtual screening. Thus, the maximum sum of formal charges decreased to 0 and the maximum XLogP increased to 6.5. Moreover, given TTR's propensity to bind halogen-bearing (electronegative) molecules, such as thyroxine, the maximum halide fraction (percent of molecular weight from halides) was increased to 0.6. Finally, iodine was added to the list of allowed elements.

The full list of parameters passed on to FILTER in order to carry out the filtering of the initial library downloaded from ZINC is given in the Appendix (Section C).

2.4. Generation of three-dimensional conformers

While many of the virtual screening methods explored in this thesis can handle ligand flexibility on-the-fly, usually by means of more or less complex functions (such as the case of genetic algorithms in molecular docking and flexible pharmacophore searches), some other methods are not able to do so. To provide meaningful results, these methods must be coupled with other programs that perform searches in ligands' conformational space by generating ensembles of conformations and scoring each conformer with an energy function. Many different algorithms have been developed to efficiently generate low-energy conformer ensembles.

Drug-like molecules usually have multiple rotatable bonds and can adopt many different conformations. Thus, molecular flexibility is a requirement of rigid ligand alignment methods, such as GH8 (see on page 73). Remark that even though GH8 was originally developed as a solution for the structural alignment of proteins, it can be used for the superimposition of small molecules as well. The flexibility of ligands can be dealt with by rigidly aligning pre-generated conformer libraries. The cost of generating the conformers only needs to be incurred once, and the library can be subsequently scored quickly against any target of interest.

The determination of the bioactive conformation of a ligand is a critical step in computational drug discovery. A ligand conformer ensemble is a set of multiple 3D conformations that is generated from a single starting conformation. Ideally the conformer ensemble will contain a conformer that is close (< 1 Å RMSD) to the correct solution. Since the bioactive conformation is not usually found at the global energy minimum in the energetic landscape, it is necessary to generate an ensemble of low-energy conformers that will hopefully contain a conformer that is sufficiently similar to the bioactive conformation. Studies evaluating the performance of several popular conformation search tools have shown that for lead-like ligands with less than eight rotatable bonds, a conformer within 1 Å of the bioactive conformation is generated in more than 80% of cases [98,558]. However, as the number of rotatable bonds increases, the chance of retrieving the bioactive conformation decreases.

For reasons of speed and availability (in our Bioinformatics lab), Openeye's OMEGA was chosen for the generation of conformers throughout this project.

2.4.1. OMEGA – Algorithm overview

OMEGA employs a rule-based, depth-first deterministic method for conformer sampling. Different ring conformations are enumerated using a fragment library, before the molecule is disassembled into fragments of up to five contiguous rotatable bonds. Different conformations of each fragment are generated based on a library of predefined torsions, and rotatable bonds are systematically adjusted using SMARTS (SMiles ARbitrary Target Specification) matching. The fragments are subsequently reassembled based on the order of their ascending energies, thereby generating a pool of alternative molecular conformations. Molecular energies are calculated using the Merck Molecular Force Field [30], and any conformations above an adjustable energy threshold (in relation to the lowest energy state), or less than an adjustable RMSD threshold (in relation to all other conformations), are rejected. OMEGA has been shown to be very reliable and effective at reproducing bioactive conformations [442].

2.4.2. Selection of parameters for 3D sampling

Throughout this project, multiple ligand conformations were generated using OMEGA. An energy window of 25 kcal.mol^{-1} was chosen since it was shown to be essential for best conformer generation performance in a study by Kirchmair et al. [536]. A lower energy window could cause valuable conformations to be discarded, whereas a higher energy threshold could produce high-energy conformers unlikely to represent bioactive confor-

mations. A diversity threshold (RMSD) of 1 Å was also used to discard similar conformations, and a maximum of 100 conformers were generated per molecule. These parameters were used in the work by Sperandio et al. [559] because they represented an optimal balance between speed and performance and agreed well with the study carried out by Kirchmair et al.

3. Results and discussion

In this section, the results of the analyses leading to the assembly of a benchmarking set and a screening set tailored for TTR are discussed. The two data sets are described as a result of the modelling operations conducted in this chapter and under the light of their relevance for the evaluation and application of VS methods.

3.1. Analysis of active TTR stabilisers using unidimensional descriptors

Several one-dimensional descriptors and physicochemical features FILTER for the known TTR stabilisers were computed with the program, in order to trace TTR binding propensities in terms of the most basic molecular properties. **Figure 4.5** shows a statistical analysis for 26 compounds using four bulk descriptors: the molecular weight (MW), number of heavy atoms, octanol-water partition coefficient (determined as XLogP), and sum of formal charges. As can be seen, TTR tends to bind molecules that are in the "lead-like" range in terms of molecular weight (median MW of approximately 325 Da). With a MW larger than 700 Da, the endogenous natural ligand of TTR, the thyroid hormone thyroxine, represents a clear outlier amongst the distribution of MWs. The tendency to bind relatively small ligands is further confirmed by looking to the median number of heavy atoms in the known binders, which is no higher than 20, a figure that is also closer to the lead-like range than to the drug-like range. Furthermore, molecules binding strongly to TTR are most likely hydrophobic; the median XlogP for the known stabilisers is around 4, a value that approximates the upper limit of drug-like properties (according to Lipiniski's Rule-of-Five). Another interesting feature characterizing the known TTR actives is their tendency to be negatively charged. Indeed, the median value for the sum of formal charges in TTR stabilisers is -1 and the maximum value 0.

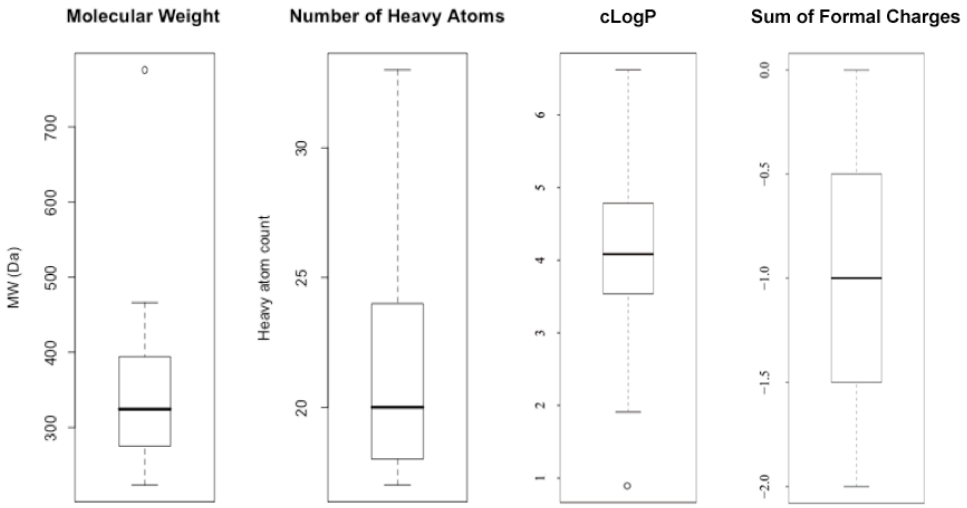


Figure 4.5. Box plots to compare selected physicochemical properties of all 26 TTR stabilisers known to date. The box region provides a graphical view of the median and quartiles of the distribution; the dotted lines extend to maximum and minimum values of the data set).

This particular feature of TTR stabilisers seems to be in line with the electrostatic properties of TTR. Even though the surface of the protein is mostly negatively charged – which is linked to its characteristic electrophoresis migration pattern; and hence its historical name of "pre-albumin" – TTR binding sites seem to counteract this characteristic, being mostly electron-depleted (**Figure 4.6**).

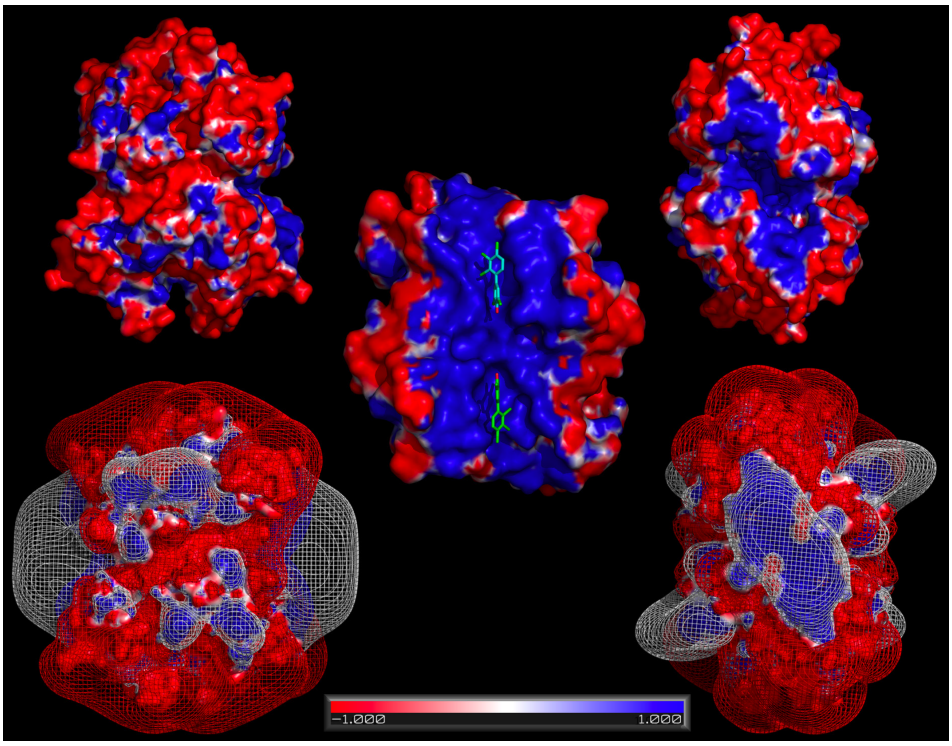


Figure 4.6. Electrostatic properties of TTR determined by Poisson-Boltzmann calculations. Dielectric constants of 4 and 80 were employed for the protein and the solvent, respectively. The electrostatic potential is represented as a colour scale (ranging from -1 to +1 kT/e) at the protein surface – top and central panels – and as isocontours at -1 kT/e (red contours), 0 kT/e (white contours) and +1 kT/e (blue contours) – bottom panes. The leftmost panes show a side view of TTR, rotated 90º from the front view shown in the rightmost panes. The central pane shows a cross section of the tetramer taken at the binding channel symmetry axis (here, the ligands bound to both pockets are displayed for reference). The Poisson-Botzmann calculations were carried out with MEAD [560] and the figure was produced with PyMol [561].

3.2. Selection of templates for ligand-based screening

The choice of the active molecules to use as reference queries (or templates) for 2D and 3D similarity searches is a difficult one. Firstly, it is very unlikely that one single active molecule can be representative of a whole series of actives or even hold all the ideal pharmacophoric features required for optimal interaction with the target binding site. Moreover, 3D similarity search methods would ideally take the bioactive conformation of the reference ligand as a query for the searches, but sometimes this is unavailable.

In the case of TTR amyloid, a comprehensive selection of TTR stabiliser ligands was carried out using criteria like potency, binding mode and state of development. On this basis, four specific molecules were chosen as reference molecules (or template ligands) for the

benchmarking (Chapter 5) and application (Chapter 7) of similarity-based virtual ligand screening.

3.2.1. Known actives as templates

For the comparison of the different similarity-based VS methods described in this chapter, we chose TTR's cognate ligand, thyroxine (compound **1**, see **Figure 4.2**) the most prominent endogenous TTR binder [562]; phenox (compound **4**) [215] a promising TTR stabiliser and representative of cluster A; a benzoxazole acid (compound **10**) [218] another potent amyloid inhibitor representative of cluster B; a polychlorinated biphenyl (compound **22**) [219] the most potent TTR binder known so far and representative of cluster C (**Figure 4.2**). In all cases, the structures of the template ligands were extracted from the corresponding X-ray coordinates deposited in the PDB: 2rox for thyroxine (T_4), 1dvy for phenox, 2f8i for the benzoxazole acid and 2g5u for the polychlorinated biphenyl (PCB18). DBF47 (compound **5c**) is a potent dibenzofuran derivative sharing structural similarity with phenox [222]. Because no X-ray coordinates are publicly available for the full structure of DBF47 (compound **5c**), we took the coordinates of the dibenzofuran-4,6-dicarboxylic acid moiety (compound **5**) in the complex with TTR (PDB code 1dvu) and modelled in the missing 3,5-dichloro aryl ether substituent.

3.2.2. Modelling of concatamers

To address the problem of *representativeness* affecting the use of single templates in VS, we took the structures of both compound **4** and compound **5c** (see **Figure 4.2**) and modelled in missing features that are believed to be critical for both shape complementarity with the receptor sites of TTR and interaction with the hydroxyl side chains of two serine residues located at the inner portion of the pockets [215,216,220]. This resulted in four new composite ligands that are concatamers of high affinity ligands: two ligands (compounds **24** and **25**) were based on the structure of DBF47 (compound **5c**) and the other two (compounds **26** and **27**) were based on the structure of compound **4**. X-ray coordinates were used as template (PDB code 1dvy) and the hydroxyl and carboxyl groups were added at the *para* position of the distal aryl ring. Since the trifluoromethyl group at the *meta* position of this ring is rather bulky to favourably allow the presence of a substituent at the *para* position, we replaced it for a chlorine atom and added an extra chlorine at position 5 (creating symmetry). This modification was inspired by the successful substitution pattern of the corresponding ring in DBF47 [222]. The composite structures were then energy minimized and conformational sampling was performed to identify low energy conformations retaining the closest similarity to the X-ray reference models. All

modelling procedures were carried out with ChemAxon's MarvinSketch [563]. Conformational sampling was performed with OMEGA. The predicted lowest-energy conformations were evaluated as potential template queries for VS.

3.3. A benchmarking set for the evaluation of VS methods

The high structural diversity found amongst the strongest TTR stabilisers has been outlined throughout this thesis and even quantified in Chapter 3, where the presence of two aromatic rings was revealed as the most unifying feature characterising TTR's pharmacophore. Understanding this diversity is a key to the correct evaluation of VS methods against TTR amyloid. Equally, the criteria for selection of inactive compounds (decoys) to use in an evaluation of VS protocols are of fundamental importance and should be able to mirror the structural diversity found amongst the active molecules. As will be further explained in Chapter 5, in this project we seek to evaluate the VS algorithms' performance at discriminating active molecules over inactive ones based on the compounds' ability to establishing complementary interactions within TTR receptor sites, rather than by inconsequential physicochemical properties, such as their molecular weight. Therefore, despite possessing dissimilar chemical topologies to the known TTR ligands, the selected decoys should share similar physicochemical properties.

To evaluate the discriminative power of different virtual screening methods, we compiled a benchmarking set comprised of TTR-active and decoy compounds. As with the construction of pharmacophore hypotheses described in Chapter 3, the selection of actives was accomplished upon knowledge of TTR binders reported in the literature and deposited in the PDB [206,214–220,222,397]. On this basis, 22 binders were then selected as "actives" making use of the available binding affinity data and clustering by maximum common substructure [542] to ensure maximal diversity and therefore avoid the inclusion of compounds coming from the same series. Subsequently, a *mimetic* set of 738 decoys was constructed by matching the physical properties of the actives within a database of topologically dissimilar drug-like compounds. Details on the construction of this set of decoys is presented in section 2.1 of this chapter.

3.3.1. Benchmarking set statistics

Table 4.2 reports global statistics for the benchmarking set. It is particularly useful to contrast the number of compounds with the number of top-level clusters determined via clustering by maximum common substructure (MCS), to get a picture of the diversity of the data set. Indeed, diversity is critical to the quality of the benchmarking set. Even

though there is no absolute definition of chemical similarity, several VS methods heavily rely on this concept; particularly, the 2D and 3D similarity methods described in this chapter, which will also be assessed in Chapter 5. Suggestions as to how to circumvent this issue include the narrowing of the sets of actives to smaller sets of representative structures, which we have attempted during the selection of TTR actives using clustering by MCS, or giving more weight to the first compound discovered in a series (this approach will be explored in Chapter 6). Although decoys are less likely to come from congeneric series, such approaches may also be important under the context of the protocols followed to select them. This will be further discussed in the following subsections.

It is also worth noting that the proportion found between the number of active molecules and the number of decoy molecules in the data set (34 decoys per ligand) is roughly identical to the average proportion found in the Directory of Useful Decoys (DUD): 36 decoys per ligand. As will be discussed in Chapter 5, a high proportion between actives and decoys (such as the one found in the DUD) is critical when the purpose of the benchmarking sets is to evaluate one VS method against other methods [564].

Table 4.2. Basic statistics for the benchmarking set assembled for TTR. The number of compounds, the number of conformers generated by OMEGA and the number of top-level clusters generated by maximum common substructure (MCS) are reported for the known actives, the selected decoys and the entire data set.

	# Compounds	# Conformers (by OMEGA)	# Top-level clusters (MCS)
Actives	22	119	14
Decoys	738	8857	387
TOTAL	760 (~ 1 A : 34 D)	8976	401

As mentioned above, a high similarity between actives and decoys in terms of physicochemical properties is an essential characteristic to a good benchmarking set. **Figure 4.7** groups together descriptive statistics for several one-dimensional descriptors computed for the actives and the decoys in our benchmarking set. The median values reveal close similarity for properties like MW, number of heavy atoms, number of hydrogen bond donors, number of rotatable bonds and atomic polarizabilities. Even though other properties have been included as criteria for the selection of decoys, less agreement was found. This mostly applies to parameters reflecting the lipophilicity/hydrophobicity of molecules: the XLogP and the TPSA. We believe these differences do not reflect limitations in the size (and diversity) of the database from which the decoys were withdrawn (comprised of approximately 4.5 million ZINC compounds), so much as the intrinsic drug-like nature of the compounds within that database. In this context, however, it is also important to bear in mind that the properties of the decoys should also mirror the global properties of the databases typically used for VS, thus reproducing the setting where the actual screenings take place. Still, looking at the upper quartiles of property

distributions and their maxima it becomes clear that "greasier" decoys are certainly represented in the benchmarking set (though in a smaller extent).

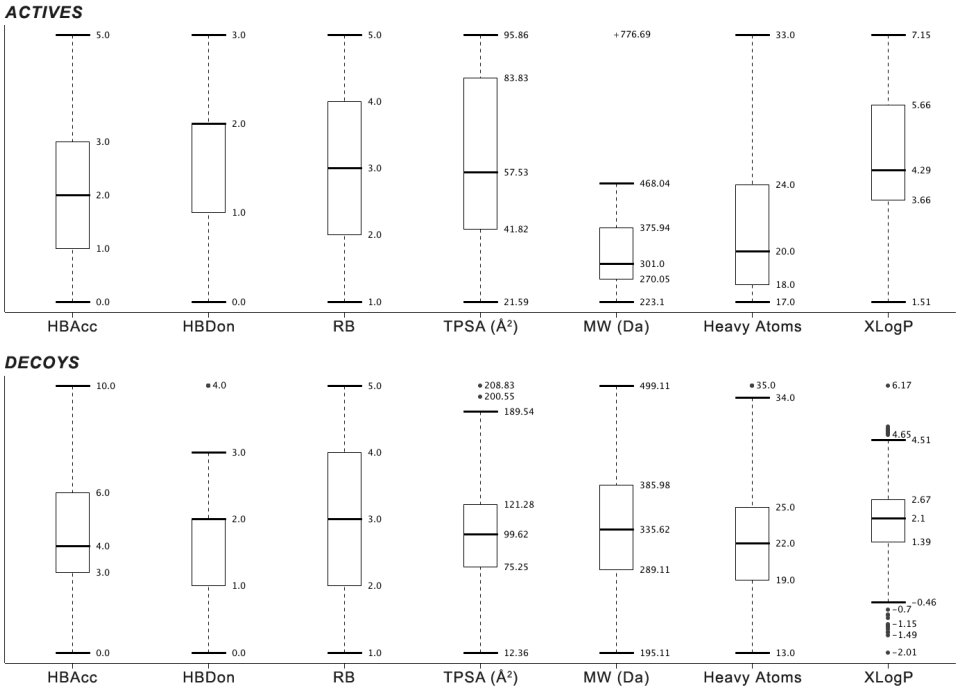


Figure 4.7. Box plots to compare a set of relevant physicochemical properties amongst TTR actives and decoys. All descriptors were computed using the program CDK. The box region provides a graphical view of the median and quartiles of the distribution; dotted lines extend to maximum and minimum values of the data set.

3.3.2. Analysis of TTR actives

Figure 4.8 provides an illustration of the clustering analysis based on maximum common substructure (MCS) performed on the 22 TTR amyloid inhibitors selected as actives. A total of 14 top-level clusters are comprised in the dendrogram. The first cluster marked in the intermediate level (cluster 1), for example, groups together three representative structures of the three families of dibenzofurans (compounds **5a**, **5b** and **5c** in **Figure 4.2**); grouped in cluster 3 are two benzoxazoles (compounds **16** and **17** in **Figure 4.2**); and so forth. The common substructures highlighted in the marked nodes of the branched trees correspond to fairly small portions of the entire structure of the compounds that populate the clusters. Interestingly, though, 10 out of the 22 ligands form unique clusters. Altogether, these observations provide an indication of the high structural diversity found amongst the selected TTR actives.

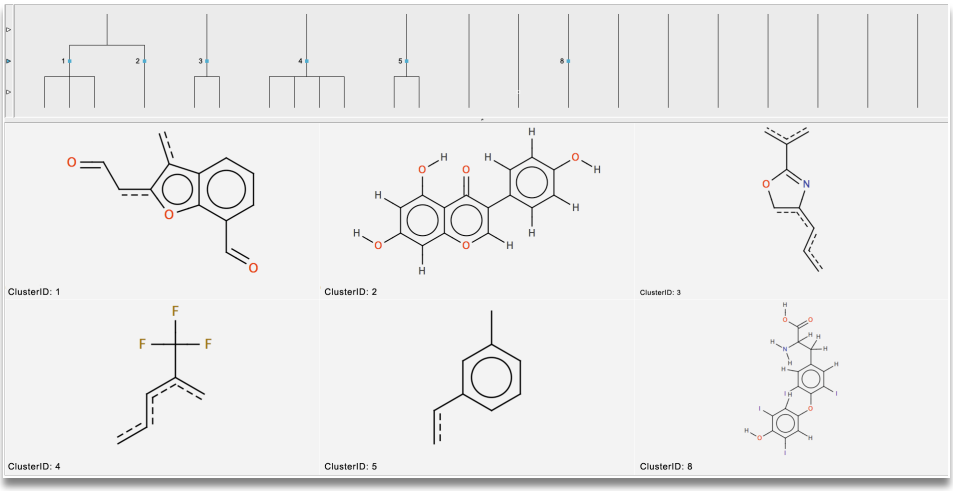


Figure 4.8. Maximum common substructure-based clustering of the 22 representatives of TTR amyloid inhibitors selected to integrate the benchmarking set as actives. The tree nodes corresponding to the intermediate level of clusters were marked to show the common substructures that are found amongst the active molecules of the data set. For comparison, TTR's natural ligand, thyroxine, was also marked, even though it integrates an individual cluster (cluster 8). The figure was produced using LibraryMCS, provided by ChemAxon [544].

Furthermore, in this chapter we devised a KNIME workflow, making use of CDK extensions, to carry out a detailed analysis of physicochemical properties of the active molecules selected to integrate the benchmarking set. In **Figure 4.9**, the bar plots show the number of compounds fitting each of several property intervals. For example, it is possible to realize that most TTR actives fall in the first two intervals of MW and number of heavy atoms, disclosing a relatively low size compared with the central value of the drug-like range (325 Dalton). Thyroxine (T_4) clearly represents an outlier due to its four (high-mass) iodine atoms. It is also evident that several of the known TTR amyloid inhibitors tend to be lipophilic in nature, falling in XLogP ranges larger than 4. Molecular polarizability walks hand in hand with molecular refractivity, and both properties have been linked with poor solubility [565]. It is interesting to note that some (but not the majority) of the known TTR actives explore higher ranges of atomic polarizabilities. Equally interesting is that this observation is in line with our groups' *wet lab* experience with some of the known actives: It has become clear that calculated octanol-water partition coefficients (such as cLogPs and XLogPs) alone are not foolproof predictors of solubility. Instead, a combination of multiple descriptors, such as TPSA, molecular refractivity and polarizability, the energy gap between the highest occupied and lowest unoccupied molecular orbitals, etc. seem to influence to different extent water solubility [565].

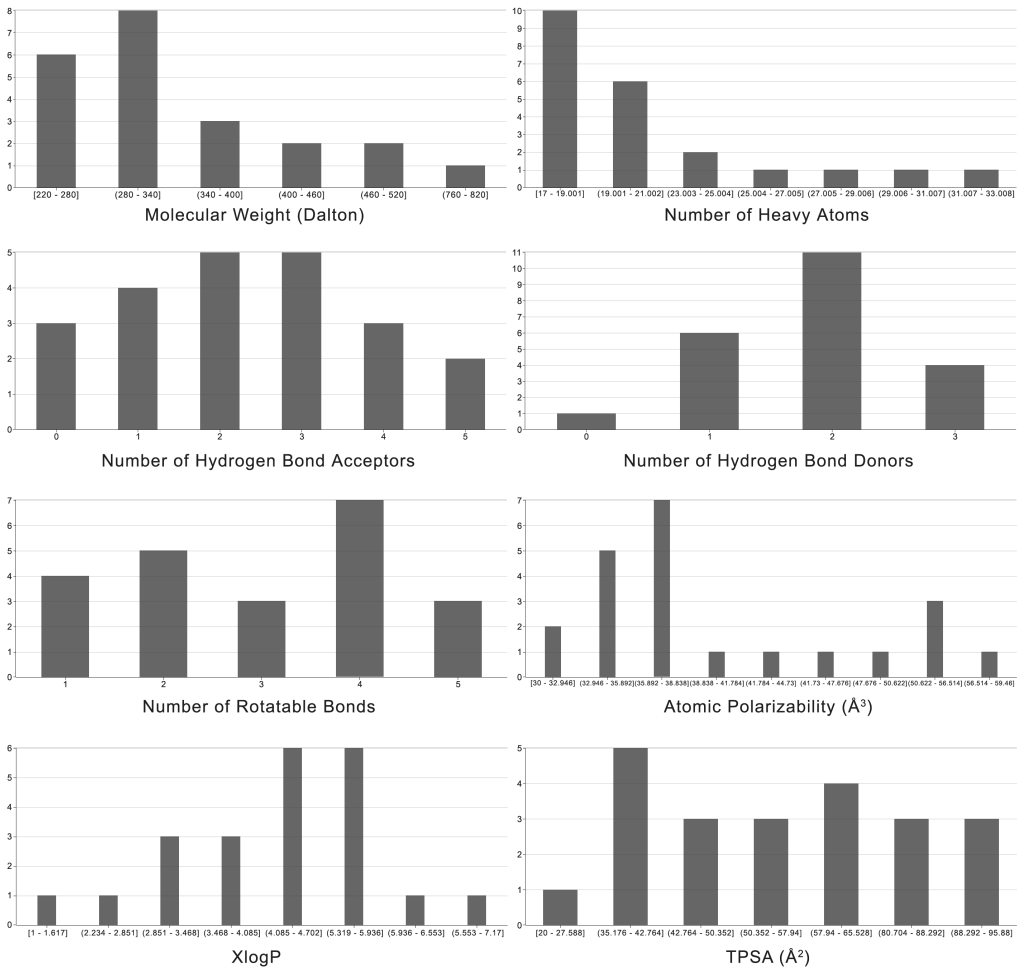


Figure 4.9. Detailed characterization of the physicochemical properties of TTR actives in the benchmarking set. The analysis and the plots were produced using the KNIME workflow presented in the Methods section.

We went further to study the positioning of the known TTR amyloid inhibitors in chemical space and contrast their physicochemical properties with those of other therapeutic drugs. This was accomplished via principal component analysis performed over 35 descriptors using ChemGPS [548,549]. The perception of dimensionality takes in the weight of individual variables in the data set, showing what molecular properties are explained by the respective orthogonal components (the so-called principal components, or PCs; see **Figure 4.10**). The four most significant PCs account for 77% of the variance: PC1 expresses size, shape and polarizability; PC2 depicts aromatic- and conjugation-related properties; PC3 represents lipophilicity, polarity and H-bonding capacity; PC4 discloses flexibility and rigidity.

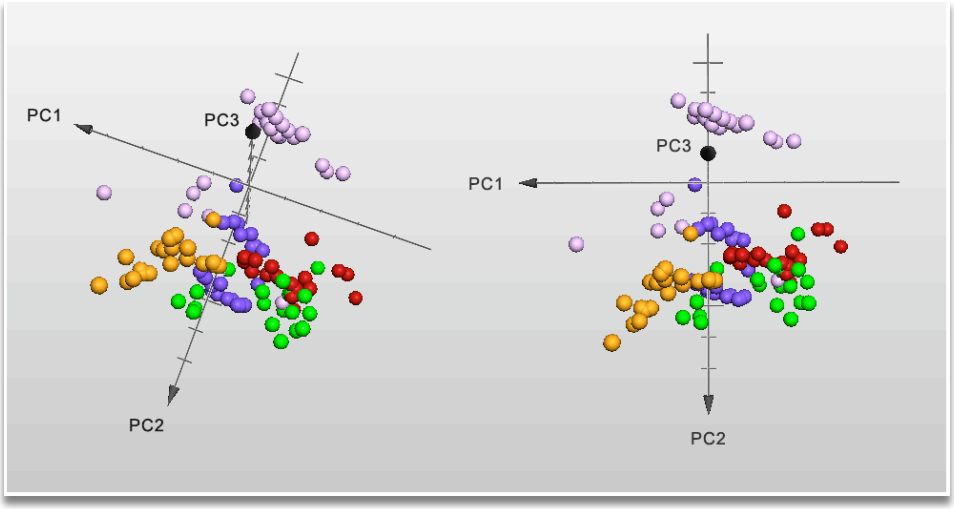


Figure 4.10. Three-dimensional projection of the *t*-scores obtained via PCA predictions based on 35 molecular descriptors for active compounds of 5 protein targets. The two panes are equivalent, differing only by a slight rotation in order to facilitate visualization in 3D. The main influences in PC1 are molecular size and shape, in PC2 is aromaticity and in PC3 is lipophilicity. Altogether, these three components explain 71% of the variance. Molecules corresponding to TTR actives are coloured in green. For comparison, active inhibitors of the enzyme cyclooxygenase-1 are shown in red, estrogen receptor antagonists are shown in orange, HMG-CoA reductase inhibitors are shown in light magenta and Heat Shock Protein 90 inhibitors in violet. PCA was carried out using ChemGPS [548,549] and the results were plotted with Apple™ system software Grapher 2.0.

Given their common "lineage", it is interesting noting that the properties of the known TTR actives map closely onto those of cyclooxygenase-1 inhibitors. Moreover, the high aromaticity of TTR actives is only overcome by that of estrogen receptor antagonists and matched by some of the Heat Shock Protein 90 (HSP90) inhibitors. Alongside with HMG-CoA reductase inhibitors, TTR actives show the highest dispersion in the coordinate space, showing that their diversity extends beyond their structural features to characterize their physicochemical features alike.

3.3.3. Analysis of TTR decoys

We applied the same set of analyses to the selected decoys, thus providing a more complete view of the characteristics of the benchmarking set and the degree of conformity between active and decoy molecules, both in terms of structural and physicochemical similarity/diversity.

Figure 4.11 illustrates the clustering analysis based on maximum common substructure (MCS) performed for the 738 ZINC compounds selected as decoys for the benchmarking set. A total of 387 top-level clusters are comprised in the dendrogram, which represents a

ratio of 1 cluster per 1.9 compounds. Indeed, more than half of the decoys are clustered individually, suggesting a reasonable degree of structural diversity in the data set. It can be noted, nonetheless, that the nitrobenzene motif is conserved across 233 compounds. This suggests that the selection the decoys may have suffered some bias caused by the use of identical filtering definitions based on actives that share very similar physicochemical properties.

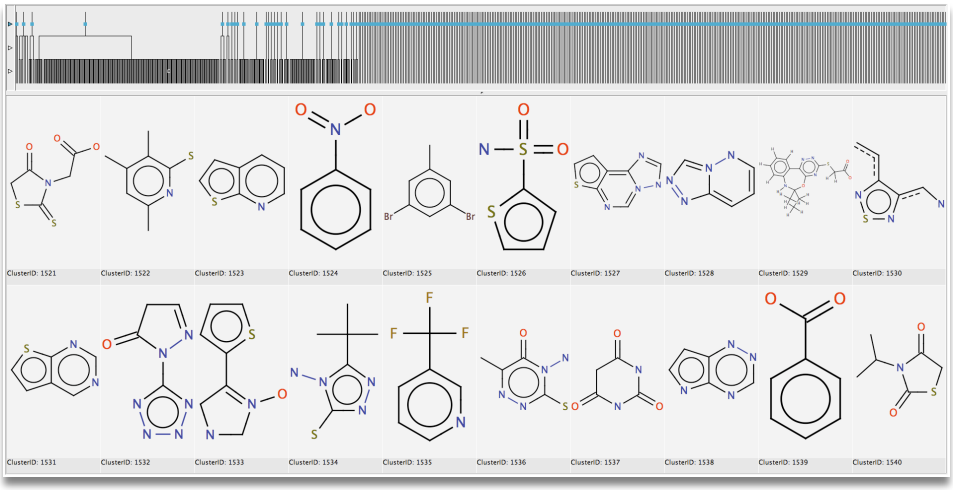


Figure 4.11. Maximum common substructure-based clustering of the 738 ZINC compounds selected to integrate the benchmarking set as decoys. The tree nodes corresponding to the top level of clusters were marked to show the common substructures that are found amongst the decoys of the data set. The figure was produced using LibraryMCS, provided by ChemAxon [544].

The bar plots in **Figure 4.12** represent the number of decoys falling inside each of several property intervals for the most relevant one-dimensional descriptors. It is interesting to see that, in general, the property distributions disclose a compromise between the patterns typically observed in general-purpose compound libraries [137] and those witnessed for TTR actives (see **Figure 4.9**). Comparing both groups (actives versus decoys), the more noticeable overlaps refer to the distributions of hydrogen bond donors and acceptors, and rotatable bonds. Looking into properties like the XLogP, TPSA and atomic polarizabilities, the plots also confirm that the proportion of decoy molecules that are likely to be soluble than insoluble is higher amongst the decoys than amongst the known TTR actives.

Figure 4.13 shows the projection of the PCA scores obtained for both TTR actives and decoys of the benchmarking set. Clearly, the diversity found amongst the physicochemical properties of the known TTR actives has been mirrored into the selected decoys, where even some undue dispersion may be argued. However, it is remarkable to see that all actives are closely surrounded in space by decoys holding similar physicochemical properties.

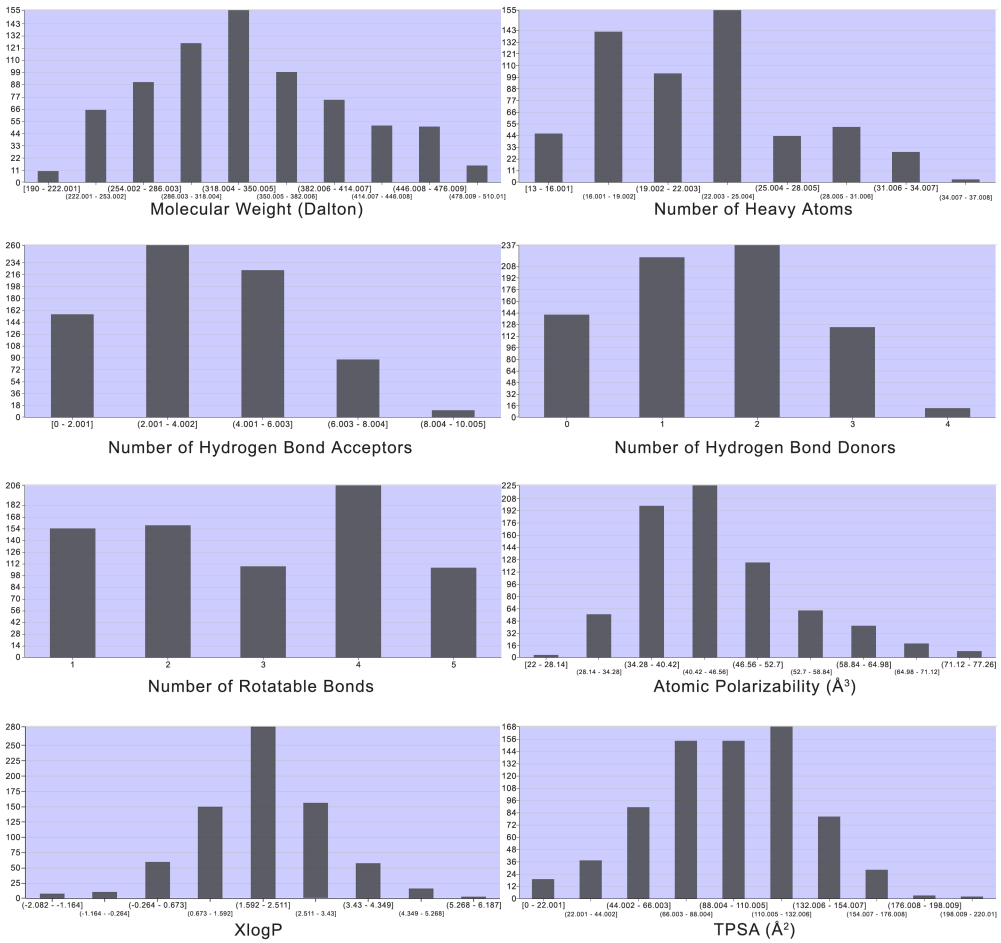


Figure 4.12. Detailed characterization of the physicochemical properties of the decoys selected to integrate the benchmarking set. The analysis and the plots were produced using the KNIME workflow presented in the Methods sections.

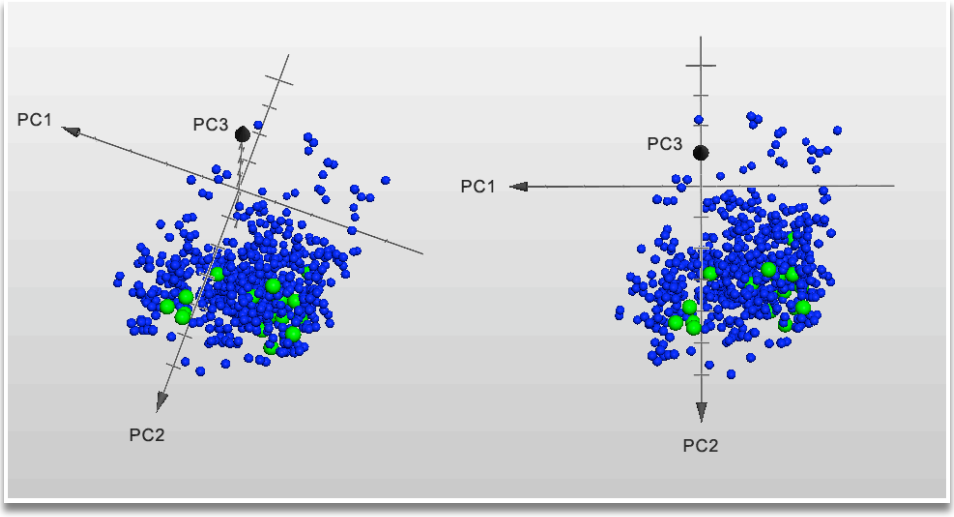


Figure 4.13. Three-dimensional projection of the *t*-scores obtained via PCA predictions based on 35 molecular descriptors for the entire TTR benchmarking set. The two panes are equivalent, differing only by a slight rotation in order to facilitate visualiza-

tion in 3D. Molecules corresponding to TTR actives are coloured in green, whereas decoys are shown in smaller spheres coloured in blue. PCA was carried out using ChemGPS [548,549] and the results were plotted with Apple™ system software Grapher 2.0.

We have contrasted these results with those obtained for some of the benchmarking sets deposited in the DUD [543]. As can be seen in **Figure 4.14**, the scattering of physicochemical properties of the decoys in chemical space is considerably lower for each of the four DUD targets analysed. In some cases this low dispersion in the decoy set is accompanied by a lower coverage of the property space of the corresponding actives. For instance, this seems to be the case of the benchmarking set deposited in the DUD for the target HMGA-CoA reductase (**Figure 4.14**).

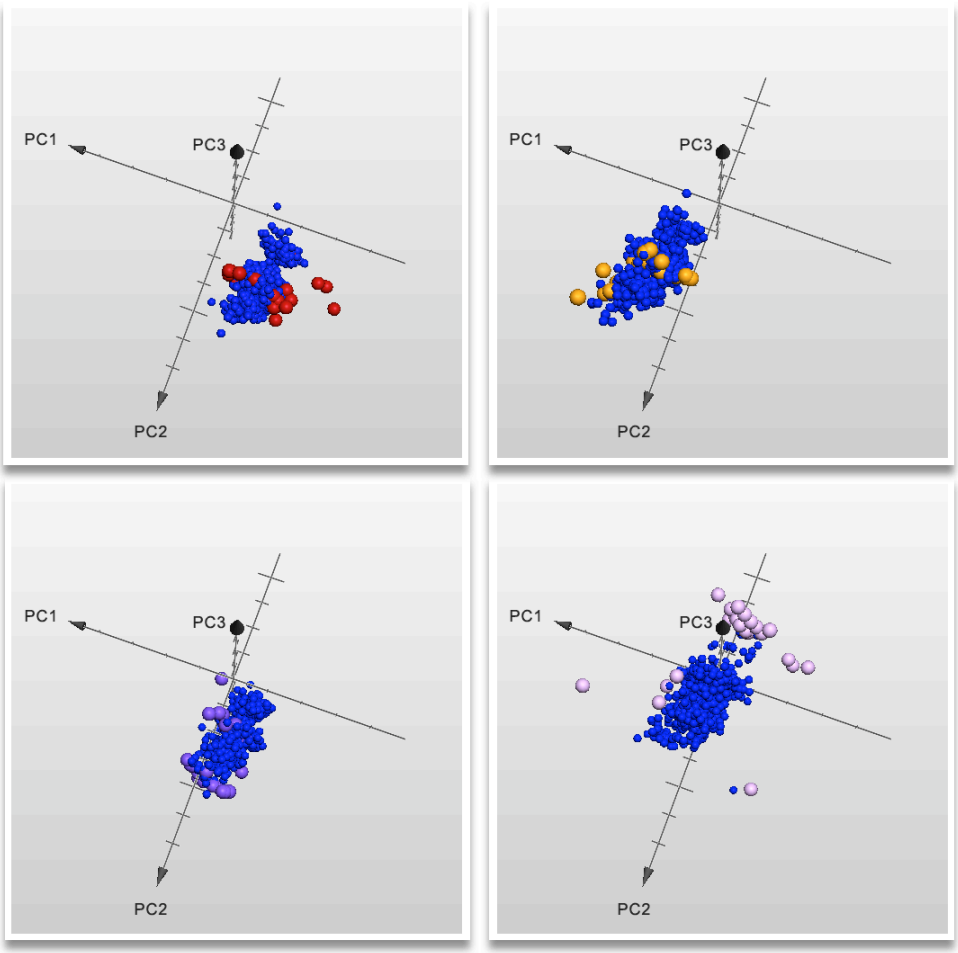


Figure 4.14. Three-dimensional projection of the *t*-scores obtained via PCA predictions based on 35 molecular descriptors for four DUD benchmarking sets. Active inhibitors of cyclooxygenase-1 are shown in red (top-left pane), estrogen receptor antagonists are shown in orange (top-right pane), Heat Shock Protein 90 inhibitors in violet (bottom-left pane), HMG-CoA reductase inhibitors are shown in light magenta (bottom-right pane). In all panes, DUD-self decoys are shown in small blue spheres. PCA was carried out using ChemGPS [548,549] and the results were plotted with Apple™ system software Grapher 2.0.

3.4. A screening set tailored for TTR

Finally, a tailored library of 2,259,573 compounds was assembled by filtering of an initial database of 10,962,930 molecules downloaded from ZINC [137]. Approximately 30 million conformers were generated with OMEGA [458] and populate the final data set to be used for virtual screening (Chapter 7).

Even though we have used multiple models that attempt to filter out compounds holding inappropriate ADMET profiles, some of the customized filter parameters used to generate our library may raise questions about the quality and safety of the included compounds. An important example is solubility. We have set a high XLogP definition to reflect the values found amongst the known TTR actives. It should be kept in mind, however, that individual descriptors are unlikely to accurately predict measurable macroscopic properties. The combination of multiple parameters and definitions is critical to obtain the expected outcomes. Moreover, some of these variables can be compensated by other alternative parameters. We have employed solubility definitions wherein all compounds predicted to be insoluble or poorly soluble are discarded.

Further assessments of molecular properties can be conducted during the post-screening stage, where screening hits predicted to hold inappropriate physicochemical features can still be deprioritised or even eliminated during the selection of the subsets of compounds to be tested experimentally.

4. Concluding remarks

"*Shoot for the moon. Even if you miss, you'll land among the stars*", author Les Brown famously said. This may sound like close enough, but given the vastness of chemical space, exploration and drug discovery need to be more meticulous and focused than that.

The concepts of molecular similarity and diversity were introduced in this central chapter as a primer to the exploration of chemical space towards the identification of new bioactive compounds. Several 1D, 2D and 3D similarity search methods were explained as complementary tools in this exploration, allowing the assembly of targeted or focused libraries of amenable size for virtual and/or experimental screening.

In this chapter, the molecular properties of the strongest TTR stabilisers were analysed and used to tweak filter definitions towards the preparation of (*i*) a benchmarking set for the evaluation of VS methods (in Chapter 5) and (*ii*) a tailored library for VS against TTR

amyloid (in Chapter 7). Together, these data sets represent the hallmark results withdrawn from this chapter.

The benchmarking set assembled for the evaluation of VS methods against TTR amyloid comprises a total of 760 compounds, of which 22 are known TTR amyloid inhibitors and 738 are carefully selected decoys molecules. Diversity is the best term to characterize the structural and physicochemical properties of both actives and decoys, resulting not only from TTR's intrinsic capacity to bind diverse compounds but also from an attempt to identify challenging decoys that match such diversity. This benchmarking set will be extensively used in Chapter 5.

The compound library assembled for VS (screening set) comprises approximately 2.3 million lead- and drug-like molecules selected through a careful filtering process. Opinions about the appropriate size of a screening library might diverge, but they should always take into consideration the type of virtual screening methods to be employed and the amount of computer power available for screening. While 2.3 million compounds can be screened overnight using 2D similarity searches, it could take years to screen using more complex methods like molecular docking. These aspects will be further discussed in Chapter 7, where the screening set prepared in this chapter will be screened by several VS protocols and using several different types of computer resources and architectures.

5. Acknowledgements

I thank ChemAxon for the free licensing of all applications included in the academic package, particularly JChem and Marvin.

Chapter 5

Evaluation of Virtual Screening methods against TTR amyloid: what we know and when we know it

"You know the truth, the brick-hard, irregular, slithery surface of truth."

[Philip K. Dick, in *Do Androids Dream of Electric Sheep?*]

1. Introduction and theory

The idea behind *in silico* screening is to use computers and software to rank virtual libraries of chemical compounds according to the probability of being active against a particular target of therapeutic interest. In principle, such prioritisation would permit that only those molecules with a considerable likelihood of showing activity would be tested experimentally. Obviously, this would reduce the costs associated with the drug discovery process [537]. With a number of success stories already reported [566], virtual screening (VS) methods are increasingly playing an important role in the field of drug discovery [559].

Traditionally, structure-based approaches are often pursuit when a crystal structure of the target protein is available. Structure-based methods usually involve docking small molecules into a receptor site of the target protein and predicting a score for the resulting interactions. Several molecular docking approaches have been introduced in Chapter 1 (page 18) and in Chapter 2. Current docking scoring methods, however, are generally poor predictors of binding affinity [419], and their ability to correctly rank candidates has been questioned [567].

When the structure of the target protein is unavailable, a ligand-based approach must be followed. There are numerous examples in the literature where ligand-based methods have been shown to outperform docking algorithms [468,496,568,569]. Ligand-based pharmacophore models and searches are perhaps the most commonly employed set of methods, often providing good results. Pharmacophore modelling has been recently attempted for the targeting of TTR through repurposing of flavonoid molecules [570]. The experimental verification of this effort has not yet been revealed. We have also devised a set of receptor- and ligand-based pharmacophore models for TTR [571]. As explained in Chapter 3, while some of our models (and particularly their combinations) seem to capture the specificities of ligand binding to TTR, their correct validation is hindered by limitations in database ranking that are intrinsic to the software available for pharmacophore-based searches. Accordingly, our pharmacophore models will be utilised only during the analysis of the VS hits put forward by protocols offering quantifiable and adequate performance (Chapter 7).

Other ligand-based approaches based on two- and three-dimensional similarity are available and require one or more known active molecules as reference. This type of methodologies was described in Chapter 4. When the structure of the target protein in complex with a ligand is available, the 3D bioactive conformation of the ligand can be

used as *template*, i.e. the reference molecule against which the database compounds are compared. Various different structure-based, ligand-centric methods often use the bioactive conformations of template ligands, including shape-matching algorithms [118], molecular field descriptors [110,121], and pharmacophore fingerprints [572–574]. Alternatively, low-energy conformations [568], or even 2D representations [117,575,576], of the known actives may be used as template ligands when no structural information is available.

One of the main challenges when using any ligand-centric approach is to find the appropriate balance between *search specificity*, which is required for target selectivity, and *search flexibility*, which is necessary for removing any dependency on the template ligand [559]. Indeed, ligand-based methods, particularly those based on simple molecular topology, are often criticised for finding close structural analogues instead of discovering novel structures (a concept known as *scaffold hopping*). Scaffold hopping is a highly desirable outcome of VS because it not only provides ways of improving the pharmacological properties of known lead compounds, but it also allows the exploration of unpatented regions of chemical space.

VS methods are only truly useful for scaffold hopping if they are able to retrieve actives that are both lead-like and structurally novel, and a major consideration when demonstrating this in a retrospective validation is the data set of actives and decoys used. As discussed in Chapter 4, it is important that the decoys have similar physicochemical properties to the actives, otherwise the actives may simply be distinguished from the decoys based on trivial differences, such as molecular weight (MW) or logP [577]. Huang et al. recently constructed the Directory of Useful Decoys (DUD), in which the decoys were specifically selected to be physicochemically similar yet topologically dissimilar to the actives, thereby providing a more realistic scenario for the validation of ligand-based VS [543].

The goal of this chapter is to test and compare the VS performance of several VS protocols designed to discover active TTR stabilisers seeded in a benchmarking set of 738 carefully selected decoys. The proposed VS protocols employ four distinct types of VS techniques: (*i*) 2D similarity searches using Chemical Hashed Fingerprints and Pharmacophore Fingerprints, and UNITY 2D fingerprints; (*ii*) 3D searches based on molecular shape, chemical complementarity and electrostatic similarity; (*iii*) LigMatch, a new ligand-based method combining 2D pre-selection from multiple templates with 3D geometric hashing; (*iv*) molecular docking into consensus TTR X-ray structures selected through the cross-docking studies presented in Chapter 2 (recall page 100).

1.1. Benchmarking and performance evaluation in VS

Benchmark studies are essential to evaluate the accuracy of in silico methods aimed at discriminating active molecules from inactive ones. As new VS methodologies, data mining and machine learning methods are adapted for the ranking of compound databases, the chemoinformatics and VS literature is being overwhelmed with performance evaluations in reasonably unified settings. Nevertheless, problems in judging the relative performance of one method on the basis of previously published reports are often encountered. Jain and Nicholls noted that "a serious weakness within the field is a lack of standards with respect to statistical evaluation of methods, data set preparation, and data set sharing" [578]. Moreover, Nicholls underlined the fact that standard procedures commonly used for the analysis of benchmarks in other fields are rarely followed in VS reports [564]. The main issues mostly concern (*i*) the experimental design, (*ii*) the preparation of data sets, (*iii*) the performance metrics, and (*iv*) the variance of analysis. In the author's own words, benchmark studies often become "anecdotal instead of systematic".

Given their relevance within the context of the present chapter, in this subsection the first two issues are briefly addressed and, in the next subsection, some of the problems surrounding the performance metrics in common use are reviewed. Point (*iv*) is more relevant within the context of Chapter 6 and will therefore be covered in the next chapter. Nonetheless, for a in-depth grasp of all these topics I highly recommend references [564,578,579].

1.1.1. Experimental design

The very first question one needs to answer when designing a VS benchmarking exercise is what it is going to demonstrate. Indeed, the hypothesis under investigation should be clearly stated: are we trying to prove that one particular methodology is "better" than others? What will we learn from the study?

The lack of a clear design goal in a retrospective VS evaluation is often reflected by the somewhat subjective way in which the decoy molecules are selected. The ultimate goal of VS boils down to a simple classification problem, holding an auspicious analogy to a police line-up [564]. Is compound (subject) X an *active* ("guilty") or an *inactive* ("innocent")? In a general assessment, there is a small set of known actives (the "suspects") and one intends to see if a method can discriminate them from a second set of presumed inactives (the so-called "fillers", or simply "innocent"). In order to ensure that arbitrary picking does not compete with real recognition, a minimum number of decoys (or innocent) is

required. On the other hand, this number cannot be so large that activity (or "guilt") is concealed within the statistical variance of the decoys (or the innocent).

Moreover, the inactivity (or "innocence") of background subjects is most often assumed through the lack of experimental activity data (or "evidence"). However, the utilisation of nondrug or lead-like decoys that are easily segregated from drug-like compounds may jeopardize the entire purpose of a benchmark. In fact, in such cases even rudimentary techniques (e.g. molecular weight filters) might be able to produce *artificial* enrichments of active compounds in database selection sets [107]. In other words, the "fillers" need to be convincing: dissimilar to the guilty (normalized via structure-unaware descriptors like molecular weight, hydrogen-bond acceptor/donor counts), but not too similar to the guilty or between each other (thus yielding false false positives).

Nicholls classified VS experiments into four types, each with its own purpose and all including the application of various filters, just before an ultimate step of random selection of decoys [564]:

i. *Universal*. Any compound available to be screened, either from chemical suppliers or corporate catalogues.

ii. *Drug-like*. Accessible and drug-like, often by employing simple filters.

iii. *Mimetics*. Accessible, drug-like and matched to *known* ligands by simple physical properties.

iv. *Modeled*. Accessible, drug-like and derived through modelling techniques on known ligands of the proposed targets.

The level of "challenge" posed by the decoys naturally increases from (*i*) to (*iv*).

The *universal* selection is based on a random withdrawal of decoys from a compound collection of a particular vendor, thus testing a method's ability to discriminate known actives without prior knowledge of the determinants of activity. However, this approach poses two problems. First, it is very unlikely that a particular vendor catalogue is able to cover the vastness of chemical space; secondly, it is likely that a random selection will include certain "characters" (in this case, decoys) that will hold idiosyncratic features and thus stand out amongst the "crowd". The latter problem has been portrayed as the 'dog' test [580]: if a dog could distinguish the known actives from the inactives what would be proved? As denoted by Nicholls, potentially a lot, but only if the rest experiment is designed with the choice of decoys in mind [564]. Indeed, the real problem is one of dynamic range. If the decoys are easily distinguishable, the only way of achieving sufficient statistical power when assessing multiple methods is to repeat the same tests multiple

times. Yet, how many times can one actually repeat the same test? The statistical error associated with the performance metrics discussed below varies not only with the number of decoys but also with the number of actives. While it is easy to (randomly) find extra decoys (presumed inactives), the number of known actives is most often very limited. This limitation hinders random decoys from being an effective experimental design.

The second approach (*drug-like* selection of decoys) is perhaps the most common and, as the name implies, consists in (randomly) withdrawing decoys from a library of "drug-like" compounds, typically assembled by following well-known rules of *drug-likeness* such as the Lipinski's Rule-of-Five. What really distinguishes this approach from the *universal* one is the idea of simulating modern drug screens that focus on compounds that look like drugs. By doing so, it tries to circumvent the second problem, assuming that all decoys look like drugs and are likely to be inactives, but the issue of limited chemical space coverage remains. In fact, most drug-like libraries are strongly biased towards specific targets that may (or may not) be linked to the retrospective experiment at hand.

The third type of experiment consists in finding *mimetic* decoys, i.e. compounds that will not be easily distinguishable from the known actives by "simple" means. The most common protocols to construct mimetic decoys include matching the physical properties of the known actives, for example their molecular weight, number of hydrogen bond acceptors and donors, lipophilicity, flexibility, etc. [421]. Recently, Irwin et al. set off community-dedicated activities towards the standardization of decoy sets, giving rise to the Directory of Useful Decoys (DUD) for docking and other VS methods [543]. Thirty-six decoys were carefully selected for each known active of each proposed target, employing a protocol identical to the one used in Chapter 4 for the assembly of a TTR-tailored benchmarking set (recall **Figure 4.3** on page 147). These decoys outline *mimetic* sets referred to as DUD-*self*, whereas the combination of all decoys across all targets form a *drug-like* set referred to as DUD-*all*. It is interesting to notice that, even though the DUD-*self* decoys are often frustratingly effective, they are not always harder to distinguish than DUD-*all* decoys. The bottom line here is that mimetic decoys are not certain to prove that expectations conform to reality for VS techniques that attempt to grasp the physics of drug-target interactions [564].

The fourth approach to decoy set construction is based on the modelling of decoys directly against the VS method under evaluation. Taking docking and scoring as an example, *modeled* decoys might try to reproduce a good shape complementarity with the target binding site, while avoiding electrostatics complementarity. This would of course demand a rigorous protocol to be followed and described (i.e. details about protein and ligand preparation, protonation, etc. must be given). Moreover, it is plausible to expect that this approach would most often cause random performances to be observed, leading

to the conclusion that docking scoring functions are worthless. Clearly, even more than *mimetic* decoys, *modeled* decoys are meant to stress-test VS methods, not to measure real-world performance. For all these reasons, this fourth approach is often avoided.

With the above in mind, what is then the most appropriate design to choose? If time and computers resources (mainly, storage) are not an important part of the equation, there is no reason why a retrospective evaluation cannot include all kinds of decoys. Researchers can indeed report the performance of the VS methods under evaluation for multiple subsets, clearly identifying the respective intent. If large sets are included, they can be provided as supplementary material and thus be compared or used by others. At this point, it is critical to emphasize that the goal of this chapter is to assess and contrast the performance of multiple VS protocols against one particular target, rather than to extract broad conclusions about the effectiveness of the methods against many targets in the VS realm. By contrast, in Chapter 6 the application of specific docking and scoring protocols to multiple targets is studied. Accordingly, specific subsets within the DUD data set will be explored, providing multiple bases for comparison.

Perhaps as important as the abovementioned *intensive* properties (i.e. properties that are intrinsic to the experimental design), are the *extensive* properties that include, for instance, the number and the proportion of actives and decoys to be used. The key question one again needs to ask is: what do we want to know? Indeed, if one intends to assess an isolated VS method on a single target, these properties will significantly differ from a case where broad assessments of the efficacy of several VS methods against multiple targets are attempted. Nicholls illustrated this with basic error analysis, while attempting to address the question of how variable the performance of VS methods is [564]. In a neat syllogism, the author demonstrates that the variances associated to the high variations in performance across the 40 DUD targets are intrinsic properties to "targets", "actives" and "inactives". In general, the observed variances linked to both actives and inactives are identical, which also allows extracting insights about the adequate proportion of actives and decoys. Moreover, following the Central Limit Theory and using the measure of the performance (the area under the ROC curve, or AUC) of FRED (with the Chemscore scoring function) across the entire DUD, Nicholls checked whether the source of the variance across targets was due to inadequate sampling of actives and decoys, or just an intrinsic property of four VS methods. The three main conclusions withdrawn from his study of extensive properties were:

i. A proportion of 4:1 between decoys and actives is associated with small error increase (11%) to the limiting value from an infinite number of inactives. Therefore, even though the number of actives is very important when evaluating the per-

formance of a VS method on a single system, the number of inactives does not need to be significantly larger.

ii. The contributions to the variance from a limited number of actives is almost negligible compared to inter-target variation. Thus, the number of systems required to evaluate the performance of a particular VS method against other methods with a 95% confidence must be very large (even larger than DUD).

iii. The number of actives per target required to evaluate the performance of a particular VS method against other methods with a 95% confidence does not need to be very large (a number of 10 actives might suffice), given the variance between systems. Therefore, the inclusion of only representatives of chemical classes might be statistically acceptable.

It is important to underline that there are no definite or "correct" answers to the design of VS experiments, and very often their goals lie between the "standardized" examples found in the literature. As aforementioned, the work reported in this chapter was not fuelled by a need to extract generally applicable conclusions about the efficacy of the VS protocols under study, but the need to evaluate their ability to discriminate TTR actives from inactives. Above all, the results presented in this chapter allow us to identify which protocols are more likely to enrich database selection sets with compounds (virtual hits) that are able to stabilise the native form of TTR and thereby prevent amyloid fibril formation. Of course, to be "better" does not necessarily mean to be "good". One of our goals was that the results could also provide a clear indication of whether our VS protocols could actually provide promising hits within reasonable size limits of the database selection sets. In order to achieve this goal, comparisons with successful examples reported in the literature are not unavoidable but highly desirable. This prompted us to follow a *mimetic* approach similar to the one proposed by Huang et al. in the construction of DUD [543], resulting in a benchmarking data set with identical proportion of actives and decoys (36 decoys per active). For more detail about the benchmarking set, recall page 159 and subsequent pages in Chapter 4.

1.1.2. Preparation of data sets

The choice of decoys is a critical aspect in the assessment of VS performance, but so is the assignment of compound activity classes. Recently, VS artefacts related with certain activity classes have emerged in the literature under terms like *analogue bias* and *artificial enrichment* [581–583]. In order to illustrate this problem, let us imagine a scenario wherein a decoy is repeated by mistake in the benchmarking data set. The gain of infor-

mation is null, but the number of decoys has increased and the associated error virtually decreased. However, the actual error has not been reduced. In the police line-up analogy, this would roughly correspond to a situation where identical twins are used as innocent. As surreal as this scenario may seem, it forces us to go back to the concept of molecular *similarity* introduced in Chapter 4. Although there is no unique definition for molecular similarity, it must be a critical aspect of the selection of the active population from congeneric series. Accordingly, techniques that rely on chemical similarity (such as 2D and 3D ligand-based methods) are more likely to be affected by biases due to failure in making independent measurements. Independence towards extensive variables is indeed a critical characteristic to pursuit both during the preparation of a decoy set and when choosing the metrics to report performance.

Good and Oprea proposed a solution to the problem of analogue bias, based on the narrowing of active populations to a smaller set of representative compounds [583]. Clark, on the other hand, suggested two weighting schemes (*arithmetic* and *harmonic*) for the active compounds retrieved within a compound series, as a way of distinguishing biased from unbiased screening statistics [584]. More recently, Rohrer and Baumann studied the problem of artificial enrichment and developed a publicly available resource for "Maximum Unbiased Validation" (MUV) of virtual screening methods [585]. Making use of a bulk descriptor space based on simple atom counts and following the principle that active-active distances in that descriptor space should be at least as large as active-decoy distances, the authors built a number of benchmarking sets of active and inactive compounds.

Besides analogue bias, more general sources of bias can contribute the phenomenon of artificial enrichment. The Similar Property Principle introduced in Chapter 4 can be regarded, in concept, as a source of *inductive bias*. The link between chemical similarity and similar activity is mirrored in compound libraries because molecules are synthesized in a biased manner, in an attempt to be similar (e.g. bioisosteres) to chemotypes known to be active. In a way, Medicinal Chemistry disseminates its basic principles to the creation of compound series that are guided towards the biologically relevant chemical space. In summary, **Figure 5.1** illustrates the concepts of artificial enrichment, analogue bias and inductive bias, showing how they relate to each other.

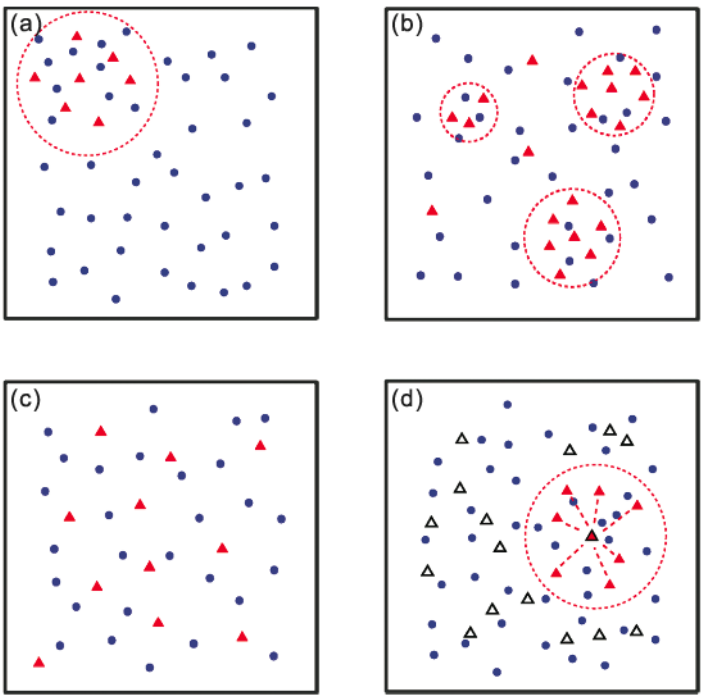


Figure 5.1. Schematic representation of the distribution of sets of active compounds in simple chemical space representations, illustrating the concepts of artificial enrichment, analogue bias and inductive bias. (a) When the screening library does not reflect the chemical character of active compounds, as assessed using molecular descriptors, the actives will only occupy a small fraction of the chemical space that is covered by the library. This makes it easy to separate active from inactive compounds, leading to *artificial enrichment* in virtual screens. (b) When active compounds consist of analogue series, the virtual screening using subsets of these analogs as reference compounds will easily detect remaining analogs. This phenomenon is known as *analogue bias*. (c) Appropriate compound data sets for virtual screening that mimic "real life" application scenarios should be composed of diverse active compounds that are widely distributed over the space covered by the screening library. (d) Inductive bias means that of the possible chemical space (blue dots) and potentially available biologically relevant chemical space (black triangles), one generally only explores regions already known to be activity relevant (black/red triangle) and chemically extrapolates from such regions (red triangles). This figure was adapted from reference [586].

In order to eschew the problem of artificial enrichment, in this project we have explored the clues provided by Good and Oprea [583] to prepare a benchmarking set tailored for TTR amyloid. Thus, instead of including all known potent TTR stabilisers in the active population, clustering by maximum common substructure was employed to ensure that only representatives of the active compounds were considered. Nevertheless, given the high structural diversity characterising TTR ligands, the benchmarking set used in this chapter contains 22 diverse actives, along with 738 property-matched decoys. For more details about the benchmarking set, the reader should review page 159 in Chapter 4. Fur-

thermore, in Chapter 6 we explore a weighting scheme similar to the arithmetic weighting proposed by Clark [584], in order to circumvent artificial enrichment while reporting results using the DUD data set.

1.2. Performance metrics

Assuming one particular experimental design, what quantities best describe the performance of a VS method? A myriad of metrics has been suggested to quantify the performance of VS models and protocols, some of which holding general application in quite different areas of data mining and machine learning. In the following subsections, some of the main metrics will be discussed. Emphasis will be given to the metrics that facilitate the detection of active compounds at the top of a ranked data set – a concept known as *early performance* or *early behaviour*.

1.2.1. Sensitivity and Specificity

Before delving into the metrics that are used to assess the performance of VS methods, two important concepts must be first introduced – sensitivity and specificity. These concepts are key characteristic qualities of any test that aims at classifying two populations.

In the context of VS, sensitivity (*Se*) corresponds to the fraction of true active molecules selected by the VS protocol under evaluation, i.e. the number of true positive results (*TP*) divided by the sum of true positives and false negatives (*FN*):

$$\mathrm{Se} = \frac{N_{\text{selected actives}}}{N_{\text{total actives}}} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

Equation 5.1

Se varies between 0 (when all actives are discarded) and 1 (when all actives are selected), thus providing insight on the active compounds that will be missed (*FN*). The lower is the sensitivity of a method, the higher will be the number of *FN* and the worst will the method at identifying actives.

By contrast, specificity (*Sp*) corresponds to the fraction of true inactive molecules correctly discarded by the VS protocol under evaluation, i.e. the number of true negative results (*TN*) divided by the sum of true negatives and false positives (*FP*):

$$\mathrm{Sp} = \frac{N_{\text{discarded inactives}}}{N_{\text{total inactives}}} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

Equation 5.2

Sp also varies between 0 (when all inactives are selected) and 1 (when all inactives are discarded), thus providing indication about the inactive compounds that will be inappropriately classified (*FP*). The lower is the specificity, the higher will be the number of *FP* and the worst will be the VS protocol at discarding inactive compounds.

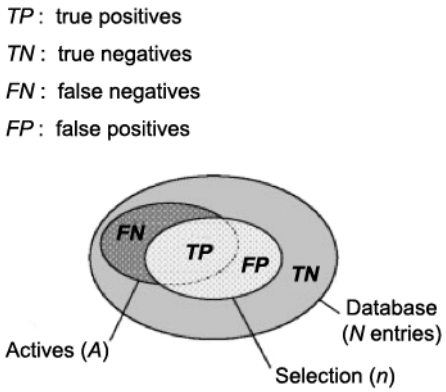


Figure 5.2. A simple diagram to help understanding the concepts of sensitivity and specificity. In a database comprised of a total of *N* entries, amongst which *A* molecules are active on the investigated target, the VS protocol selects *n* compounds as being actives. This figure was extracted from reference [475].

The diagram provided in **Figure 5.2** illustrates the relationship between different classifiers, which helps understanding the concepts of *Se* and *Sp*, and also facilitates the comparison between the most common metrics reported in **Table 5.1**.

Table 5.1. Metrics for performance evaluation commonly used in virtual screening.

Metric	Expression
Sensitivity	$Se = \frac{TP}{A} = \frac{TP}{TP + FN}$
Specificity	$Sp = \frac{TN}{N - A} = \frac{TN}{TN + FP}$
Yield of actives	$Ya = \frac{TP}{n}$
Enrichment	$E = \frac{TP/n}{A/N}$
Statistical significance	$S = \sum_{k=TP}^{A} \frac{\binom{A}{k}\binom{N-A}{n-k}}{\binom{N}{n}}$
Accuracy	$Acc = \frac{TP + TN}{N} = \frac{A}{N} Se + \left(1 - \frac{A}{N}\right) \cdot Sp$

Metric	Expression
Discrimination ratio	$DR = \frac{TP/A}{TN/(N-A)} = \frac{Se}{Sp}$
"Matthews" correlation coefficient	$C = \frac{TP \cdot TN - FN \cdot FP}{((TN+FN)(TN+FP)(TP+FN)(TP+FP))^{1/2}}$

Different metrics provide different pictures about the performance of a VS protocol at identifying active compounds or discarding inactive ones. The enrichment and the yield of actives are perhaps the most common: the enrichment discloses how many times the VS protocol performs better than a random selection, whereas the yield of actives provides an indication of the hit rate one would get if *n* selected molecules were screened (for a second time).

1.2.2. Properties of metrics used in VS evaluations

An obvious problem with the high diversity of the available metrics is incomparability, or, as bluntly put by Nicholls, the "apples and oranges" problem [564]. There are always two sides to a coin. On the one hand, one may agree that a good metric is the one that everyone uses. A good example is the root-mean-square deviation (RMSD): everyone knows there are problems and limitations to the use of the RMSD, but everyone continues using it as a standard measure (because it allows the comparison of results). On the other hand, one may even recognize that the answer to the problem depends on the complete sharing of all data, instead of the imposition of a specific standard, but a scenario where screening data is openly shared is still far from sight.

Nicholls proposed a set of required properties for a good metric [564]:

i. independence to extensive variables (as formerly mentioned);
ii. robustness;
iii. straightforward assessment of error bounds;
iv. no free parameters;
v. easily understood and interpretable.

Some of these characteristics will be discussed on the following subsections while the reader is introduced to the VS performance metrics used throughout this project.

1.2.3. Early performance

The enrichment factor (EF) is commonly used to compare VS results [583], expressing the proportion of active molecules retrieved by a VS method from a chemical database at defined intervals (i.e. at X%) of the VS ranking:

$$EF_{(X\%)} = (\text{Fraction of Actives Recovered}) / (X/100)$$

Equation 5.3

or

$$EF = (TP/(TP+FP)) / ((TP+FN)/(TP+FP+TN+FN))$$

Equation 5.4

The EF is particularly useful to quantify the early enrichment provided by the method, for example at the top 1 % and 2 % levels, which is more likely to reflect a drug discovery process guided by VS, wherein only a very small percentage of the screening library needs to be experimentally tested. It has been recognized that VS methods should provide at least a 10-fold increase in enrichment against a random selection in order to hold a reasonable chance of identifying true hits [496].

Because of its functional form, however, a major shortcoming of the EF metric is that it depends on the relative number of actives and decoys in the database: the enrichment becomes smaller if fewer inactives are initially present in the benchmarking set. Indeed, the "Fraction of Actives Found" term does not hold property (*i*) proposed by Nicholls.

To quantify early performance, in this project we used ROC enrichment (ROCE), which avoids this dependency by expressing the percentage of actives recovered as a proportion of the percentage of decoys recovered, rather than as a proportion of the percentage of all compounds observed [587,588].

$$ROCE_{(X\%)} = (\text{Fraction of Actives Recovered}) / (\text{Fraction of Decoys Recovered})$$

Equation 5.5

or

$$ROCE = (TP/FP) / ((TP+FN)/(FP+TN))$$

Equation 5.6

This modification alone makes the enrichment independent of extensive quantities, more robust, accessible to analytic error approximation, and suffering only a minor loss of in-

terpretability. As with the EF, a ROCE greater than 1.0 represents enrichment relative to a random selection.

1.2.4. ROC curves and enrichment curves

The receiver operating characteristic (ROC) curve reports the evolution of *Se* as a function of (1 – *Sp*) as the detection threshold (i.e. the size of *n*) varies. In analogy to signal detection theory developed during World War II (a pioneering use of the ROC curve), *Se* is the perceived signal (here the activity) and (1 – *Sp*) relates to the detected background "noise" emitted by inactive molecules. ROC curves have been recognized as an adequate approach for characterizing the ability of a VS workflow to distinguish active molecules from inactive ones [587,588]. **Figure 5.3**c provides an illustration of the construction of one of such curves. The 45° diagonal rising from the origin to the upper-right corner represents a random selection of the molecules, whereas the ROC curve above the diagonal illustrates the performance of a VS protocol capable of detecting the correct signal, i.e. of discriminating active compounds over inactive ones. In an ideal distribution (where actives are totally segregated from inactives) the ROC curve takes off vertically from the origin towards the upper-left corner (*Se* = *Sp* = 1) and then migrates horizontally to the upper-right corner. In other words, the more a ROC curve bows towards the upper left corner of the graph, the clearer the signal looks.

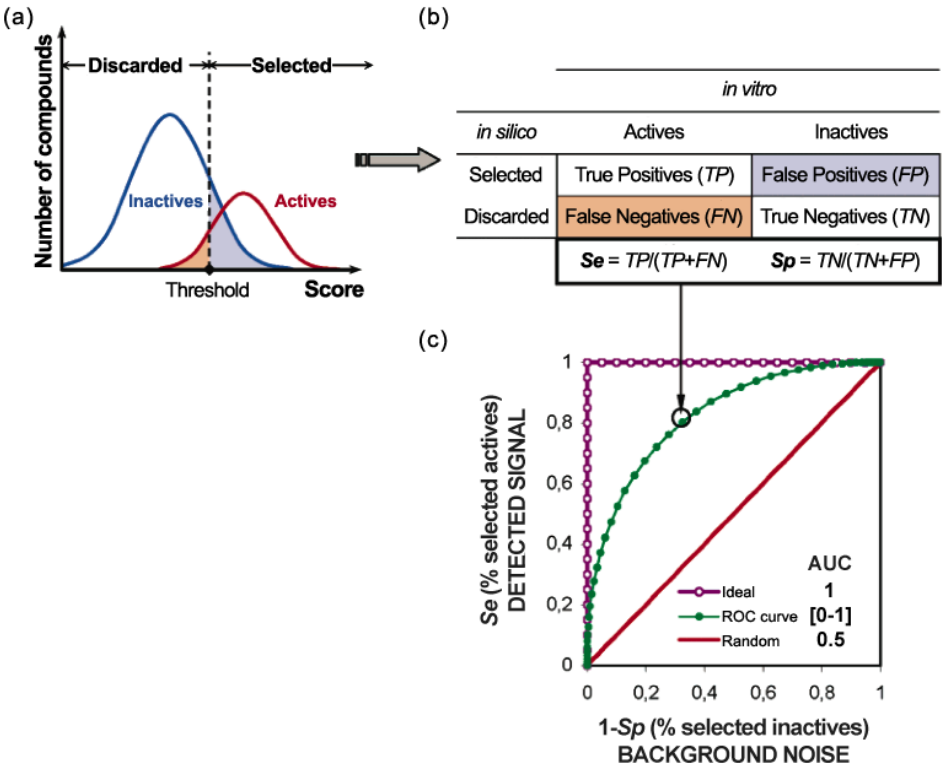


Figure 5.3. Receiver operating characteristic curves in a nutshell. (a) Theoretical distributions of scores obtained for both active (red) and inactive compounds (blue) using a given VS protocol. For the sake of simplicity, it is assumed that the scores assigned to both active and inactive compounds follow a Gaussian distribution, even though this is unlikely to happen in a real setting. Typically, there is a certain amount of overlap between the two distributions, revealing false predictions (coloured areas). Upon modification of the selection threshold (dashed line), the fractions of such erroneous classifications vary considerably. (b) Confusion matrix illustrating the relationship between the elements of a computer test attempting to classify two populations. Sensitivity (*Se*) and specificity (*Sp*) are also contextualized. (c) Construction of a traditional ROC curve: for all possible selection thresholds, the progression of the inferred *Se* and *Sp* is reported on a ROC plot, *Se* as a function of 1 – *Sp*. The area under the ROC curve is another commonly used metric to quantify the overall performance of the VS protocol. This figure was adapted from reference [587].

In the context of VS, at least, *Se* and *Sp* can only assume discrete values, and the corresponding confusion matrices are filled with integers. As a result, as the selection threshold varies, the inclusion of a true positive will insert a vertical shift to the ROC curve, while the inclusion of a false positive will produce a horizontal dislocation. As a result, the ROC curves used to evaluate VS methods usually look rather serrated and "bumpy", compared with the one presented in **Figure 5.3**c. Nevertheless, because the relative positions of the ROC plots provide indication of the accuracy of the method, the area under the ROC curve (ROC AUC) is often used to measure the global performance of VS protocols. When the ROC AUC is close to 0.5, the method under evaluation is accomplishing a near-random selection and its discriminative power is poor. On the other hand, when the ROC AUC ap-

proaches 1, the VS protocol efficiently discriminates actives from inactives, corresponding to an ideal scenario. Caution should be taken, however, when interpreting the meaning of AUC values. ROC plots disclose the intrinsic quality of a given VS protocol and not the quality of any given molecules. For instance, an AUC value of 0.7 only implies that a randomly selected active will have a better score than a randomly selected inactive 7 times out of 10.

Inspired by the ROC curve approach, VS researchers developed a useful graphical method that represents the evolution of enrichment as the database selection threshold changes. The resulting plots are known as *cumulative recall curves* (or simply *enrichment curves*) and hold a similar outlook as the ROC plots (see **Figure 5.4**). Enrichment curves report the yield of actives (i.e. the actual *Se*) as a function of the fraction of the ranked database, and not as a function of (1 – *Sp*) as in "traditional" ROC curves.

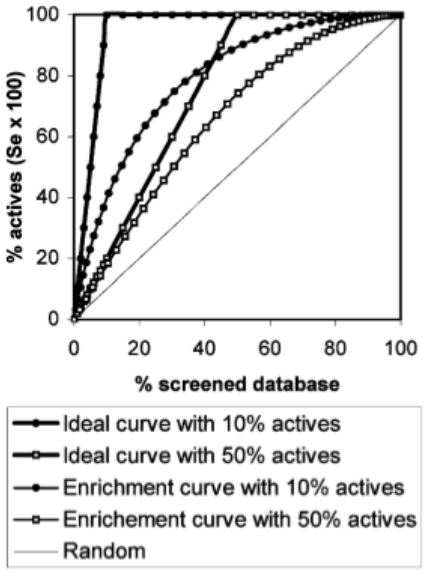


Figure 5.4. Enrichment curves for the same VS protocol using two alternative benchmarking sets with different ratios of actives versus inactives. The plots illustrate the difficulty that arises when comparing benchmarking sets holding significantly different actives vs. decoys ratios, due to the fact that enrichment curves are influenced by this extensive property. This figure was extracted from reference [587].

Like with the enrichment factor (EF), enrichment curves show dependence to extensive variables. Indeed, as disclosed in **Figure 5.4** ideal enrichment curves directly depend on the ratio of actives in the database. Consequentially, they are limited to narrower assessment space between ideal and random performances, especially when the number of actives becomes too close to the number of decoys. On the other hand, enrichment curves are straightforward to plot and provide a *real life* view of both early and overall performance of the VS protocols under evaluation (when performing VS on a given screening data

set in real life, investigators do not know at the outset which compounds are active or inactive).

Since the main goal of this chapter is to evaluate multiple VS protocols in a real life case – the screening of novel TTR amyloid inhibitors – we made use of enrichment curves to assess and contrast both their early and overall behaviour, thus guiding the fine tuning and the selection of the best VS protocols for large-scale VS campaigns. Because one unique benchmarking set (comprised of known TTR stabilisers and carefully selected *mimetic* decoys) was used, the dependence on extensive variables affecting enrichment curves is not relevant to the interpretation of the results. The area under the enrichment curves (AUC) was also calculated for each protocol, providing a measure of the VS performance across the whole data set. Since the known TTR actives represent only 3% of the entire data set, like with ROC curves ideal distributions of actives and inactives will approach 1.0, whereas random behaviour will approach 0.5.

2. Computational methods

Ligand-centric VS methods based on 2D (along with their combinations) were introduced in Chapter 4, starting from page 142. Molecular docking and scoring methods were introduced earlier in the first two chapters of this thesis, within the context of receptor-based studies of TTR-ligand interactions. In this chapter, we combine all insights extracted thus far to assemble a large set of VS protocols aimed at the identification of novel TTR amyloid inhibitors. All protocols were submitted to successive rounds of performance evaluations, allowing not only their fine-tuning but also a rational selection of the protocols with higher likelihood of detecting active TTR stabilisers. Throughout this section, the key details about the tested protocols are succinctly described and the results of their performance evaluation are provided in the next section.

2.1. 2D similarity searching with ScreenMD and UNITY

Thirty-two VS protocols based on 2D similarity were setup using ChemAxon's ScreenMD and Tripos' UNITY data miner. First, chemical hashed fingerprints (CHF) and 2D pharmacophore fingerprints (2D-PF) were generated for all ligands in the TTR benchmarking and screening sets, using GenerateMD. The two types of 2D descriptors were used to compute compound similarities against the four query structures (or template ligands) selected in Chapter 4 – phenox, benzoxazole, PCB18 and thyroxine (see page 158). Two different metrics of similarity were employed: the Euclidean distance and the Tanimoto

coefficient. Both GenerateMD and ScreenMD are part of ChemAxon's JChem package [542]. Additionally, UNITY was used within SYBYL [457] to generate 2D standard screens (UNITY 2D fingerprints) and perform 2D similarity searches on the benchmarking and screening sets. Moreover, the four concatamers assembled in Chapter 4 (see page 158) were also used individually both for evaluation (in this chapter) and screening purposes (in Chapter 7).

2D-similarity search parameters were passed on to program ScreenMD via XML-format files defining the type and the characteristics of the fingerprints used in 2D similarity searches (see **Figure 5.5**). These configuration files are relatively simple because they only contain three critical fingerprint attributes: Length, BondCount and BitCount. The Length parameter corresponds to the number of bits in the fingerprint (usually a multiple of 32); BondCount sets the longest path taken into account, and BitCount defines the number of bits to be turned into 1 for each identified feature. In this work, three different fingerprint lengths were tested: 512 (the default length), 1024 and 2048. The BondCount and BitCount parameters were kept at their defaults, respectively 7 and 3.

Besides fingerprint attributes, ChemAxon's XML configuration file stores definitions of the metrics to be used in the screenings, along with their corresponding *dissimilarity* thresholds. Structures with a dissimilarity value below the defined threshold are retained during the screening process, forming the so-called *hit set*. Because compounds holding a dissimilarity value higher than this threshold are discarded from the list, we opted for setting the threshold at its maximum allowable value (1.0 for the Tanimoto coefficient and 100 for the Euclidean distance), so that every compound comprised in the sets could be listed with an associated dissimilarity value. This is particularly important for the construction of ROC and enrichment curves, where all database compounds must have an associated score. In the case of ScreenMD, the scores were easily obtained by calculating the inverse of the dissimilarity values (the scores thus expressing the level of similarity to the template ligands and concatamers).

```
<?xml version="1.0" encoding="UTF-8"?>

<ChemicalFingerprintConfiguration Version ="0.3"
        schemaLocation="cfp.xsd">
        <Parameters Length="1024" BondCount="7" BitCount="3"/>
        <StandardizerConfiguration Version ="0.1">
        <Actions>
            <Action ID="aromatize" Act="aromatize"/>
        </Actions>
</StandardizerConfiguration>

<ScreeningConfiguration>
        <ParametrizedMetrics>
        <ParametrizedMetric Name="Tanimoto" Active Family="Generic"
        Metric="Tanimoto" Threshold="1.0"/>
        <ParametrizedMetric Name="Euclidean" ActiveFamily="Generic"
Metric="Euclidean" Threshold="100" />
</ParametrizedMetrics>
</ScreeningConfiguration>
</ChemicalFingerprintConfiguration>
```

Figure 5.5. A 2D-similarity search parameters file given as input for 2D similarity searches with ChemAxon's ScreenMD. The three most important attributes are Length, BondCount and BitCount. Illustrated in this example is the set of attribute values that provides best VS performance on TTR's benchmarking set.

2.2. 3D similarity searches

The Similarity Property Principle introduced in Chapter 4 has been questioned by some experiments [589,590]. Doucet et al., for example, performed search of similarity between compounds binding the adenosine A1 receptor and reported conflicting results. Even though the molecules in their study looked structurally similar, they exhibited significant differences in terms of steric, hydrophobic and electrostatic features [590]. Indeed, these differences are often exposed by means of 3D superimposition of the molecules under comparison. Pepperrell et al. conducted a comparison of a 3D search method with a 2D similarity search method on a database of 4500 structures [591]. The outputs differed significantly and highlighted some complementarities of the two substructure searching systems. Several 2D hits showed little overall similarity to the query (target) molecule, even though containing the same substructure. On the other hand, 3D hits presented strong topographical resemblance yet showing great difference from the target in terms of topology (see **Figure 5.6**).

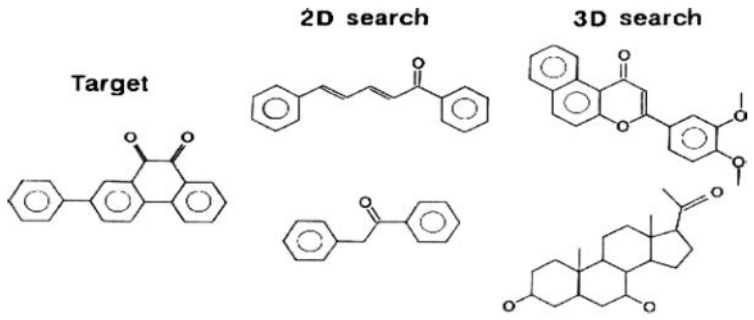


Figure 5.6. Difference in two top hits for the same template (target) molecule in 2D and 3D searches conducted by Pepperrell [591].

In recent years, the problems of the Similarity Property Principle have been addressed [592]. In this paper, the authors provided examples where a small modification in the chemical structure leads to drastic changes in physicochemical properties and biological activity. Their observations are in agreement with the so-called *Similarity Paradox* proposed by Bajorath, who stated that even if similar structures have generally similar activity, minor modifications can make active molecules lose their activities completely [593].

Biological activity is usually the result of the interplay of several complex processes, which cannot be easily depicted by a set of linear relationships. Non-linear variable mapping, where the activity is represented by a function of structural, topological and molecular descriptors, must often be used in order to describe those processes more accurately. One of the first prospective application of a 3D search method in the virtual screening of therapeutic compounds aimed at identifying novel scaffolds for inhibitors of a bacterial protein-protein interaction of therapeutic interest [118]. The program used in this study is called ROCS and, more recently, it was tested against a data set of 21 protein systems and performed at least as well as the seven competing docking programs [568].

2.2.1. Shape and chemical similarity searching with ROCS

Shape complementarity is a crucial determinant of protein-ligand interactions. In this project, 3D-similarity searches were performed using ROCS version 2.3.1 [594] (ROCS is the acronym for "Rapid Overlay of Chemical Structures"). ROCS is a 'ligand-centric' shape-based fitting method that represents molecular volume using a set of atom-centred Gaussian functions, and aligns molecules by optimising the overlap of their volumes. [118]. In general, the overlap of shape between two molecules (or any two objects) A and B, $O_{A,B}$, is given by

$$O_{A,B}(\vec{q}^{A},\vec{q}^{B}) = \int\int\int \chi^{A}(\vec{r},\vec{q}^{A})\chi^{B}(\vec{r},\vec{q}^{B})\, d\vec{r}$$

Equation 5.7

Here, *r* is a position in space, *q* is a set of variables that determine orientation and position, and χ is the *characteristic volume* function. The volume integral emerging from equation x is over the entire space (d*r* = d*x* d*y* d*z*). Determining $O_{A,B}$ allows for the calculation of a fundamental quantity between any two molecules: the shape distance, $D_{A,B}$ which is given by

$$D_{A,B} = \sqrt{O_{A,A} + O_{B,B} - 2O_{A,B}}$$

Equation 5.8

This property is a true metric (obeying the triangle inequality) and predicts that shape is actually an intrinsic, nota an extrinsic or relative property. One can also produce a related property that has more familiarity, a Tanimoto, $T_{A,B}$:

$$T_{A,B} = O_{A,B}/(O_{A,A} + O_{B,B} - O_{A,B})$$

Equation 5.9

This shape Tanimoto (ST) has the recognizable quantity that it is 1.0 if two shapes are identical, and 0.0 if completely different. Two shapes are never completely different, i.e., have zero overlap, but shapes may be identical for different molecular.

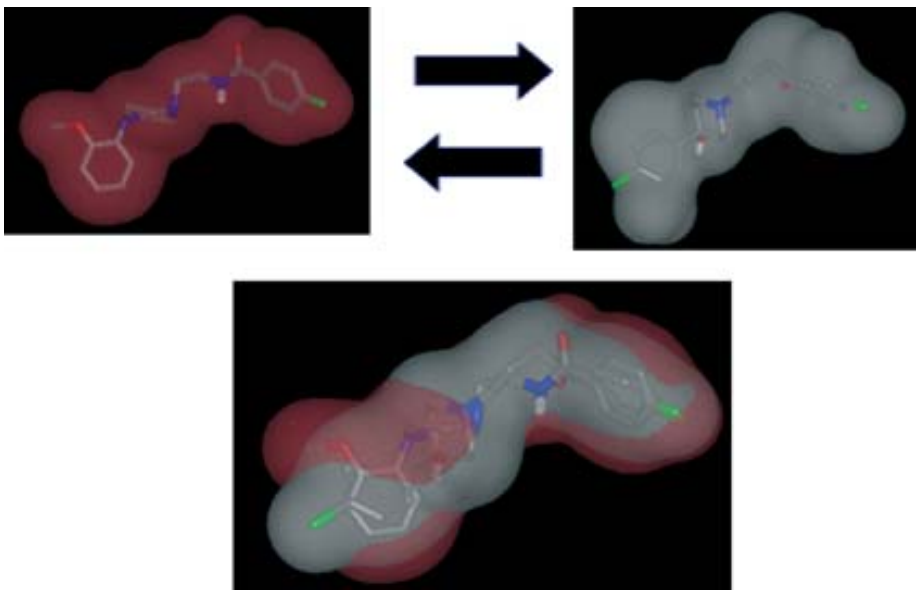


Figure 5.7. Schematic representation of the superposition of two molecules performed by ROCS. The shape Tanimoto between the two molecules represented in the top

panels is 0.78, and the shape overlap is portrayed by the exposed colour of one structure against that of the other. This figure was extracted from reference [595].

Besides basing comparisons between molecules exclusively on shape overlap, ROCS uses a *colour score* to take the chemical properties of overlapping groups, such as H-bond donors, H-bond acceptors, etc., into consideration. Thus, we made use of the so-called *Combo Score*, the sum of two measures of similarity: the shape Tanimoto and the *Scaled Colour* score, the former being a measure of the shape similarity and the latter a measure of the chemical match between two molecules.

Sixteen VS protocols were assembled making use of both the Combo Score and the Scaled Colour score to carry out 3D similarity searches against all X-ray template ligands and the composite ligands (concatamers).

2.2.2. Electrostatic similarity searching with EON

Electrostatics seems to play a particularly important role in the binding of ligands to TTR [384]. Analyses of the electrostatic potential have shown that the surface potential is more positive in the two binding sites than in other regions of the protein surface, due to the influence of the ammonium groups of two Lys-15 residues [235,596]. In addition, the median of the sum of formal charges for the 22 active ligands selected for the test set is minus one (-1), disclosing a tendency to bind negatively charged molecules. This is also in agreement with TTR's well-known propensity to bind halogen-bearing (electronegative) molecules. It is therefore tempting to search for small molecules holding electrostatic properties similar to the template ligands.

The program EON takes the molecular overlays produced by ROCS as input and a field-based measure of Tanimoto to compare the electrostatic potential of a target (database) molecule with a template molecule. The electrostatic potential is calculated internally using Zap, OpenEye's Poisson-Boltzman (PB) electrostatics toolkit. The basic equation for a the quantification of the overlap between the electrostatic potential for two molecules A and B is

$$Tanimoto_{A,B} = \frac{\int A(\vec{r}) * B(\vec{r})}{\int A(\vec{r}) * A(\vec{r}) + \int B(\vec{r}) * B(\vec{r}) - \int A(\vec{r}) * B(\vec{r})}$$

Equation 5.10

In EON, two different Electrostatic Tanimoto (ET) measures are computed, based on the outer dielectric used in the PB calculation: the *ET_pb* (an abbreviation of Poisson-Boltzmann's Electrostatic Tanimoto) uses an outer dielectric of 80, while *ET_coul* (an

abbreviation of Coulomb's Electrostatic Tanimoto) uses a value of 2.0. The rational for using a PB electrostatic field is that the external potential is dampened by orientation of the aqueous solvent. It is a common observation that proteins essentially act to reproduce the aqueous desolvation of well-bound ligands. As a result a PB electrostatic field is more likely to correctly capture the essential elements of binding than that from the Coulombic field. However, this would still seem to be a point to be proven. For the ranking of hit lists, EON outputs both electrostatic Tanimoto measures (which typically track each other closely) and an electrostatic Tanimoto combo (*ET_combo*), which is a sum of the shape Tanimoto (ST) and the Poisson-Boltzmann Electrostatic Tanimoto (ET_pb).

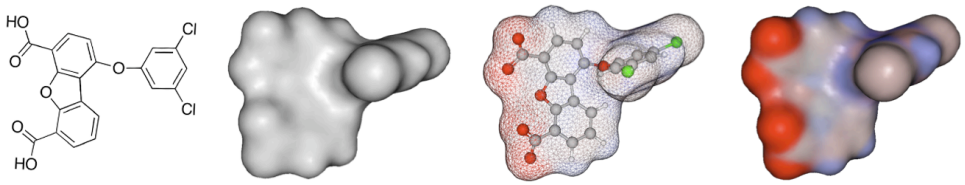


Figure 5.8. Multiple representations of a dibenzofuran (DBF47), a strong TTR amyloid inhibitor, as processed by ROCS and EON. From left to right, the chemical formulae, the shape, the chemical properties and the electrostatic properties are illustrated. The figure was prepared using OpenEye's VIDA [597].

In this chapter, we assembled eight VS protocols that make use of ROCS' alignments and EON's ET_combo score to perform 3D shape electrostatic similarity searches within the benchmarking set against all X-ray templates and concatamers.

2.3. 3D geometric matching upon 2D selection from multiple templates: LigMatch

Given the considerable structural diversity of the known TTR binders, it seems clear that no single template can entirely represent a global pharmacophore for binding to TTR. In this project, we explored a new method developed in our group at the University of Leeds – LigMatch – that combines 2D pre-selection from multiple templates with 3D geometric hashing [598].

In the first step, LigMatch uses a 2D fingerprint based method to calculate Tanimoto coefficients between each molecule in a database to be screened and each of the available template ligands. Then, following the MAX fusion rule explained in Chapter 4 (page 146), the 3D ligand template with the highest Tanimoto coefficient is selected as the template against which all the conformers of that molecule are aligned to and scored using a modified version of the geometric hashing algorithm GH8. Finally, the compounds are ranked

according to a normalized mean 3D similarity score. If only one template is available, the 2D selection step is skipped (single template mode, or STM for short).

In this chapter, the atom–atom score was determined for every conformer of every molecule in the benchmarking set, and each molecule was assigned both a best (conformer) and a mean atom–atom score (based on all conformers). To deal with the maximum possible number of matching atoms, these were then normalized by dividing by the number of atoms present in that given molecule. The resulting lists were then ranked by both the descending normalized best score and the descending normalized mean score. To allow the direct comparison of the performance of LigMatch with the other similarity search methods introduced above, we used the same X-ray templates to perform the searches on the benchmarking set, resulting in a total of twelve VS protocols.

2.4. Docking and Scoring

The high number of deposited crystal structures of TTR complexes encourages the exploration of structure-based virtual screening methods. As shown in Chapter 2, molecular docking can be used to generate sensible geometries of TTR-ligand complexes. AutoDock 4, Vina, GOLD and FRED were tested and all docking programs were able to generate near-native binding modes. However, it was also shown that, when sorted by score, the position of the pose closest to the experimental one is often arbitrarily distributed within the pool of generated poses. Moreover, the inability to predict experimental binding affinities was clearly demonstrated.

A reliable scoring function should be able to achieve three tasks: (*i*) correctly identify the pose closest to the native conformation amongst a pool of docking-generated poses for a given compound (docking power); (*ii*) provide a linear correlation between predicted scores and binding affinities on a random set of protein-ligand complexes (scoring power); and (*iii*) rank a set of different ligands binding to the same target according to their binding affinities (ranking power). In an ideal world, a scoring function that could reach perfection in terms of scoring power would also be perfect in terms of docking power and ranking power. However, in reality, scoring functions are designed to be fast and used in the context of high-throughput virtual screening. Thus, the simplifications of the underlying biophysics render scoring functions in current use far from perfect at binding affinity predictions. Aspects like desolvation and entropic effects (torsional, translational and rotational degrees of freedom) and protein flexibility are not meticulously considered.

Mainly over the last two decades, some computational methods have emerged as being able of providing accurate binding affinity predictions. These include free energy perturbations [599,600], thermodynamic integration [601] and MM-PBSA calculations [602]. Nevertheless, the amount of computational power required to achieve an adequate conformational sampling for the systems under study hinder their utilisation in high-throughput virtual screening campaigns. However, absolute affinities (which can be ascertained experimentally for the most promising hits) are not a requirement, so much as the relative ranking of a data set of ligands with respect to their affinity to one particular target. In other words, scoring power has no direct relevance for rescoring [603]. Yet, and in principle, a good ranking power cannot be achieved without a high docking power.

Alongside ligand-centric methods, in this chapter eighteen docking and scoring VS protocols were assembled and evaluated for their ability to discriminate TTR actives. Based on the results obtained in Chapter 2, AD4 and Vina were elected for the docking procedure and, besides their built-in scoring functions, two additional knowledge-based scoring functions, DrugScore [93,604] and DSX [603], were explored. For every ligand, the lowest-energy pose was re-scored with both scoring functions. In DrugScore, only the distance-dependent pair-preferences derived from the Cambridge Structural Database (CSD) were used (DrugScore^{CSD}). The pair potentials in DSX inherit the statistical philosophy of DrugScore. However, the latest scoring function features a more detailed atom type assignment and avoids problems associated with the *reference state* of no interaction in DrugScore. In the next chapter, we provide a more detailed and comprehensive description of DrugScore and DSX, to justify their choice over other freely available stand-alone scoring functions.

3. Results

In the following subsections we report the results of performance evaluation obtained for each of the VS protocols setup to identify TTR amyloid inhibitors. The results are grouped and presented in increasing order of complexity of the underlying VS methods.

3.1. 2D similarity searches

Figure 5.9 shows the enrichment curves for the 2D fingerprint-based VS protocols employing different similarity metrics. Throughout this chapter, the x-axes are shown in logarithmic scale in order to facilitate the analysis of early behaviour. Additionally, ROC

enrichment (ROCE) values at 0.5, 1, 2 and 5% of the ranked benchmarking set are given in **Table 5.2**.

Overall, the enrichment achieved using chemical hashed fingerprints (CHF) as molecular descriptors and the Euclidean distance (ED) as the measure of similarity between the benchmarking set molecules and the template ligands is the best of all VS methods explored for TTR. The use of the Tanimoto coefficient with CHF yields poor global enrichments, as disclosed by the low area under the enrichment curve (AUC) for most templates. However, the results are highly template-dependent in terms of early enrichment. In fact, when using CHF with the Tanimoto metric, the PCB18 template (compound **22** in **Figure 4.2**) is an outstanding exception, attaining the best results of the whole benchmark when compared to all other X-ray derived templates.

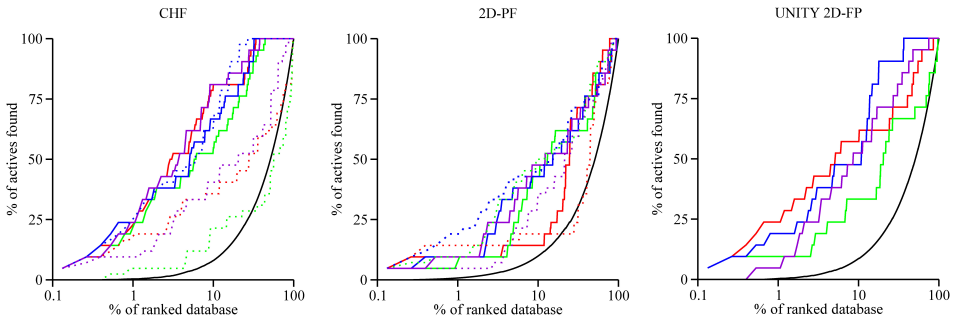


Figure 5.9. Enrichment curves to compare the VS performance of three different 2D fingerprint-based methods on a benchmarking set created for TTR: ChemAxon Chemical Hashed Fingerprints (CHF) and 2D Pharmacophore Fingerprints (2D-PF), and Tripos UNITY 2D Fingerprints. For CHF and 2D-PF, the continuous lines represent the use of the Euclidean distance as a measure of similarity, whereas the dotted lines represent the use of a Tanimoto coefficient. For UNITY 2D-FP only the Tanimoto coefficient was used. The line colours correspond to different templates as follows; phenox (1dvy) in red, benzoxazole acid (2f8i) in green, PCB18 (2g5u) in blue, and T_4 (2rox) in violet. Random performance is depicted by the black line.

In contrast to CHF, the 2D Pharmacophore Fingerprints (2D-PF) seem to provide better early enrichment when using the Tanimoto metric as opposed to the ED metric, although similar AUC values are obtained for the two metrics. UNITY 2D Fingerprints provide the best results using the Tanimoto coefficient, yielding high early enrichment with the phenox template (compound **4**) and high overall enrichment with the PCB18 template. As verified for the CHF descriptors, the PCB18 template seems to hold the highest level of similarity to TTR binders within the 2D descriptor space. This may relate to this compound's high affinity for TTR [219,378].

Table 5.2. ROC Enrichment (ROCE) values at 0.5, 1, 2 and 5% false positive rates (FPR) and AUC values of the enrichment curves for the VS protocols based on 2D similarity tested on TTR. Grey shading identifies the best performing protocols.

Method / Similarity measure / Template			*ROCE 0.5%*[a]	*ROCE 1%*	*ROCE 2%*	*ROCE 5%*	*AUC*
CHF	Euclidean distance	Phenox (1dvy)	**142.0**	**59.5**	**42.2**	**16.3**	**0.92**
		Benzoxazole (2f8i)	142.0	59.5	42.2	11.0	0.88
		PCB18 (2g5u)	**Max.**	**59.5**	**42.2**	**12.9**	**0.90**
		Thyroxine (2rox)	**142.0**	**59.5**	**42.2**	**18.8**	**0.92**
	Tanimoto	Phenox (1dvy)	94.8	38.0	13.6	7.9	0.61
		Benzoxazole (2f8i)	5.9	5.2	2.6	2.5	0.44
		PCB18 (2g5u)	**Max.**	**101.0**	**42.2**	**12.9**	**0.93**
		Thyroxine (2rox)	47.5	11.9	11.2	7.9	0.70
2D-FP	Euclidean distance	Phenox (1dvy)	47.5	11.9	5.6	3.0	0.72
		Benzoxazole (2f8i)	11.8	11.9	5.6	5.4	0.72
		PCB18 (2g5u)	47.5	11.9	5.6	9.5	0.73
		Thyroxine (2rox)	47.5	11.9	8.9	6.6	0.73
	Tanimoto	Phenox (1dvy)	142.0	23.7	8.9	3.6	0.58
		Benzoxazole (2f8i)	47.5	11.9	9.6	9.5	0.76
		PCB18 (2g5u)	**Max.**	**38.0**	**25.8**	**11.0**	**0.76**
		Thyroxine (2rox)	11.8	5.2	2.6	4.1	0.70
UNITY 2D-FP	Tanimoto	Phenox (1dvy)	**Max.**	**59.5**	**33.3**	**12.5**	**0.80**
		Benzoxazole (2f8i)	47.5	11.9	5.6	5.3	0.66
		PCB18 (2g5u)	**142.0**	**38.0**	**18.3**	**12.5**	**0.90**
		Thyroxine (2rox)	11.8	5.2	12.7	9.5	0.83

[a] This percentage of the ranked benchmarking set of 760 ligands corresponds to only 4 ligands. When these top-scoring ligands are all actives, the ROC enrichment value becomes the result of a division by zero. Accordingly, the result is presented as "Max."

3.2. 3D similarity searches and LigMatch

The enrichment curves for all single-template 3D similarity-based VS protocols are given in **Figure 5.10**. For comparison purposes, we tested LigMatch's single template mode against each of the available templates.

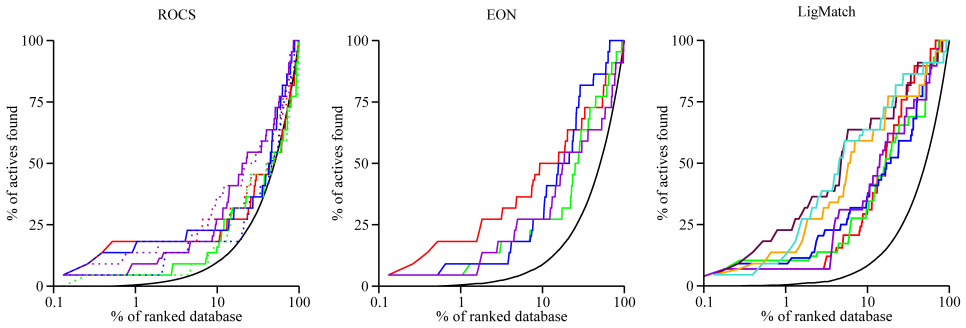


Figure 5.10. Enrichment curves to compare the performance of different single-template 3D similarity-based VS protocols (ROCS, EON and LigMatch's single template mode) on the benchmarking set created for TTR. The performance of LigMatch's multiple template mode is also illustrated. For ROCS (leftmost panel), the continuous lines represent the use of the Combo Score as a similarity score, whereas the dotted lines represent the use of the Scaled Colour score. In LigMatch, the line colours correspond to different tem-

plates; for the single template methods: phenox (1dvy) in red, benzoxazole acid (2f8i) in green, PCB18 (2g5u) in blue, and T_4 (2rox) in violet; and for the multiple-template LigMatch schemes: phenox + benzoxazole acid + PCB18 + T_4 in maroon, phenox + benzoxazole acid + PCB18 in orange, and phenox + PCB18 + T_4 in turquoise.

In its multiple template mode, LigMatch employs a preliminary 2D selection from the available templates prior to the 3D alignment and similarity comparison of the molecules with the templates. We present its results along with those obtained from ROCS and EON to illustrate the benefit of using multiple templates in a single VS run. ROCE values at 0.5, 1, 2 and 5% of the ranked benchmarking set and the area under the enrichment curves are given in **Table 5.3**.

Table 5.3. ROCE values at 0.5, 1, 2 and 5% FPR and AUC values of the enrichment plots for the 3D similarity methods and LigMatch. Grey shading identifies the best performing VS protocols.

Method / Similarity measure / Template			*ROCE 0.5%*[a]	*ROCE 1%*	*ROCE 2%*	*ROCE 5%*	*AUC*
ROCS	Combo-Score	Phenox (1dvy)	**Max.**	**36.2**	**12.9**	**4.0**	**0.54**
		Benzoxazole (2f8i)	11.3	5.0	2.5	1.9	0.51
		PCB18 (2g5u)	**136.0**	**36.2**	**12.9**	**5.2**	**0.61**
		Thyroxine (2rox)	11.3	11.3	6.9	4.0	0.65
	ScaledColor	Phenox (1dvy)	**Max.**	**36.2**	**12.9**	**4.0**	**0.58**
		Benzoxazole (2f8i)	11.3	5.0	2.5	1.9	0.51
		PCB18 (2g5u)	11.3	5.0	2.5	4.0	0.56
		Thyroxine (2rox)	**45.0**	**22.7**	**12.9**	**5.2**	**0.66**
EON	ET_combo	Phenox (1dvy)	**Max.**	**36.2**	**22.7**	**9.1**	**0.73**
		Benzoxazole (2f8i)	11.3	5.0	8.5	4.0	0.68
		PCB18 (2g5u)	11.3	11.3	5.3	4.0	0.78
		Thyroxine (2rox)	11.3	5.0	8.5	6.3	0.65
Lig-Match	Single template mode	Phenox (1dvy)	22.7	8.5	3.8	3.8	0.79
		Benzoxazole (2f8i)	51.5	14.7	7.3	3.7	0.74
		PCB18 (2g5u)	45.0	11.3	6.9	5.7	0.74
		Thyroxine (2rox)	22.7	8.5	3.8	7.4	0.76
	Multiple template mode	Phenox (1dvy) + Benzoxazole (2f8i) + PCB18 (2g5u) + Thyroxine (2rox)	**Max.**	**56.8**	**36.1**	**15.6**	**0.84**
		Phenox (1dvy) + Benzoxazole (2f8i) + PCB18 (2g5u)	45.0	22.7	22.7	9.1	0.82
		Phenox (1dvy) + PCB18 (2g5u) + Thyroxine (2rox)	45.0	23.7	22.7	14.7	0.83

[a] This percentage of the ranked benchmarking set of 760 ligands corresponds to only 4 ligands. When these top-score ligands are all actives, the ROCE value becomes the result of a division by zero. Accordingly, the result is presented as "Max."

3.2.1. ROCS

The best overall enrichments for ROCS are obtained with the T_4 template when using the chemical complementarity descriptor alone (as implemented in the Scaled Colour scoring function), with an AUC of 0.66, and when using the combination of the molecular shape

and the chemical complementarity descriptors (as implemented in the Combo Score function), with an AUC of 0.65. Interestingly, these results seem to agree with the binding characteristics of T_4 to TTR. T_4 (compound **1**) presents a more complete set of pharmacophoric features to interact with TTR binding sites (including a hydroxyl group at the inner aromatic ring and a positively charged amino group at the outer tail), but the presence of the four iodine atoms renders it too bulky and may hinder its diffusion into TTR binding sites (thus, decreasing its affinity). On the other hand, the best early enrichments are attained with the phenox and the PCB18 templates when applying the Combo Score, yet sharply decreasing beyond the 1% threshold and detecting no more than 25% of the actives at 10% of the ranked database. Holding less bulky halogen atoms and fewer degrees of rotational freedom in their structure, these templates present the shape and the chemical features that better represent the strongest TTR binders, therefore capturing them at the top of the ranked lists. The results obtained with ROCS reveal poor overall enrichments compared to all other methods tested in this study, which suggests that molecular shape and chemical complementarity descriptors alone may be insufficient to capture the structural and chemical diversity of all TTR actives.

3.2.2. EON

The inclusion of a descriptor of the electrostatic properties of TTR binders was expected to be beneficial, given TTR's propensity to bind neutral to negatively charged molecules or molecules bearing electronegative elements (halogens). EON takes molecular alignments generated by ROCS and computes a Poisson-Boltzmann Electrostatic Tanimoto (ET) score between each database molecule and the query template. In **Figure 5.10** and **Table 5.3** we report the results obtained with EON using the ET combo score, which is a sum of the ET and a shape Tanimoto.

Analysing the differences between EON and ROCS we observe a slight improvement in terms of over-all enrichment, with the PCB18 template showing the highest increase in AUC. In terms of early enrichment, the only considerable improvement occurs for the phenox template, specifically at 2 and 5% of the ranked database (ROCE of 22.7 and 9.1, respectively). It is worth noting that 50% percent of the actives (11 hits) are retrieved within 10% of the top-scored compounds. Indeed, this is the best result obtained with a 3D similarity search method using a single template derived from an X-ray complex. A reasonable decline in early enrichment is observed for the PCB18 template. This is probably linked to this compound's peculiar electronegative potential, which is due not only to the presence of four halogen atoms but also to their electron-withdrawing inductive effect over the two hydroxyl groups, which may even render them deprotonated at neutral *p*H [605]. This distinctive feature seems to separate this template from actives with

similar shape, namely compounds **6** and **19** (recall **Figure 4.2**), causing the loss of early enrichment. However, potent inhibitors from cluster B (compounds **8**, **9**, **10** and **12**) and compounds **3**, **4**, **5b**, **5c**, and **23**, all receiving very weak scores by ROCS with the PCB18 template, are placed at much higher levels of the ranked list with EON, explaining the increase in overall enrichment.

3.2.3. LigMatch: single template mode

The average AUC for the four single templates tested with LigMatch's single template mode is 0.76, a result that represents an improvement with respect to EON (average AUC of 0.71) and ROCS (average AUC of 0.58), but not with respect to those obtained with the 2D CHF employing the Euclidean distance (average AUC of 0.91). In terms of early enrichment, while other ligand-based protocols perform very well with at least one of the templates, LigMatch's single template mode offers a more modest/equilibrated performance. Interestingly, it performs better than any other 3D search methods when using the benzoxazole (2f8i) template.

3.2.4. LigMatch: 2D selection from multiple templates

Given the considerable structural diversity amongst the TTR actives in the benchmarking set, and the fact that these compounds explore different sub-regions of TTR pockets, we explored LigMatch's multiple template mode to assess the influence of adding extra layers of information in a single 3D search VS run.

Considerable improvements in terms of both early and overall enrichments are achieved when using the 2D selection from multiple templates and employing the strongest inhibitors of each of the three clusters as templates (PDB codes 1dvy, 2f8i and 2g5u). However, further increase is attained by inclusion of T_4 (2rox) as a fourth template (see **Figure 5.10** and **Table 5.3**).

This observation supports the notion that, although T_4 is neither a high affinity binder nor a potent amyloid inhibitor, the endogenous TTR binder may hold specific features that are critical to the interaction with TTR and/or to better represent the chemical space of known TTR binders than other features presented by stronger inhibitors (whose binding may be merely more favourable entropically due to higher hidrophobicity). To test this idea, we replaced the strong binder benzoxazole acid (compound **10**), which lacks some key substituent groups, by T_4 (compound **1**), in a third multiple-template scheme. A slight increase in both early and overall enrichment was obtained (see **Figure 5.10**),

though much more modest than that obtained with the four-template scheme (described above). In all cases, and as pointed out in the LigMatch paper [598], ranking by the normalized mean score provides better results than ranking by the normalized best score.

3.3. 2D and 3D similarity searches using concatamers

Another possible strategy to address the problem of low "representativeness" of single templates and improve the performance of similarity searches would be to designate certain features as being essential, for example, the negatively charged carboxylates interacting with Lys-15 residues at the entrance of TTR pockets, or the ionizable hydroxyl interacting with Ser-117 residues at the opposite end of the pockets. Conceptually, this would be equivalent to a 3D pharmacophore. Therefore we modelled four composite templates (concatamers), by adding a few features believed to be essential to strong reference templates, and tested them using all similarity search methods available. The results are shown in **Figure 5.11** and **Table 5.4**.

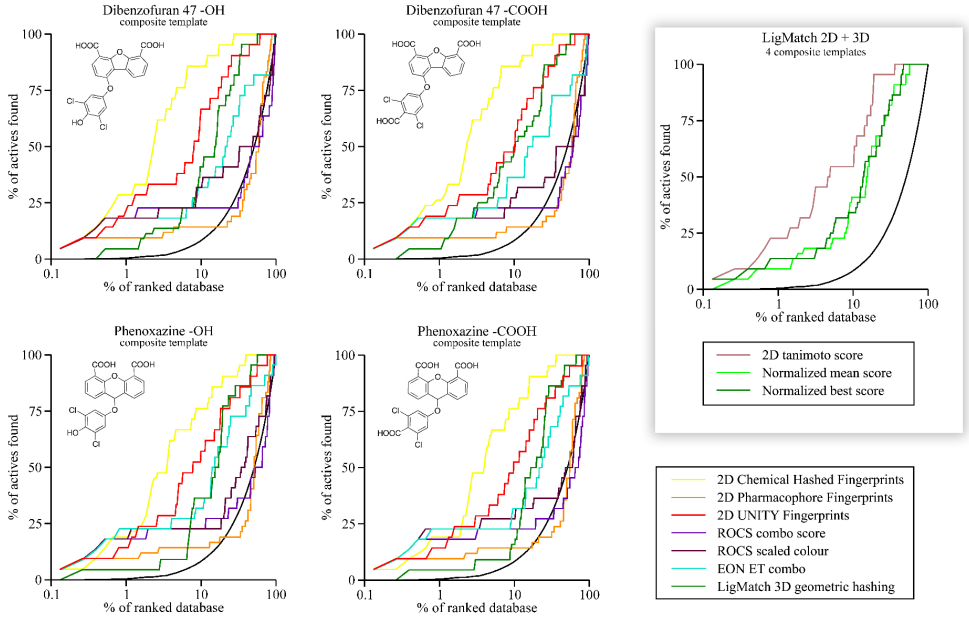


Figure 5.11. Enrichment curves to compare the performance of different VS protocols using four different composite ligands as single templates (panels in the left and centre). The use of a 2D selection from the four composite ligands as implemented in LigMatch's multiple template mode (panel in the right), is also compared with the single template mode using a normalized best score (in dark green), a normalized mean score (in green), and a 2D Tanimoto score (in brown) for ranking.

In spite of their very close similarity, the two families of concatamers provide slightly different results in terms of early and overall enrichment. The templates derived from the DBF47 (inhibitor **5c**) yield the highest enrichment profiles of the whole benchmark

with the CHF method (and the Euclidean distance metric). Indeed, ROCE values of 142.5, 42.2 and 26.2 respectively at 1, 2 and 5% of the ranked benchmarking set, along with an AUC of 0.95, are obtained with the DBF47-OH template (compound **24**), which are quite impressive results. Conversely, significant improvement to 3D similarity searches is only obtained with the phenox-derived templates. In particular, the phenox-OH template (compound **26**) positively influences the results of EON, yielding the highest early enrichment at 1% achieved with a 3D similarity search protocol and an acceptable overall enrichment. These results are only comparable with those obtained with the PCB18 and the phenox templates.

With an average AUC of 0.82 across the four concatamers, the geometric hashing procedure of LigMatch provides the best overall enrichment of all the 3D methods. However, poor early enrichments are generally obtained using single templates, contrasting with the results achieved using the 2D selection procedure from the more diverse templates (as also seen in the previous subsection). To study the influence of these two factors (i.e. the 2D selection step and the diversity amongst the template molecules) we re-ranked the benchmark set using the multiple template protocol with 2D selection from the four concatamers. Although considerable improvement in early enrichment is observed, the values are still far below those obtained using the four templates 1dvy, 2f8i, 2g5u and 2rox (see last row in **Table 5.3**). These results confirm that, unlike the tested single-template similarity search protocols, LigMatch (and particularly its single template mode) does not benefit from the use of the composite templates.

Table 5.4. ROCE values at 0.5, 1, 2 and 5% FPR and AUC values of the enrichment curves for four composite templates using different methods on TTR. Grey shading identifies the best performing protocols.

Original template / Key R-group added / Method			*ROCE 0.5%[a]*	*ROCE 1%*	*ROCE 2%*	*ROCE 5%*	*AUC*
DBF47	OH	2D CHF Euclidean	**Max.**	**142.5**	**42.2**	**26.2**	**0.95**
		2D PF Tanimoto	47.5	11.9	5.6	3.0	0.50
		2D UNITY Tanimoto	142.0	59.5	33.3	9.5	0.87
		ROCS Combo-Score	Max.	36.2	17.5	5.2	0.47
		ROCS Scaled-Color	Max.	36.2	12.9	5.2	0.57
		EON ET_combo	Max.	36.2	12.9	4.0	0.69
		LigMatch STM	11.3	5.0	6.9	2.9	0.83
	COOH	2D CHF Euclidean	**Max.**	**101.0**	**42.2**	**23.0**	**0.95**
		2D PF Tanimoto	47.5	11.9	5.6	2.0	0.49
		2D UNITY Tanimoto	142.0	38.0	23.8	9.5	0.86
		ROCS Combo-Score	Max.	36.2	12.9	5.2	0.45
		ROCS Scaled-Color	Max.	36.2	12.9	4.0	0.53
		EON ET_combo	Max.	36.2	22.7	9.1	0.69
		LigMatch STM	11.3	5.0	12.9	6.3	0.84

Original template / Key R-group added / Method			*ROCE 0.5%*[a]	*ROCE 1%*	*ROCE 2%*	*ROCE 5%*	*AUC*
Phenox	OH	2D CHF Euclidean	142.0	38.0	33.3	20.8	0.92
		2D PF Tanimoto	47.5	11.9	7.2	3.0	0.52
		2D UNITY Tanimoto	47.5	23.7	18.3	11.0	0.84
		ROCS Combo-Score	Max.	36.2	17.5	5.2	0.51
		ROCS Scaled-Color	Max.	36.2	17.5	5.2	0.60
		EON ET_combo	**Max.**	**113.5**	**17.5**	**6.3**	**0.75**
		LigMatch STM	11.3	5.0	2.5	1.9	0.82
	COOH	2D CHF Euclidean	94.8	38.0	18.3	20.8	0.93
		2D PF Tanimoto	47.5	11.9	5.6	3.0	0.51
		2D UNITY Tanimoto	47.5	23.7	18.3	7.9	0.83
		ROCS Combo-Score	Max.	36.2	12.9	5.2	0.45
		ROCS Scaled-Color	Max.	56.8	17.5	6.3	0.56
		EON ET_combo	**Max.**	**56.8**	**17.5**	**5.2**	**0.75**
		LigMatch STM	11.3	5.0	2.5	1.9	0.79
All 4 *concatamers*		LigMatch Multiple Template Mode	45.0	11.3	10.7	3.9	0.82

[a] This percentage of the ranked benchmarking set of 760 ligands corresponds to only 4 ligands. When these top-score ligands are all actives, the ROCE value becomes the result of a division by zero. Accordingly, the result is presented as "Max."

3.4. Docking and Scoring

Two best-performing X-ray structures of TTR identified through the crossdocking studies – 1bm7 and 2g5u – were selected to devise docking and scoring VS protocols based on AutoDock 4 (AD4) and Autodock Vina, with re-scoring of the predicted complexes with DrugScore^{CSD} [93,604] and DSX [603]. A third TTR structure was manually modelled on top of the 1bm7 structure, to reproduce the optimal positioning of one of the three hydroxyl rotamers of Ser-117 residues found in 2g5u (and absent in the original 1bm7). This particular side-chain conformation facilitates interaction with ligands holding polar features that can hydrogen-bond either with the protein or with a water molecule arrested at the inner end of TTR's binding tunnel. This is a relevant aspect from the viewpoint of ligand-binding thermodynamics, but can also have a great impact in the accuracy and performance of docking programs. We call this structure *1bm7opt*.

Figure 5.12 shows the enrichment curves for both TTR binding sites found in 1bm7, 2g5u and *1bm7opt*, and with each docking and scoring protocol under evaluation. Since in Chapter 6 we compare the performance of several docking-based VS protocols against multiple DUD targets using the ROC curve approach, the corresponding ROC curves for TTR's benchmarking set are provided in Appendix (**Figure D.1**).

The visual inspection of the curves indicates poor enrichments, in general, only marginally better than random selection. This is confirmed by a mean AUC of 0.53 (± 0.17) across all structures and protocols. Strikingly, even though no palpable structural differences were detected in the comparison of the two TTR binding sites, it is clear that most docking and scoring protocols were able to capture minor structural discrepancies and translate them into significant decreases in VS performance. In fact, the median AUC obtained on TTR "AC" binding site was 0.63 against 0.46 obtained on the "BD" binding site. The Vina Scoring Function and the Vina + DrugScore^{CSD} protocols offered the most consistent results on both binding sites across the three TTR structures.

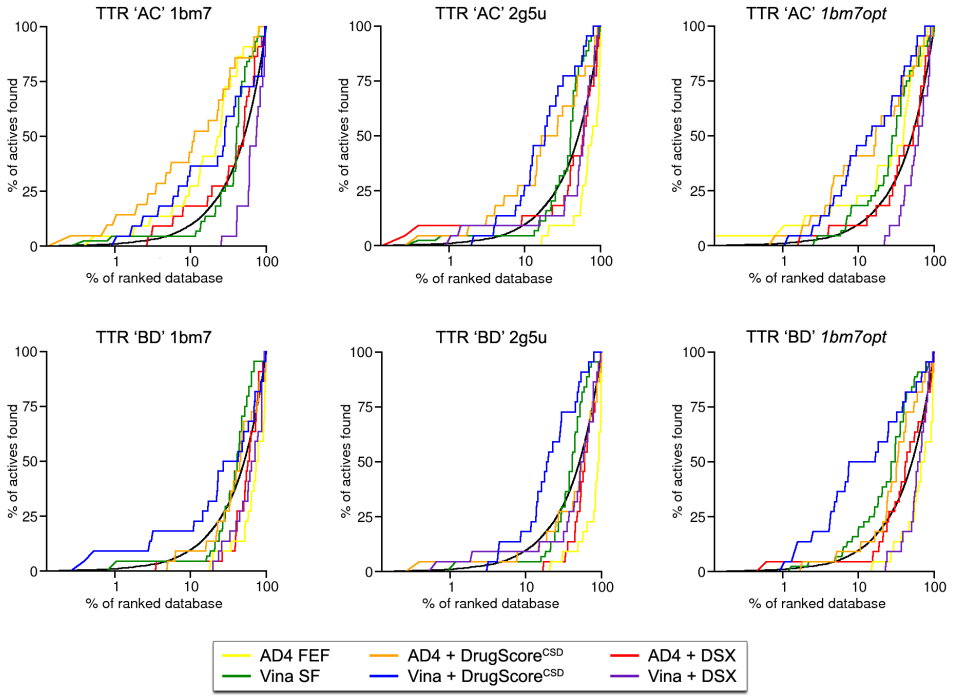


Figure 5.12. Enrichment curves comparing the VS performance of six docking and scoring protocols against TTR. The results obtained with the 1bm7 X-ray structure are shown in the left panels, the middle panels show the results with structure 2g5u, and the right panels correspond to the results of the *1bm7opt* model. In all cases, the results obtained for the "AC" binding site of the protein are in the top panels and the bottom panels show the results for the "BD" binding site. The protocols are coloured as follows: AD4 Free Energy Function (yellow), AD4 + DrugScore^{CSD} (orange), AD4 + DSX (red), Vina Scoring Function (green), Vina + DrugScore^{CSD} (blue), Vina + DSX (purple), and random performance (black).

The influence of TTR's Ser-117 residues on the affinity grids of grid-based methods (such as AutoDock 4 and Vina) was previously evidenced in MIF analyses (Chapter 3). Because no visible gains in enrichment were drawn from the use of the modelled *1bm7opt* structure (see **Figure 5.12** and **Table 5.5**), this experiment raised new doubts about the role of Ser-117 residues in the binding and interaction of the known actives to TTR.

Failures during the docking procedure are quite recurring, often motivated by inconsistencies in the input files or just inability to handle specific atomic elements, tautomeric forms, etcetera. These failures decrease the amount of information in the benchmarking set, affecting the profiles of enrichment curves. At this point, it is worth mentioning that both AD4 and Vina flawlessly handled the entire TTR benchmarking set of ligands. Moreover, all docking poses generated by AD4 and Vina were successfully re-scored with both DrugScore^{CSD} and DSX.

Table 5.5. ROCE values at 0.5, 1, 2, and 5 % and AUC of the enrichment curves for AutoDock 4 and AutoDock Vina against TTR using built-in scoring functions. The best combinations of AUC and ROCE values are highlighted in bold and grey shades.

		AUC	Early enrichment (ROCE)				AUC	Early enrichment (ROCE)			
			0.5%	1%	2%	5%		0.5%	1%	2%	5%
TTR structural template	TTR binding site	AD4 FEF					Vina SF				
1bm7	"AC"	**0.74**	**11.3**	**5.0**	**5.3**	**2.9**	**0.63**	**5.6**	**5.0**	**2.5**	**0.9**
	"BD"	0.28	0.0	0.0	0.0	0.0	0.58	0.0	5.0	2.5	0.9
2g5u	"AC"	0.26	0.0	0.0	0.0	0.0	0.58	5.6	5.0	2.5	0.9
	"BD"	0.19	0.0	0.0	0.0	0.0	0.60	0.0	2.5	2.5	0.9
1bm7opt	"AC"	0.65	11.3	11.3	8.5	3.9	0.65	0.0	0.0	0.9	1.9
	"BD"	0.32	0.0	0.0	0.0	0.0	0.68	0.0	0.0	1.3	1.4

Looking at the early enrichment profiles across **Table 5.5**, **Table 5.6** and **Table 5.7**, it is clear that the best performance was obtained with the AD4 + DrugScore^{CSD} protocol using the "AC" binding site of TTR structure 1bm7: ROCE of 11.8, 23.7, 13.6 and 7.9 at 0.5, 1, 2 and 5% of ranked data set. Along with an AUC of 0.78, these results confirm that the coupling of the AD4 + DrugScore^{CSD} protocol with the 1bm7 X-ray template provided the best overall VS performance for TTR. Nevertheless, and as evidenced by the plots in **Figure 5.12**, other docking and scoring protocols offered respectable performance, such as the Vina + DrugScore^{CSD} protocol with the "BD" binding site of the *1bm7opt* template.

Table 5.6. ROCE values at 0.5, 1, 2, and 5 % and AUC of the enrichment curves for AutoDock 4 and AutoDock Vina against TTR upon re-scoring with DrugScore^{CSD}. The best combinations of AUC and ROCE values are highlighted in bold and grey shades.

		AUC	Early enrichment (ROCE)				AUC	Early enrichment (ROCE)			
			0.5%	1%	2%	5%		0.5%	1%	2%	5%
TTR structural template	TTR binding site	AD4 + DrugScore^{CSD}					Vina + DrugScore^{CSD}				
1bm7	"AC"	**0.78**	**11.8**	**23.7**	**13.6**	**7.9**	0.63	0.0	5.0	5.3	3.9
	"BD"	0.51	0.0	0.0	0.0	0.9	0.58	45.0	11.3	5.3	3.9
2g5u	"AC"	0.67	11.3	5.0	5.3	3.9	0.75	0.0	0.0	0.0	2.9
	"BD"	0.46	11.3	5.0	2.5	0.9	0.73	0.0	0.0	0.0	2.9
1bm7opt	"AC"	0.73	0.0	5.0	5.3	7.8	0.73	0.0	0.0	2.5	5.2
	"BD"	0.31	0.0	0.0	2.5	0.9	**0.75**	**0.0**	**5.0**	**8.5**	**7.6**

Table 5.7. ROCE values at 0.5, 1, 2, and 5 % and AUC of the enrichment curves for AutoDock 4 and AutoDock Vina against TTR upon re-scoring with DSX. The best combinations of AUC and ROCE values are highlighted in bold and grey shades.

		AUC	Early enrichment (ROCE)				AUC	Early enrichment (ROCE)			
			0.5%	1%	2%	5%		0.5%	1%	2%	5%
TTR structural template	TTR binding site	AutoDock 4 + DSX					Vina + DSX				
1bm7	"AC"	0.55	0.0	0.0	0.0	1.9	0.30	0.0	0.0	0.0	0.0
	"BD"	0.42	0.0	0.0	0.0	0.9	0.36	0.0	0.0	0.0	0.0
2g5u	"AC"	0.45	0.0	0.0	0.0	0.0	0.44	0.0	5.0	5.3	1.9
	"BD"	0.38	0.0	0.0	0.0	0.0	**0.45**	**0.0**	**5.0**	**5.3**	**1.9**
1bm7opt	"AC"	0.51	0.0	0.0	2.5	1.9	0.39	0.0	0.0	0.0	0.0
	"BD"	**0.50**	**0.0**	**5.0**	**2.5**	**0.9**	0.38	0.0	0.0	0.0	0.0

Molecular weight (and more generally molecular size) is known to affect the scores assigned by docking scoring functions. This issue is particularly important in the case of TTR, given its natural propensity to bind iodine-bearing molecules, whose molecular weight may be considerably higher than that of the most common drug-like compounds. To conclude our analysis, we delved deeper into the ranking produced by DrugScore^{CSD} to find the strongest amyloid inhibitor discovered to date, a polychlorinated biphenyl (PCB) [378], ranked at the top 0.5% of the list. This was a compelling result because the molecular weight of this compound (322 Daltons) is amongst the lowest of all active inhibitors in the benchmarking set.

We tested this apparently low bias displayed by DrugScore^{CSD} by plotting the scores assigned to each compound in our diverse decoy set against its respective molecular weight, both for AD4 and DrugScore^{CSD} (**Figure 5.13**). In order to compare the two profiles quantitavely, we computed Pearson correlation coeffiecients for each, providing a normalized measurement of how the two variables (i.e. score and molecular weight) are linearly related. While the free energy function of AD4 shows a strong dependence on the compounds' molecular weight, with a Pearson correlation coefficient of -0.40, DrugScore^{CSD} exhibits a minor (and even positive) dependence, with a correlation coefficient of 0.11. This is another clear improvement offered by the DrugScore^{CSD} scoring function to the quality of the ranking produced by our docking-based VS protocols.

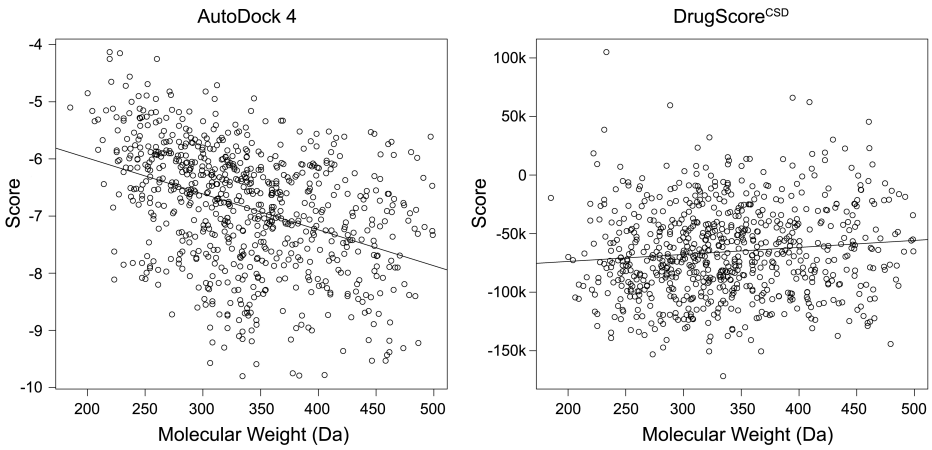


Figure 5.13. Scatter plots to evaluate the influence of molecular weight (MW) on docking scores. The data was based on the docking scores assigned by AutoDock 4 (left panel) and DrugScore^{CSD} (right panel) to the 738 decoys of the TTR benchmarking set. A least squares regression line fitting the data is shown on both plots, disclosing a negative slope for AD4 and a near neutral slope for DrugScore^{CSD}, allowing for a qualitative comparison of the scoring functions.

4. Discussion and concluding remarks

In this chapter we presented a comparative analysis of several virtual screening protocols assembled to identify new inhibitors of amyloid fibril formation by TTR. Using a benchmarking set comprised of the most important TTR stabilisers discovered to date and property-mimetic decoy molecules, we compared the performance of several VS protocols ranging from simple 2D fingerprint-based methods, through more complex 3D similarity searches, all the way to molecular docking and scoring. The performance of the different protocols was assessed by carefully selected metrics of enrichment of known actives at the top-levels of the ranked benchmarking set.

Focusing only on the enrichment profiles obtained, ligand-based protocols performed better than protein-based ones (i.e., docking into TTR structures). 2D similarity search methods, in particular, were able to offer the best discrimination between the known TTR actives and the decoy molecules. Given the reasonable structural diversity found amongst the known actives, these methods provide a good balance between search specificity, a key aspect for target selectivity, and search flexibility, which is necessary for eliminating the dependency on the used templates. While producing worse enrichments than 2D methods, 3D similarity search methods based on shape and chemical complementarity were able to retrieve more diverse scaffolds, with noticeable added value when electrostatic similarity was included as a descriptor. Yet, because these methods use one single bioactive template molecule per VS run they often overlook strong TTR stabilisers with dissimilar structure. The use of concatamers to address the problem of representativeness of single templates brought improvements to the performance of similarity search methods, and even more balanced outcomes were attained using LigMatch, a method that can employ multiple templates and combine 3D geometric hashing with a 2D pre-selection process. It would be very interesting to see whether the application of the same 2D pre-selection protocol to other 3D similarity search methods (e.g. ROCS and EON) would improve matters.

Several docking and scoring VS protocols were assembled and tested using the same metrics of evaluation. Some of these protocols included a step of re-scoring of the complexes predicted by docking, using knowledge-based scoring functions, and the improvements in the ability to correctly rank known ligands over decoys were clear. Nevertheless, their VS performance is still far from ideal and our results highlight limitations when handling hydrophobic pockets such as TTR. We hypothesize that non-polar/non-specific interactions may exhibit cooperative behaviour at "hidden" levels of physics/chemistry that is only linearly captured by the scoring functions (in their "traditionally" additive form), or, most likely, not captured at all. Indeed, several recent studies have emphasized the need for scoring functions that incorporate descriptors that model cooperative effects [606–608]. Still, we have successfully identified at least one VS protocol that combines docking (with AutoDock 4) and re-scoring (with DrugScore^{CSD}) to provide encouraging enrichment profiles for TTR and minimal bias towards the molecular weight of the ligands. This protocol will be explored under the aim of the screening of new TTR amyloid inhibitors. Still, it would be interesting in the future to see whether the use of machine learning techniques to train knowledge-based (statistical) potentials – as others have attempted using Kernel PLS (partial least-squares) [609] and random forest classifiers [610] – would provide more accurate binding affinity predictions and better rankings.

In this chapter, my self-criticism goes to the lack of complementary VS experiments employing alternative benchmarking data sets (of different sizes and characteristics) in order to test the assembled VS protocols more exhaustively and obtain additional statistical evidence. One possibility would be to use the DUD-*all* set, i.e. to seed the known TTR amyloid inhibitors amongst the entire library of actives and decoys deposited in the Directory of Useful Decoys. Given that all molecules in the DUD have been screened through drug-like filters, such VS experiment could (*i*) provide a picture of the behaviour of the methods in a real life context, (*ii*) offer interesting clues on the propensity of the assembled protocols to prioritize known actives against other targets (and thus retrieve hits that are more likely to cause side effects). Once again, time is always against us, but these are ideas to keep in mind and test in the near future.

To wrap up some of the key conclusions on the application of VS methods to TTR, the increase in complexity of the protocols by inclusion (or combination) of additional layers of information – such as molecular shape, electrostatic properties, the use of composite templates or multiple templates, and receptor information in molecular docking – may be accompanied by decreases in performance on a given benchmarking set. However, this increase in complexity seems to be accompanied by an increase of the likelihood of retrieving new and promising scaffolds with distinctive physicochemical properties (at increasing costs of false positives).

While the re-scoring of thousands of predicted protein-ligand conformations can be accomplished in a matter of CPU hours, the docking of a single ligand into a receptor structure itself can be a computationally intensive task. This, of course, depends on the types of algorithms employed by the docking programs for conformational sampling, coupled with the complexity/flexibility of the ligands to be docked. Unfortunately, AutoDock 4 has been shown to perform particularly slowly when compared with other programs [440]. However, the promising results attained with the AD4 + DrugScore^{CSD} protocol, along with the requirement of an enormous amount of computer power to screen over a virtual library of approximately 2.3 million compounds, prompted us to port this protocol onto the Ibercivis volunteer computing platform, as a first implementation phase of the AMILOIDE subproject in the Ibercivis volunteer computing platform. The details of this effort will be described in Chapter 7.

In conclusion, this chapter highlights the importance of a thorough validation of VS protocols as an essential step to a VS campaign against any target of biological relevance. The lessons learned in this study can be broadly considered in further attempts to identify novel and safe TTR amyloid inhibitors. It should be stressed, however, that this study does *not* provide an exhaustive comparison of all virtual screening techniques currently available, and different or better results could have been obtained using other methods.

This chapter does offer a practical discussion (and application) of many central aspects of experimental design and performance metrics for virtual screening. Most importantly, it reflects an attempt to implement well-sought "standards" in the relatively young fields of molecular modelling and virtual screening, allowing reliable assessments to be made. It is our belief that only through such standards can these fields prove how important they really are in saving money and resources in the challenging (yet wonderful) effort of drug discovery.

Application of the docking-based VS protocols tested in this chapter to multiple protein targets will be discussed in the following chapter. Further validation of the best performing VS protocols through the experimental evaluation of virtual hits for their ability to prevent TTR fibril formation *in vitro* is presented in Chapter 7.

5. Acknowledgements

My first acknowledegement goes to Sarah Kinnings for her work with LigMatch and her support during the porting of the program onto the SeARCH cluster at the University of Minho, in Portugal. I thank Gerd Neudert for the helpful discussions and support with DSX and DrugScore^{CSD}. I also thank Prof. Gerhard Klebe for the stimulating discussions at the International Workshop "New Approaches in Drug Design & Discovery", in 2007, in Marburg (Germany), and for the free academic licensing of DrugScore scoring functions.

Chapter 6

Application of docking and scoring protocols to multiple protein targets: one recipe does not fit all

"I believe that the extraordinary should be pursued. But extraordinary claims require extraordinary evidence."

[Carl Sagan]

1. Introduction and theory

Even though TTR is associated with more than 90 amyloidogenic mutations (some of which with distinctive ligand-binding characteristics), to engage in a laborious endeavour as the setting up of dedicated infrastructures for high-throughput docking of millions of small molecules into one single pharmaceutical target soon appeared a short-sighted perspective for a platform that could become useful to many academic researchers. This realization incited us to test some of the docking and scoring protocols assembled for TTR on a broader set of targets of pharmaceutical interest. It is important emphasising, however, that the focus of this exercise was not the comparison of docking accuracy or the VS performance of the protocols against other docking and scoring engines in common use, but to analyse their application to different targets and thus understand in which cases can alternative scoring functions enhance/strengthen the discriminative power of the docking methodologies. In this context, we attempt to identify, amongst a limited group of academic-free tools, which docking and scoring-based protocols are most suitable for a given protein target, providing examples of successful cases that include well-known "workhorses" in the drug discovery field.

As denoted by Chang et al. [440], amongst the large number of docking programs available, AutoDock 4 and Vina are unique in the sense that not only are they freely accessible to academic and industrial researchers, hence being widely used, as they are also released under open source licenses (GNU General Public License and Apache Open Source License). In their article, the authors provided a fairly comprehensive description of the main differences between the methodologies used by the two programs, which have been summarized in Chapter 2. A variation of particular importance to our work is in the local search function: while AD4 employs stochastic search to generate random conformations for evaluation, Vina generates a gradient to search for local minima. This difference alone renders Vina a much faster program, which is a key aspect for high-throughput virtual screening. Moreover, Vina's authors demonstrated that the program offers higher accuracy over a set of 190 protein-ligand complexes (78% of cases within 2 Å RMSD) compared to AD4 (49%).

The essential goal of virtual screening is the prospective prediction of active ligands. It is, therefore, reasonable that one would like to evaluate docking by its ability to accurately estimate ligand binding affinities, but this is now beyond the field. Realistically, and as demonstrated in Chapters 2 and 5 of this thesis, the performance of docking and scoring methods is assessed in retrospective calculations using two main criteria. The ability to reproduce experimentally observed ligand poses within a certain tolerance limit (com-

monly 2 Å RMSD) – often referred to as *pose-fidelity* – is fundamental and well documented [416,418,421,447,448,611–615]. Just as important is *enrichment*, i.e. the ability to prioritize actives from amongst a data set of decoys, where a decoy is a member of the data set that does not bind to the target. Enrichment is critical for docking to be of any use in virtual screening [89,421,616,617]. Indeed, no matter how neat certain poses may be, the compounds that do not score well are unlikely to be experimentally tested.

In order to be meaningful while making comparisons between docking techniques, benchmarking sets must have a sufficiently high proportion of decoys, challenging to each of the methods when compared to active ligands. Property matching of decoys to ligands can be a helpful rationale while building adequate benchmarking sets. As pointed out in Chapters 4 and 5, if the ligands and decoys differ by physical properties alone, differentiation of ligands from decoys by docking can result from trivial properties like molecular weight, hydrophobicity, polarity, etc. The Directory of Useful Decoys (DUD) represents an attempt to reduce these biases and to provide the community of modellers a common touchstone by which to evaluate the performance of their docking and scoring protocols.

Exploring some of the concepts explained in Chapter 5 with respect to the design of VS experiments, in this chapter we make use of the DUD to evaluate the application of several VS protocols based on high-throughput docking (HTD) to multiple protein targets of pharmaceutical interest, and thus guide their selection and implementation on a large computational resource comprised by the desktop computers of thousands of volunteer citizens. Even though such effort is mainly fuelled by the high computational demands of the docking-based VS protocols assembled for TTR, the work of this chapter will assert its translational relevance in the realm of academic drug discovery.

1.1. The Directory of Useful Decoys (DUD)

The DUD (http://dud.docking.org/, accessed 06/06/2011) provides a publicly-available, unbiased set of active and decoy molecules, which is a widely accepted standard for the benchmarking of docking programs. The DUD includes a wide representation from several protein families such as protein kinases and nuclear hormone receptors, and a reasonable representation of other families such as metalloenzymes and serine proteases. In total, it comprises 40 target proteins along with their (respective) known actives and decoys that have been carefully selected in order to match the physicochemical properties of the actives, while holding dissimilar chemical topology.

As useful as virtual screening benchmarks may be, there are a few caveats justifying a very careful use. First and foremost, when training is to be carried out by selecting the coefficients of a scoring function based on the best retrospective performance, the program becomes trained to meet the benchmarks and there is a considerable risk of overfitting. A second important issue has to due with the specificity of the benchmarks. As explained in Chapter 4, the DUD, for example, has been tailored to address the weaknesses of docking methods and not those of any other virtual screening approaches, such as ligand-centric methods. Third, there are critical and hard-to-quantify biases in compounds comprising any benchmark. The pool from which the compounds are drawn is very unlikely representative of chemical space, and even the annotated actives are certainly an incomplete index of what will bind to a target (see **Figure 6.1**).

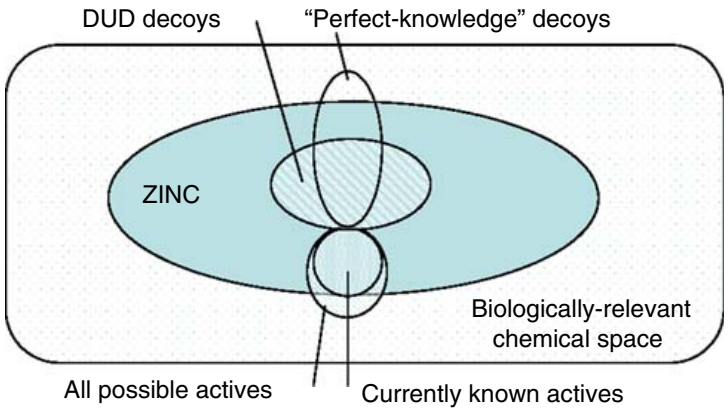


Figure 6.1. Schematic representation illustrating the problem of chemical space and decoys drawn from the ZINC database or derived from known actives, in both cases suffering from incompleteness bias.

Another known problem of the DUD is that of analogue bias (explained in Chapter 5, page 178), which can lead to deceptively high enrichment rates. Indeed, in its raw state, there are many targets in the DUD for which almost every active is a trivial analogue of a central structure [618]. To address this problem, Good and Oprea filtered the active molecules in the DUD using a loose lead-like cut-off (MW < 450 and AlogP < 4.5 for all targets except the Nuclear Hormone Receptors (NHR); AlogP < 5.5 for NHR targets), before using a reduced graph representation to cluster the molecules by chemotype. A more detailed description of the clustering procedures employed for DUD can found at:
http://dud.docking.org/clusters/summary.pdf ,accessed 06/06/2011.
These procedures resulted in the removal of large molecules with inadequate physicochemical properties and in a clustered annotation of the DUD data set:
http://dud.docking.org/clusters/, accessed 06/06/2011.
In this chapter, we make use of this clustered annotation to employ an arithmetic weighting scheme similar to the one proposed by Clark and Webster-Clark, wherein a weight is

assigned to each active molecule to reflect the size of the cluster to which it belonged [584]. The use of this procedure is particularly important when analysing subsets of targets with a high number of actives, where the problem of analogue bias is more significant.

1.2. DUD – Description by protein family

The DUD incorporates 40 protein targets that have been chosen based on the availability of known ligands, crystal structures and also former docking studies. The number of known ligands varies from 12 (~ 0.01% of the data set) to 416 (~ 0.4% of the data set), and, overall, a total of 2950 ligands are comprised. Except for the platelet-derived growth factor receptor kinase and the vascular endothelial growth factor receptor, all proteins have ligand-bound X-ray crystal structure deposited in the PDB. The targets were grouped into 6 families: nuclear hormone receptors, kinases, serine proteases, metalloenzymes, folate enzymes, and other enzymes [543].

Each of the following subsections provides a brief description of these target protein families, which were used in this chapter to test our VS protocols based on high-throughput docking.

1.2.1. Nuclear Hormone Receptors

Nuclear hormone receptors (NHR) represent a family of ligand-activated transcription factors that control multiple biological processes, such as lipid and glucose homeostasis, detoxification, cellular differentiation, embryonic development and organ physiology [619]. Accordingly, mutations in NHR affect these critical regulatory roles and cause several pathologies, including cancer, diabetes, and osteoporosis. NHR are therefore regarded as highly relevant therapeutic targets for the pharmaceutical industry [620].

The NHR present in the DUD data set are:

- Androgen receptor (AR);
- Estrogen receptor (ER);
- Glucocorticoid receptor (GR);
- Mineralocorticoid receptor (MR);
- Peroxisome proliferator activated receptor γ (PPARg);
- Progesterone receptor (PR);

- Retinoic X receptor α (RXRa).

Of all these targets, the ER is perhaps the one that has been more explored to benchmark docking methods and scoring functions [89,621–623]. A summary description of these targets – including their PDB entry, X-ray structure resolution, number of annotated active and decoy molecules – is provided in **Table E.1** of the Appendix.

1.2.2. Protein Kinases

Protein kinases represent one of the largest protein families, comprising 518 genes and approximately 1.7% of the human genome. These proteins share a conserved catalytic domain of ca. 300 amino acids responsible for phosphorylation mechanisms. Phosphorylation is associated with conformational changes in proteins, activating or deactivating them, which in turn can regulate key cellular processes, such as cell growth and differentiation, apoptosis, metabolic pathways, and membrane transport. Protein kinases regulate the spatial and temporal control of phosphorylation and are regulated by cellular regulatory mechanisms. Their constitutive activation, which can be caused by mutations, overexpression or the failure of the appropriate regulatory mechanisms, is a critical factor behind pathologies like cancer, diabetes and inflammatory disorders. Therefore, the targeted inhibition of kinases is a prominent strategy for the development of drugs to treat a wide range of diseases, and has received a lot of attention of the pharmaceutical industry [624–626].

The protein kinases comprised in the DUD data set are:

- Cyclin-dependent kinase 2 (CDK2);
- Epidermal growth factor receptor (EGFr);
- Fibroblast growth factor receptor kinase (FGFr1);
- Human heat shock protein 90 (HSP90);
- P38 mitogen-activated protein (P38 MAP);
- Platelet-derived growth factor receptor kinase (PDGFrb);
- Src(sarcoma)-family tyrosine kinase (SRC);
- Thymidine kinase (TK);
- Vascular endothelial growth factor receptor (VEGFr2).

TK, CDK2 and P38 MAP are perhaps the targets of greater focus and several docking and enrichment studies have been published [416,627–629]. In our lab at the University of

Leeds, Kinnings and Jackson conducted a large-scale structural comparison of protein kinase ATP-binding sites using GH8 (introduced in Chapter 2, page 73), which resulted in a functional classification of kinases based on the structural similarity of their binding sites [630].

A summary description of all protein kinases present in the DUD data set is given in **Table E.2** of the Appendix.

1.2.3. Serine Proteases

Serine proteases are enzymes that catalyse the hydrolysis of peptide bonds in proteins. They represent nearly one-third of all proteolytic enzymes and probably the most widely studied group of proteins in biology. This particular mechanistic class was initially differentiated by the presence of an Asp-His-Ser catalytic triad, which has evolved into at least four different *clans* (epitomised by chymotrypsin, subtilisin, carboxypeptidase Y and Clp protease). Serine proteases use the serine residue located in their active site for a nucleophilic attack of the targeted peptidic bond. Due to this ability, they are involved in several physiological functions in humans, such as digestion, immune response, blood clotting and reproduction. Consequently, several disorders have been linked to serine proteases. The most prominent examples include blood coagulation abnormalities, emphysema and chronic bronchitis, cystic fibrosis and chronic obstructive pulmonary disease, and also hepatitis C and herpes [631,632].

The serine proteases comprised in the DUD data set are:

- Coagulation factor Xa, also known as Stuart-Prower factor (FXa);
- Thrombin;
- Trypsin.

Several enrichment studies have been published on serine protease systems, mainly on FXa [612,621] and thrombin [629,633]. A summary description of the DUD target subset corresponding to the serine protease family is provided in **Table E.3** of the Appendix.

1.2.4. Metalloenzymes

Proteins and enzymes that contain a metal ion cofactor are ubiquitous in biology and it is estimated that nearly one-third of all proteins require metals to accomplish their functions. Indeed, the so-called metalloproteome is extremely diverse: metalloproteins and metalloenzymes are involved in numberless biological processes, from the catalysis of

bioorganic reactions, through transport and storage functions, all the way to signal transduction. One of the most prominent examples of a metalloprotein is that of hemoglobin, along with myoglobin, one of the very first protein structures to be determined and studied. These proteins use their iron-containing prosthetic group (heme) to bind and transport oxygen in red blood cells and muscle tissue (respectively). Another important example to the drug discovery field is the cytochrome P450 superfamily (CYP), which catalyses the oxidation of organic substances, from metabolic intermediates (e.g. lipids, steroidal hormones) to xenobiotics (e.g. drugs, toxic compounds). Indeed, CYPs are the major enzymes responsible for drug metabolism and bioactivation, being involved in approximately 75% of the total number of metabolic reactions [634–636].

The records of involvement of metalloenzymes in human disorders and pathologies are perhaps as vast as their repertoire of biological functions. Interestingly, one of the most well-known and sought pharmaceutical targets is a metallopeptidase called angiotensin-converting enzyme (ACE) that is involved in hypertension and heart disease [102,637].

The metalloenzymes comprised in the DUD data set are:

- Angiotensin-converting enzyme (ACE);
- Adenosine deaminase (ADA);
- Catechol *O*-methyltransferase (COMT);
- Phosphodiesterase 5 (PDE5).

Given the often important role of metals in protein-ligand interactions, there are several examples in the literature suggesting that metalloenzymes are difficult to model [638,639]. In a recent report, however, a statistic analysis demystified the presence of metal ions as a major "brick wall" in the molecular modelling field [640] (which, by no means, means they are easy to model).

A summary description of the protein targets listed as metalloenzymes in the DUD is given in **Table E.4** of the Appendix.

1.2.5. Folate Enzymes

Folate enzymes are involved in the conversion of folates (or more generally, folic acid and its derivatives). Folic acid is itself inactive, but its derivatives are fundamental for several biological functions across many organisms, namely DNA synthesis, repair and methylation. In humans, it is particularly important during pregnancy and infancy, were rapid cell division and growth must occur. The specific targeting of folate enzymes of

pathogenic organisms has been explored as forms of antibiotic and antiprotozoal therapies (malaria being a outstanding example). Several antifolate agents have also been designed for cancer chemotherapy [641,642].

The folate enzymes comprised in the DUD data set are:

- Dihydrofolate reductase (DHFR);
- Glycinamide ribonucleotide transformylase (GART).

DHFR is the most eminent folate enzyme target and some enrichment studies have been reported [633,643]. A summary description of DHFR and GART is provided in **Table E.5** of the Appendix.

1.2.6. Other Enzymes

A few other enzymes that do not fall inside the previous categories are also included in the DUD data set. These are also well-known cases, involved in different human disorders and/or pathologies.

The "other" enzymes comprised in the DUD data set are:

- Acetylcholinesterase (AChE);
- Aldose reductase (ALR2);
- AmpC β-lactamase (AmpC);
- Cyclooxygenase-1 (COX-1);
- Cyclooxygenase-2 (COX-2);
- Glycogen phosphorylase (GPB);
- HIV protease (HIVPR);
- HIV reverse transcriptase (HIVRT);
- Hydroxymethylglutaryl-CoA reductase (HMGR);
- *Mycobacterium tuberculosis* enoyl-acyl carrier protein reductase (InhA);
- Neuraminidase (NA);
- Poly(ADP-ribose) polymerase (PARP);
- Purine nucleoside phosphorylase (PNP);
- S-adenosyl-homocysteine hydrolase (SAHH).

Amongst the "other enzymes" group, HIV protease [447,628,643], neuraminidase [644], aldose reductase [633], HIV-RT [645], AChE [621] and COX-2 [644,646] are the most well studied cases with respect to the performance of VS methods. A summary description of this group of enzymes is given in **Table E.6** of the Appendix.

2. Computational methods

Molecular docking and scoring methods were introduced in the first chapter of this thesis (from page 18 to page 22). Moreover, in Chapter 2 the docking programs AutoDock 4 and AutoDock Vina were presented within the context of receptor-based studies of TTR-ligand interactions (from page 77 to page 82). In this chapter, the DrugScore family of stand-alone scoring functions is briefly described and we explore the application of a set of docking-based VS protocols assembled for the identification of novel TTR amyloid inhibitors to the forty targets in the DUD data set. The protocols were submitted to successive rounds of evaluation, employing different weighting schemes and performance metrics. In this section, the main details of the modelling procedures are briefly described and the results of the performance evaluation are provided in the next section.

2.1. Receptor and ligand preparation

Coordinates for all DUD protein targets were directly taken from the DUD website. All protein structures were processed by the *Dock Prep* tool of UCSF Chimera [463,647] in order to add hydrogen atoms and to fix problems, such as incomplete side-chains. The *prepare_receptor4.py* script included in MGLTools 1.5.4 was used to prepare the PDBQT format receptor files supported by AD4 and Vina [438,439]. All DUD decoys and ligands were obtained from the DUD website in the MOL2 format. Open Babel 2.2.3 was used for format conversion as required by the different docking packages [648,649]. PDB-formatted ligand files were also converted to PDBQT files using the *prepare_ligand4.py* script from MGLTools.

2.2. Docking and Scoring

All ligands in the DUD were docked into their respective receptor structures with AutoDock 4 version 4.2.2 and AutoDock Vina version 1.0.2. Ten independent genetic algorithm (GA) runs were executed on each AD4 run and processed using the native clustering

analysis with a 2.0 Å cut-off. The default *exhaustiveness*[1] of 8 was used in all Vina runs. A detailed description of AD4 and Vina was respectively given in subsections 1.4.1 and 1.4.2 of Chapter 2. In order to stress-test the docking programs and provide comparable results, we used docking simulation boxes of the exact same size of those employed by the authors of DUD (when using the DOCK docking program) [543]. These simulation boxes are fairly large, allowing not only for a full coverage of the target binding sites but also for an inclusion of neighbouring regions where compounds lacking the appropriate pharmacophore (decoys) may bind.

As in the previous chapter, the DrugScore^{CSD} and DSX standalone scoring functions were used to re-score the lowest-energy poses generated by AD4 (LE) and Vina, and the lowest-energy conformations in the largest clusters of poses generated by AD4 (LC). Given the relevance of this re-scoring procedure within the context of this chapter, a deeper description of the DrugScore and DSX scoring functions is given in this subsection.

2.2.1. DrugScore

Unlike empirical scoring functions that are trained by (and often biased towards) affinity data, knowledge-based scoring functions make use of large amounts of crystallographic data. Therefore, they are often more accurate at pose prediction and faster than empirical scoring functions. Knowledge-based scoring functions stem from the original concept that led to *knowledge-based potentials*, which were in turn based on the Boltzmann distribution:

$$\frac{n(i)}{N} = \rho(i) = \frac{e^{-E(i)/kT}}{Z(T)}$$

Equation 6.1

where $n(i)$ is the number of particles in a set of states i with energy E(i), N is the total number of particles in the system, T is the absolute temperature, k is the Boltzmann constant, $Z(T)$ is the partition function, and $\rho(i)$ is a state-dependent density (probability) function. Rearrangements of Equation 6.1 under the light of the theory of liquids [650] lead to the concept of *potential of mean force*, which basically corresponds to the Helmholtz free energy of two particles or, alternatively, the mean force acting on two particles due to the interactions with each other and their surroundings. The application of potentials of mean force to protein folding prediction [651–653] and to the scoring of protein-

[1] In Vina, the exhaustiveness parameter is linked to the time spent searching for the global minimum of the scoring function, which is varied heuristically according to features like the number of atoms and the flexibility of the ligand molecule.

ligand complexes has been reported [654]. In these studies, contact data extracted from structure repositories (e.g. PDB) are used to derive contact densities. However, the corresponding distribution functions are derived from particles that reside on different environments, and it has been stated that the derived statistical potentials are not potentials of mean force *per se* [655,656]. Indeed, even if a complete protein system is at an equilibrium state, two particular atoms may not inevitably be at thermodynamic equilibrium. This implies that the partition of the total free energy into pairwise atom–atom contributions is not valid, and that the distribution of atom–atom contact distances does not follow a Boltzmann-like distribution. As a result, terms like "potential of mean force" or "energy" have been eschewed and replaced by terms like "preference" or "quantity" [656]. Furthermore, since statistical potentials are not energies, the term "score" has replaced the term "energy, and the linear factor kT has been discarded in the master equation for knowledge-based scoring functions:

$$\text{score}(i) = -\ln\left(\frac{\rho(i)}{\rho_{\text{ref}}}\right)$$

Equation 6.2

In pairwise distance-dependent contributions, the total score for a complex of protein atoms a_{p} and ligand atoms a_{l} is determined as

$$\text{total score}_{\text{pair}} = \sum_{a_{\text{p}}}\sum_{a_{\text{l}}} \text{score}(p(a_{\text{p}}), l(a_{\text{l}}), r(a_{\text{p}}, a_{\text{l}}))$$

Equation 6.3

$$\text{score}_{\text{pair}}(p, l, r) = -\ln\left(\frac{\rho(p, l, r)}{\rho_{\text{ref}}}\right)$$

Equation 6.4

where $p(a_{\text{p}})$ and $l(a_{\text{l}})$ are the atom types and $r(a_{\text{p}},a_{\text{l}})$ is the distance of a_{p} and a_{l}. This set of equations can be applied to other structural features rather than distance dependent atom–atom scores alone, but it should be kept in mind that statistical potentials are more likely a class of heuristics and only through experiment can their meaning and utility be demonstrated [603].

Most knowledge-based scoring functions, such as PMF [654] and ASP [460], use a state of no interaction as reference state, ρ_{ref}, which can be seen as a weighting function for $\rho(i)$ while applying Equation 6.2. Instead, DrugScore uses a state of mean interaction:

$$\rho^{\mathrm{DS}}(p,l,r) = \frac{N(p,l,r)}{\Delta V(r)\sum_{r'} N(p,l,r')/\Delta V(r')}$$

Equation 6.5

$$\rho_{\mathrm{ref}}^{\mathrm{DS}} = \rho_{\mathrm{ref}}^{\mathrm{DS}}(r) = \frac{\sum_{p'}\sum_{l'}\rho(p',l',r)}{n_{\mathrm{p}} n_{\mathrm{l}}}$$

Equation 6.6

Here, n_{p} is the number of different protein atom types and n_{l} is the number of different ligand atom types. Moreover, in DrugScore the density functions are also probability functions. More important than the choice of the reference state, this difference is critical because averaging over all density functions without normalization would cause the reference to be dictated by contact types of high occurrence frequencies.

Another important advantage of the DrugScore reference state is the implicit inclusion of a volume correction (see Equation 6.5), which compensates for the fact that fewer than expected contacts are found at short distances. This fact is linked to the low accessibility of space occupied by other ligand or protein atoms. By contrast, since the quality of all DrugScore potentials is linked to one unique reference function, a possible disadvantage of its reference state is the "contagious" effect of an inaccurate density function over all resulting potentials.

Since all DrugScore potentials were derived from crystallographic information, only non-hydrogen atoms are used within this scoring function, avoiding problems such as pK_{a} shifts of ionizable groups due to changes in the electrostatic field induced by ligand binding. Of course, this may be regarded as a loss of information about directionality of polar interactions, but this is coped by the many-fold pair-preferences incorporated in the potentials. The first set of DrugScore potentials were derived from crystallographic information deposited in the Protein Data Bank, but important gains in accuracy were achieved when the data stored in the Cambridge structural database [657] was used instead. The two implementations are respectively referred to as DrugScore$^{\mathrm{PDB}}$ and DrugScore$^{\mathrm{CSD}}$. In this project, we have tested both versions. However, given the poor results obtained with DrugScore$^{\mathrm{PDB}}$, only the results obtained with the latest version of DrugScore$^{\mathrm{CSD}}$ are reported in this thesis.

2.2.2. DSX

DrugScore eXtended (or simply DSX) extends the DrugScore formalism to a more detailed atom type assignment. Indeed, the DSX pair potentials are based on Equation 6.5 and Equation 6.6, but instead of using *Sybyl* atom types (as in DrugScore) it uses *fconv*'s 158 atom types [658]. The importance and the impact of the inclusion of a wider *repertoire* of atom types in solving some of the problems associated to the reference state was well demonstrated by Neudert et al. [603]. As mentioned by the authors, the more atom types are available, the more the problem posed by particles residing in dissimilar environments (that should be treated as particles of different types) is waived. However, this strategy poses a possible shortcoming with respect to Equation 6.6: if a particular atom type is now split into two atom types, these can be considered twice. This is a problem when two contact types are equal in essence, in which case the effect of splitting up a particle is to double the weight of its interaction in the reference state.

To cope with this problem, the authors implemented a step of clustering of the density functions through a similarity measure. The definition of the density functions of DSX thus became

$$\rho^{DSX}(c,r) = \frac{\sum_{p\_l \in c} N(p,l,r)}{\Delta V(r) \sum_{p\_l \in c} \sum_{r'} N(p,l,r')/\Delta V(r')}$$

Equation 6.7

$$\rho_{\text{ref}}^{DSX} = \rho_{\text{ref}}^{DSX}(r) = \frac{\sum_{c'} \rho(c',r)}{n_{\text{c}}}$$

Equation 6.8

where c defines a cluster of contact types and n_{c} the number of clusters.

Besides this key modification, DSX substitutes the former DrugScore statistical potential to account for desolvation effects (based on "average" information about the changes in solvent-accessible surface upon complexation) by more sophisticated "SR" potentials derived from the PDB (score_{SR}), which account for the actual variations of SAS. Moreover, DSX features a newly developed knowledge-based torsion angle-dependent potential ($\text{score}_{\text{tors}}$) that allows for a local relaxation of docking poses, while handling atypical torsion angles produced by docking algorithms. This latter innovation will be explored later in Chapter 6. Overall, the total DSX-score for a specific protein-ligand complex is given by

$$
\begin{aligned}
\text{score}_{\text{total}} &= w_{\text{p}}\text{score}_{\text{pair}} + w_{\text{t}}\text{score}_{\text{tors}} + w_{\text{s}}\text{score}_{\text{SR}} \\
\text{score}_{\text{pair}} &= \sum_{a_i \in P} \sum_{a_j \in L} \text{score}_{\text{pair}}^{DSX}(c(a_i, a_j), r(a_i, a_j)) \\
\text{score}_{\text{tors}} &= \sum_{b} \sum_{T \in b} \frac{\text{score}_{\text{tors}}^{DSX}(t(T), \phi(t))}{n_{\text{T}}} \\
\text{score}_{\text{SR}} &= \sum_{a \in P} \text{score}_{\text{SR}}^{DSX}(c(a), \text{SR}(a)) \\
&\quad + \sum_{a \in L} \text{score}_{\text{SR}}^{DSX}(c(a), \text{SR}(a))
\end{aligned}
$$

Equation 6.9

where a is an atom from either the set of protein atoms P or the set of ligand atoms L, c is a cluster type, b is a central bond of a torsion T, t is a torsion type, n_{T} is the number of torsions for a given bond, SR is the SAS-ratio for a protein or ligand atom, and the $w_{\text{p/t/s}}$ are the weighting factors used [603].

Besides re-ranking by total score, the per-atom score (PAS) and per-contact score (PCS) options implemented in DSX were also explored in this chapter. These correspond just to the total score divided by the number of ligand atoms (PAS) or by the number of atom-atom interactions (PCS), respectively. Moreover, the scoring of torsion angles option in DSX was also tested to evaluate its ability to differentiate multiple poses generated by AD4 during the redocking of DUD cognate ligands.

2.3. Analysis of docking and scoring results

The accuracy of docking generated poses was quantified by the root-mean-squared deviation (RMSD) from experimentally determined ligand coordinates. All RMSD values were calculated with *fconv*. This program employs clique-based graph matching (based on the Bron-Kerbosch algorithm) to perform functional alignments between the molecules under comparison, and thus circumvent problems associated with molecular symmetry [658]. The VS performance of the docking and scoring protocols was assessed by plotting ROC curves [564], both with the (traditional) harmonic and the arithmetic weight (aw) scheme. Both the area under the ROC curves (ROC AUC) and area under the awROC curves were calculated to provide a measure of the overall VS performance across the entire data sets. Ideal distributions of actives and decoys yield AUC values near 1.0, while random distributions yield AUC values near 0.5. Normality within the entire DUD data set and each protein family was tested using the Shapiro-Wilk method. Given the levels of variance detected in the AUC values, all confidence intervals were based on 10,000 bootstrap replicates.

ROC Enrichment (ROCE) and arithmetich weighted ROCE (awROCE) were used for quantification of early active recovery. As described in the previous chapter (page 184), the ROCE metric expresses the percentage of actives recovered as a proportion of the percentage of decoys observed. Therefore, and unlike the enrichment factor metric, the ROCE is not affected by the ratio of actives to decoys in the benchmark sets [564]. This is particularly important in this chapter, since the size of the benchmarking sets varies considerable across the several DUD targets (even though the authors of DUD tried to keep the ratio of actives to decoys fix). Like with the EF, ROCE values greater than 1.0 represent enrichment against random selection. ROCE values at different false positive rates (FPR), corresponding to (top) fractions (0.5, 1, 2 and 5%) of the ranked list of decoys, were reported to show the effectiveness of the methods in terms of early enrichment. In some cases, the top fractions (i.e. the earliest stages of VS) correspond to a small number of ligands, especially in smaller data sets. In the cases where the top-scoring compounds are entirely (or almost entirely) populated by active molecules, the respective ROCE assumes an infinite value (or unrealistically high values). Accordingly, besides the mean values, median values were determined to describe early enrichment across the DUD target families and the entire data set (see **Table E.7** in the Appendix). All statistical data and ROC analyses were produced with the R software package.

3. Results and discussion

In the following subsections we report the results of performance evaluation obtained for each of the docking-based VS protocols applied to the DUD data set. Further information is provided in Section E of the Appendix.

3.1. Evaluation of pose prediction accuracy

Each cognate ligand included with the DUD targets was docked back into its respective protein structure using AD4 and Vina. Details about the protein structures, such as PDB accession codes and experimental resolution, are provided in the Appendix, from **Table E.1** to **Table E.6**. As for TTR, the heavy atom RMSD was calculated against the X-ray ligand conformation. **Figure 6.2** briefly illustrates notable examples where both AD4 and Vina performed well or failed at pose prediction, along with cases with differing success. All results were tabulated for both the top-ranked poses and the best poses (lowest-RMSD) generated by the docking programs. Since only one single experimental complex was available for each target in the DUD (with the exception of PDGFrb, where no experimental structure is available), and considering that many of the obtained RMSD values

seemed unrealistically high and of limited meaning[1], we were particularly wary while generating descriptive statistics for our results on docking accuracy against the DUD targets. Therefore, instead of reporting what might be regarded as deceptive statistical averages, **Figure 6.3** shows the percentage of complexes over a distribution of RMSD values, contrasting the redocking performance of AD4 and Vina.

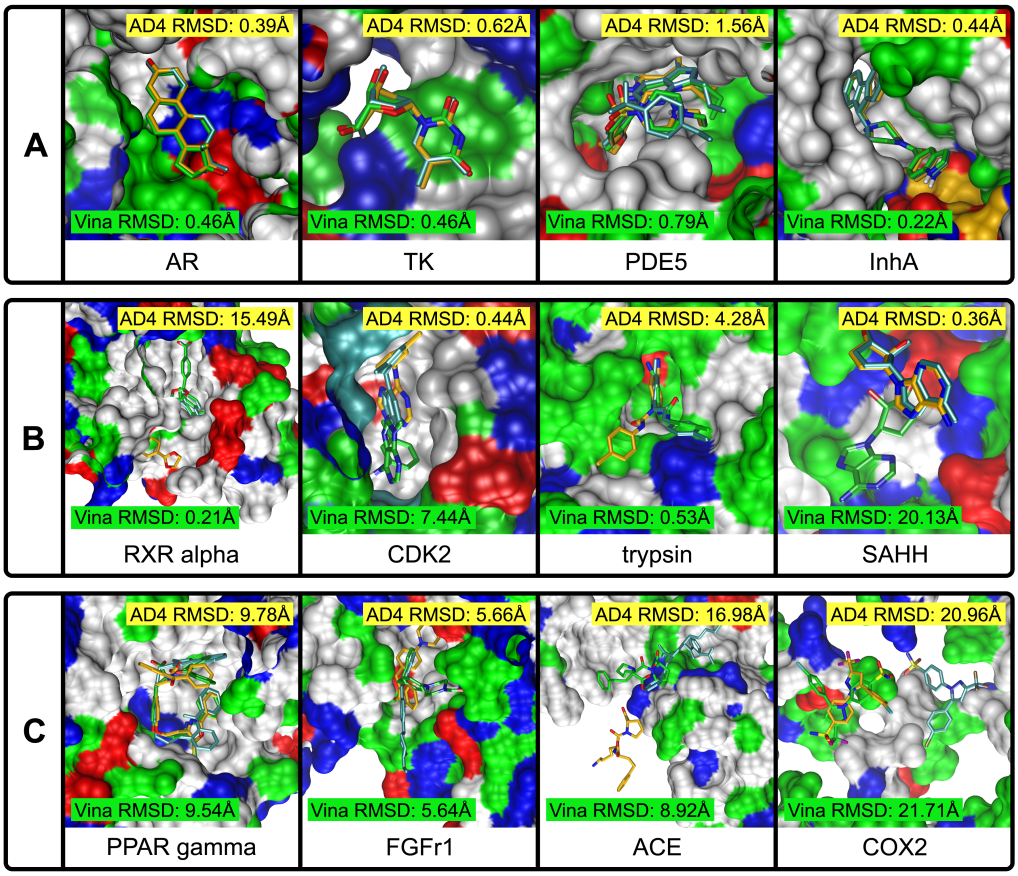


Figure 6.2. Comparison between experimental ligand binding modes of selected DUD targets and the top-scoring poses generated by AD4 and Vina. Pane A: cases where both AD4 and Vina offered good or exceptional predictions (RMSD values ≤ 2.0 Å or ≤ 0.5 Å, respectively). Pane B: cases where AD4 and Vina had contrasting success at pose prediction. Pane C: cases where both AD4 and Vina failed to predict correct ligand poses. The images were generated using VMD. Abbreviations: androgen receptor (AR), thymidine kinase (TK), phosphodiesterase 5 (PDE5) and enoyl-ACP reductase (InhA), retinoic X receptor α (RXRa), cyclin-dependent kinase 2 (CDK2), trypsin, S-adenosyl-homocysteine hydrolase (SAHH). peroxisome proliferator activated receptor γ (PPARg), fibroblast growth factor receptor kinase (FGFr1), angiotensin-converting enzyme (ACE) and cyclooxygenase-2 (COX2).

[1] As discussed in Chapter 2, RMSD values can often be unrealistically high simply because a small part of the ligand adopts an alternative, wrongly-placed conformation, even though a considerable part of the ligand structure is correctly docked.

Interestingly, the profiles obtained for the best-predicted poses suggest that the local search algorithms of AD4 and Vina preformed very similarly, with both docking programs providing near-native ligand poses (RMSD ≤ 2.0 Å) for approximately 60% of the complexes. Amongst the cases where both programs consistently failed to generate low-RMSD poses are ACE, COMT, COX2, PPAR, PR and SRC. These failures may be the result of insufficiencies of the sampling algorithms rather than of the scoring functions [417]. Less success and more noticeable differences were found while analysing the discriminative power of scoring functions towards the ranking of the generated poses. Indeed, low-RMSD poses were not systematically scored favourably over high-RMSD poses. Compared to the free energy function of AD4, the Vina scoring function was able to identify a higher percentage of low-RMSD conformations by ranking them first: Approximately 53% of Vina's top-ranked poses were below 2.0 Å, whereas only 43% of AD4's top-ranked poses were below this reference threshold.

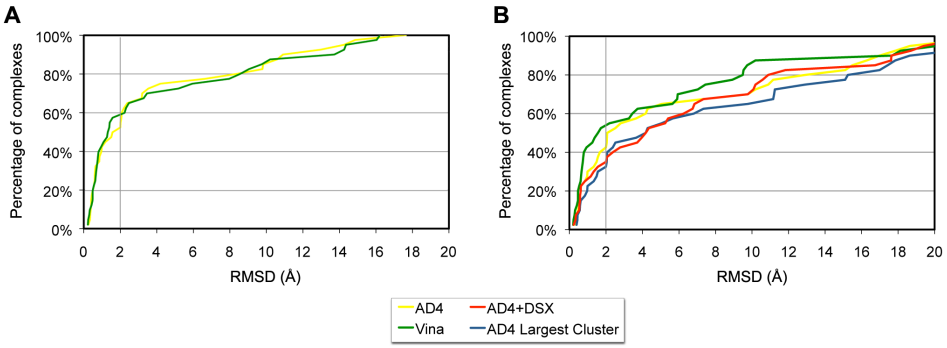


Figure 6.3. Cumulative RMSD distribution plots for X-ray ligand re-docking with 40 DUD complexes. Pane A: the RMSD values refer to the best ligand poses generated by AD4 and Vina. Pane B: the RMSD values refer to the top-ranked poses scored by AD4 (in yellow), Vina (in green) and DSX upon docking with AD4 (in red). Cumulative RMSD distributions for the lowest energy conformations belonging to the largest clusters of poses generated by AD4 are also plotted (in blue).

While considerably smaller than that found by Vina's authors for their test set of protein complexes, this difference supports the global conclusion that Vina represents a promising compromise between docking accuracy and speed [439,440]. Globally, these results also seem to corroborate the view that alternative docking conformations, rather than just the lowest-energy ones, should be considered when evaluating docking programs. Indeed, it has been suggested that better discrimination could be attained by considering the lowest-energy conformation amongst the largest pool of generated poses [417], instead of the lowest-energy conformation overall. To test this idea, RMSD values were computed for the lowest-energy pose of the largest cluster of poses returned by AD4. While better discrimination of correct poses was observed for certain targets (namely, CDK2, FGFr1 and PARP), in general the results indicate that no real benefit is obtained

from this strategy (see pane B in **Figure 6.3**). Finally, DSX's ability to discriminate correct ligand poses was also evaluated, by turning on the scoring of torsion angles upon minimization of the ligands to a close local minimum, but, again, no significant advantages were detected.

3.2. Virtual screening performance – Overview

Even though the protocol followed by the authors of the DUD attempted to yield benchmarking sets with similar ratios of actives to decoys, differences in these proportions are to be expected when comparing multiple sets, due, for example, to failures during the docking of ligands/decoys into their respective target structures. Therefore, to eschew biases associated to enrichment plots when the ratio of actives to decoys grows large, in this chapter we used receiver operating characteristic (ROC) curves to compare the performance of our docking-based VS protocols on multiple targets. For a deeper understanding of the differences between ROC and enrichment curves, recall subsection 1.2.4 in Chapter 5.

Figure 6.4 illustrates a group of cases where the correct prediction of bound ligand poses (i.e. good pose fidelity by the docking algorithm), assessed against the available X-ray ligand conformation, was accompanied by moderate or even good VS performances. Here, a correctly predicted pose was defined as having an RMSD inferior to 2.5 Å, and a moderate (or good) VS performance was defined by a ROC AUC greater than 0.65. Worth highlighting are the performance of AD4's FEF against the estrogen receptor (ER) and the poly(ADP-ribose) polymerase (PARP), and the perfomance of the AD4+DSX (PCS) protocol and the Vina+DSX (PCS) protocol respectively against the mineralocorticoid receptor (MR) and against HIV protease (HIV protease). While still far from what may be perceived as ideal VS performances (i.e. ROC AUCs approaching 1.0), these cases epitomise the most expected outcome of a docking-based VS experiment, wherein good docking accuracy is accompanied by reasonable or good discrimination of active over inactive compounds.

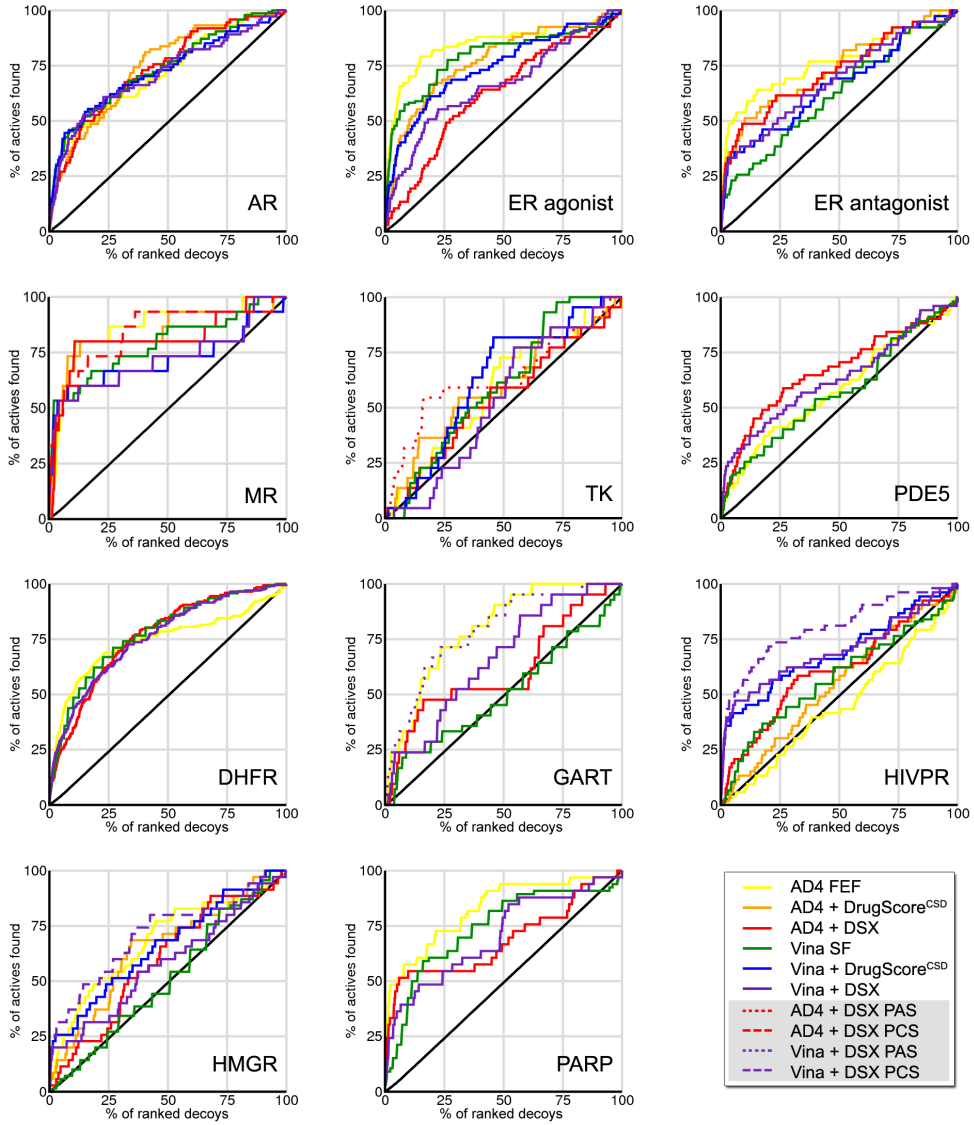


Figure 6.4. ROC curves illustrating the overall performance of docking and scoring protocols against selected DUD targets. The plots shown correspond to cases of successful docking pose prediction for the cognate X-ray ligand (here defined as having an RMSD < 2.5 Å) that are associated with modest or good VS performances (ROC AUC > 0.65) within their respective protein family. The six main protocols are coloured as follows: AD4 Free Energy Function (yellow), AD4 + DrugScore^{CSD} (orange), AD4 + DSX (red), Vina Scoring Function (green), Vina + DrugScore^{CSD} (blue), Vina + DSX (purple) and random selection (black). Variations to DSX scoring are shown in dotted lines (per-atom score) and in dashed lines (per-contact score). Abbreviations: androgen receptor (AR), estrogen receptor (ER), mineralocorticoid receptor (MR), thymidine kinase (TK), phosphodiesterase 5 (PDE5), dihydrofolate reductase (DHFR), glycinamide ribonucleotide transformylase (GART), HIV protease (HIVPR), hydroxymethylglutaryl-CoA reductase (HMGR) and poly(ADP-ribose) polymerase (PARP).

By contrast, **Figure 6.5** presents cases where acceptable or even good overall VS performances did not depend on a correct docking prediction of the X-ray ligand conformation.

Here, the most important examples are the peroxisome proliferator activated receptor γ (PPAR), the retinoic X receptor α (RXRa) and the purine nucleoside phosphorylase (PNP), where protocols using DrugScore^{CSD}, DSX and AD4's FEF were able to offer reasonable or even good enrichment in actives. In principle, these cases represent an unexpected outcome of a docking-based VS protocol, for its ability to discriminate active molecules over inactive ones should be coupled with its ability to identify the most favourable interactions between the active compounds and target's active/binding site.

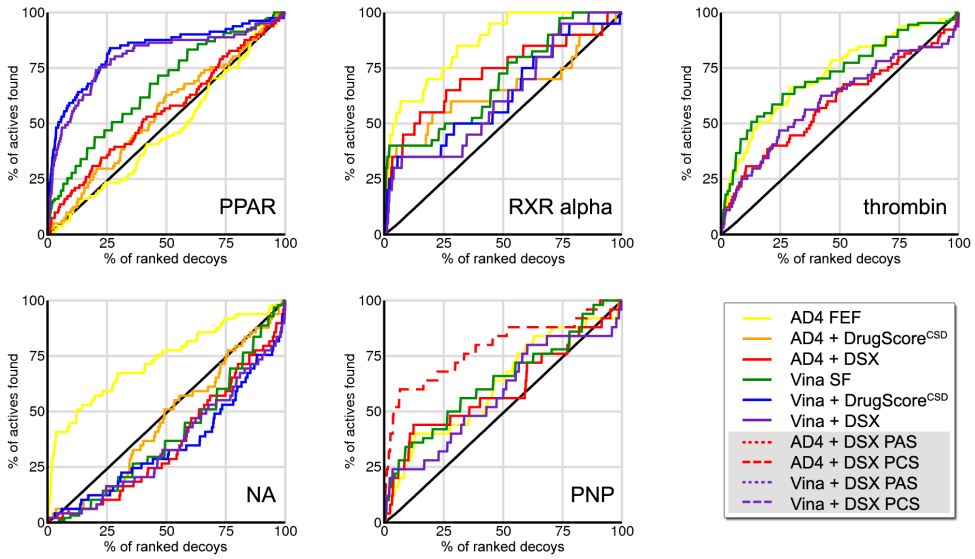


Figure 6.5. ROC curves illustrating the overall performance of docking and scoring protocols against selected DUD targets. The shown plots correspond to cases of unsuccessful docking pose predictions for the cognate X-ray ligand (RMSD > 2.5 Å) that are associated with at least one VS protocol of acceptable or good VS performance (ROC AUC > 0.70) within their respective protein family. The six main protocols are coloured as follows: AD4 Free Energy Function (yellow), AD4 + DrugScore^{CSD} (orange), AD4 + DSX (red), Vina Scoring Function (green), Vina + DrugScore^{CSD} (blue), Vina + DSX (purple) and random selection (black). Variations to DSX scoring are shown in dotted lines (per-atom score) and in dashed lines (per-contact score). Abbreviations: peroxisome proliferator activated receptor γ (PPAR), retinoic X receptor α (RXRa), neuraminidase (NA), and purine nucleoside phosphorylase (PNP).

In **Figure 6.6**, an analysis focused on early enrichment rates is provided for cases where higher discrepancy between early recovery rates and overall performance was verified.

Table 6.2 to **Table 6.7** of the following subsections report the best performing docking and scoring protocols for each target across five protein families in the DUD, including different performance metrics such as the AUC and ROCE. **Figure 6.8** reports the mean area under the ROC curves obtained for all tested protocols broken down by protein family, and a table comprising all mean AUC and median/mean ROCE values is provided

in the Appendix (**Table E.7**). The mean AUC values for each docking and scoring protocol across the entire DUD are tabulated in **Table 6.11**, along with the values of 95% confidence intervals.

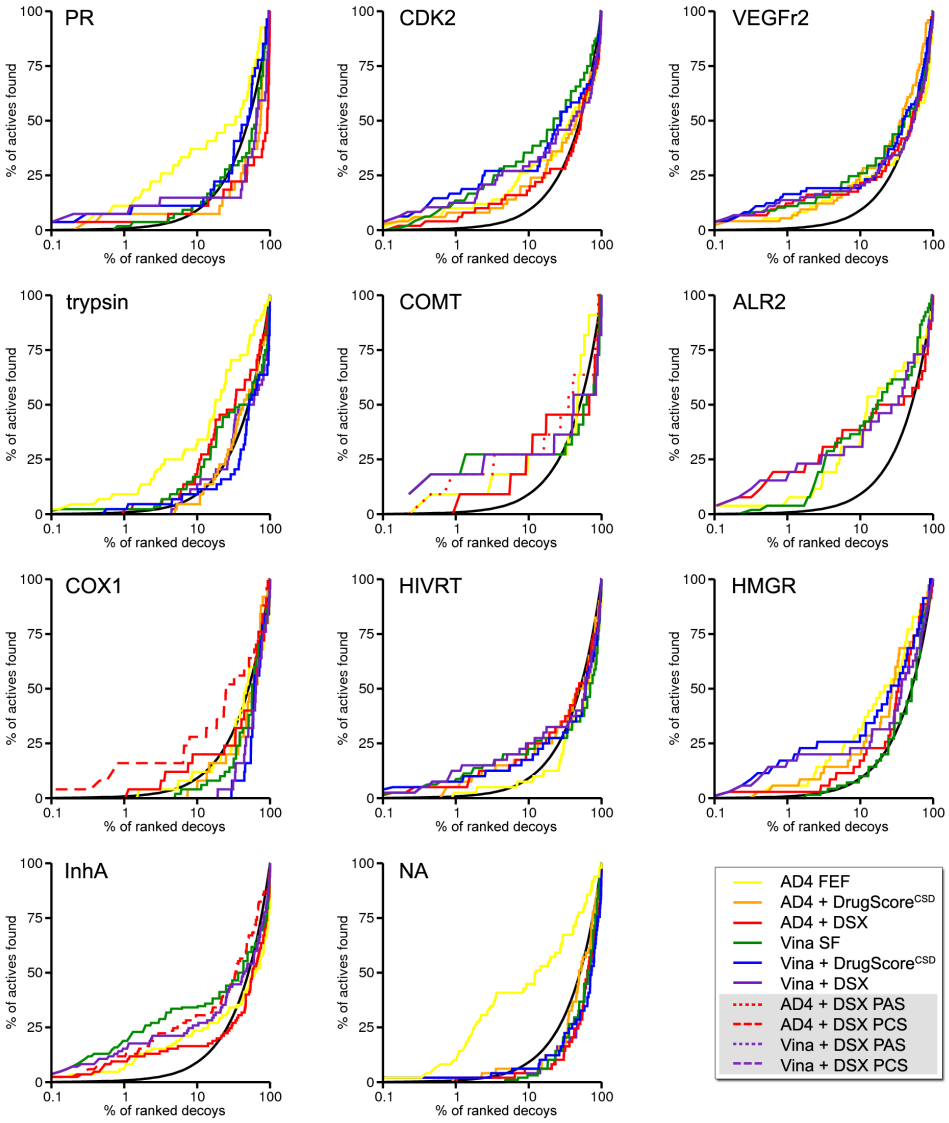


Figure 6.6. ROC curves focusing on the early performance of several docking and scoring protocols against selected DUD targets. The cases shown are associated with some modest or poor overall VS performances (ROC AUC < 0.65) yet presenting some good early recovery rates within their respective protein family. In order to facilitate the visualization of early performance, the x-axis is plotted on logarithmic scale. The six main protocols are coloured as follows: AD4 Free Energy Function (yellow), AD4 + DrugScore^{CSD} (orange), AD4 + DSX (red), Vina Scoring Function (green), Vina + DrugScore^{CSD} (blue), Vina + DSX (purple) and random selection (black). Variations to DSX scoring are shown in dotted lines (per-atom score) and in dashed lines (per-contact score). Abbreviations: progesterone receptor (PR), cathecol *o*-methyltransferase (COMT), aldose reductase (ALR2), cyclooxygenase-1 (COX1), HIV reverse transcriptase (HIVRT), hydroxymethylglutaryl-CoA reductase (HMGR), enoyl-ACP reductase (InhA) and neurominidase (NA).

The relative performance of each protocol varied considerably when compared with the entire DUD data set, and even though all mean AUCs were above the 0.5 threshold (corresponding to no discrimination), several protocols yielded AUCs below this value for many targets. Indeed, variability in the results was verified not only between docking and scoring protocols but also between targets. Some targets gave distinct results with the same set of VS protocols (COX1, COX2, NA, PNP, PR, RXR alpha, SRC, trypsin), whereas some other targets had very similar results with all protocols (ACE, ALR2, AR, DHFR, MR, SAHH, VEGFr2). Only a few targets yielded good or acceptable enrichment profiles for all protocols (MR and DHFR), or consistently random or worst than random for all protocols (EGFr, FGFr1, p38 MAP and PDGFrb). Most DUD targets returned a range of AUC values, showing that at least some of the docking and scoring protocols were reasonably successful for one of the targets. Variations to the default scoring of DSX were tested in this study, whereby re-ranking was produced based on per-atom scores (PAS) and per-contact scores (PCS). While no noticeable effects were observed in the case of TTR, significant improvements in enrichment were witnessed for several DUD targets (particularly for COMT, FXa, HIVPR, MR, PDE5 and PNP).

3.3. Performance analysis using arithmetic weighting

Although docking and scoring protocols are less affected by the analogue bias problem than similarity-based methods, here we provide a complementary analysis by exploring an arithmetic weighting procedure whereby a weight is assigned to each active molecule to reflect the size of the cluster to which it belonged. For example, if a given cluster contained seven actives of similar chemotypes, each of those molecules would be given a weight of one-seventh. As explained in the methods section (subsection 1.1 starting on page 215), this weighting procedure was based on the clustered annotation of the DUD data set produced by Good and Oprea, which can be found at:
http://dud.docking.org/clusters/summary.pdf (accessed 06/06/2011).
From the viewpoint of performance analysis, the complementary view provided by the arithmethic weighting procedure has an obvious advantage over the non-weighted procedure when analysing subsets of targets with a high number of actives and especially a high number of clusters of actives. In these subsets, the problem of analogue bias may contribute to an overestimation or an underestimation of the performance of a VS protocol: if one particular active compound (chemotype) is assigned with a favourable or disfavourable score, it is likely that all compounds that hold high similarity to that active will be equally selected or discarded by the protocol. Compare to the non-weighted metric, the arithmetic weighted metric provides more clear insights about the protocols' ability to prioritize structurally diverse actives, which is an expected outcome of virtual screen-

ing. **Table 6.1** and **Figure 6.7** report results for proeminent examples of each target family possessing a high number of clusters of actives (at least 10 clusters). The most distinctive cases are CDK2 (32 clusters of actives), EGFr (40 clusters) and COX2 (44 clusters).

Table 6.1. Arithmetic weighted ROC AUC (awAUC) and ROC Enrichment (awROCE) values at 0.5, 1, 2, and 5% for the best performing docking and scoring protocols across selected DUD targets. The best protocols (awAUC ≥ 0.75) are marked in bold and grey shade.

Target	Best performing protocol	Overall Enrichment (awAUC)	Early Enrichment (awROCE)			
			0.5%	1%	2%	5%
ER agonist	**AD4 FEF**	0.82	47.5	79.3	45.8	17.2
CDK2	Vina SF	0.65	31.8	22.9	11.7	6.4
EGFr	Vina + DrugScore^{CSD}	0.70	8.0	5.1	3.2	2.0
FXa	**Vina + DSX PAS**	0.76	7.8	3.4	1.6	0.6
ACE	Vina + DrugScore^{CSD}	0.58	21.7	15.4	8.1	3.2
PDE5	**Vina + DSX PAS**	0.76	38.3	21.9	16.8	8.5
DHFR	**AD4 FEF**	0.75	17.3	15.6	11.8	8.9
ALR2	AD4 + DSX	0.65	153.0	48.0	16.4	8.4
COX2	Vina + DSX PCS	0.65	3.5	2.0	1.9	1.4
InhA	Vina + DSX PCS	0.64	23.3	10.9	6.8	5.3

Over the following subsections, the results reported in **Table 6.1** will be contrasted with those obtained using non-weighted metrics. It is worth mentioning that, albeit rare, there are a few cases where the two quantification approaches (arithmetic versus non-arithmetic weighting) show disagreement. The clearest examples are EGFr and ACE, where the non-weighted procedure seems to overestimate the protocols' overall performance, while the arithmetic weighting scheme reveals limited ability to select diverse actives, and FXa, PDE5 and COX2, where the non-weighted procedure seems to underestimate the protocols' overall performance, while the arithmetic weighting scheme suggests some ability to select diverse actives. ACE and ALR2, on the other hand, are two interesting examples were the non-weighted procedure clearly underestimates the early performance of the protocols, and particularly their ability to populate the top fraction of the ranked lists with structurally diverse actives.

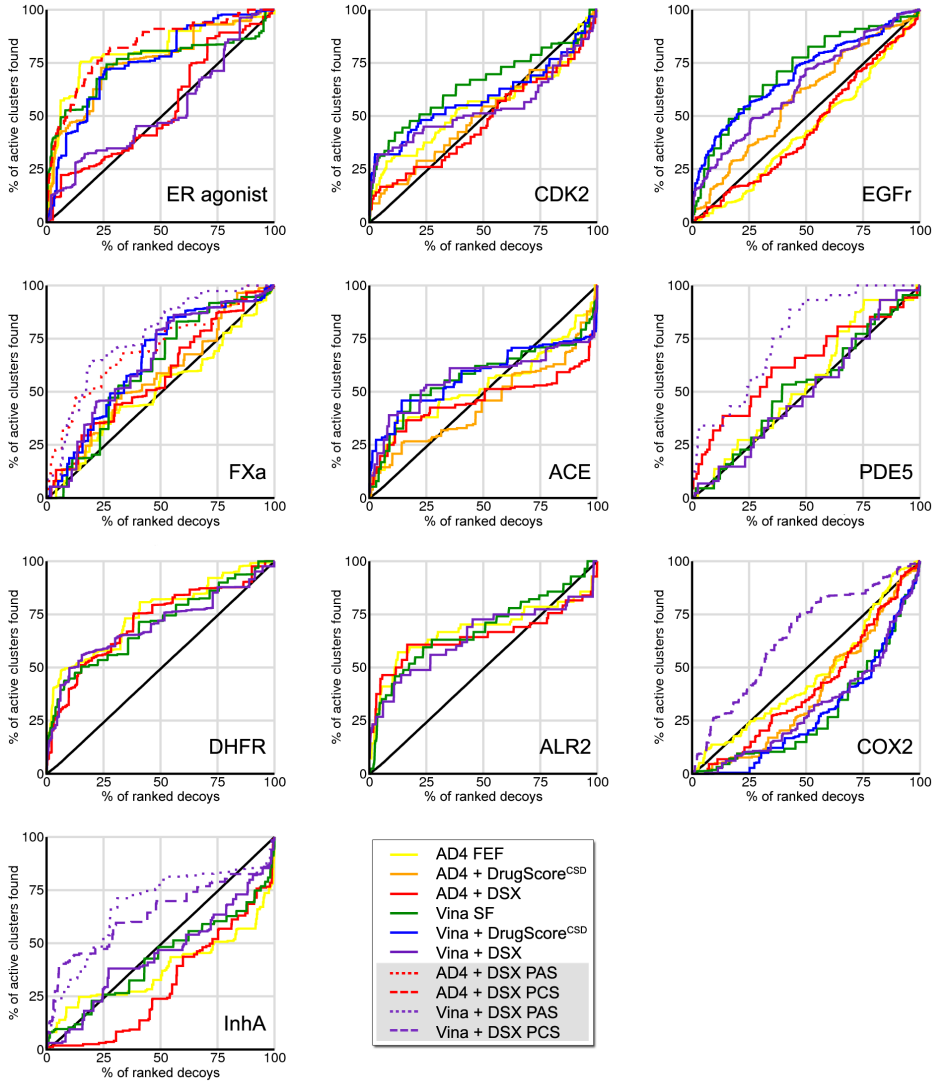


Figure 6.7. Arithmetic weighted ROC curves illustrating the performance of docking and scoring protocols against selected DUD targets. The cases shown are subsets of each protein family possessing at least 10 clusters of actives, according to the assignments by Good and Oprea. The six main protocols are coloured as follows: AD4 Free Energy Function (yellow), AD4 + DrugScore^{CSD} (orange), AD4 + DSX (red), Vina Scoring Function (green), Vina + DrugScore^{CSD} (blue), Vina + DSX (purple) and random performance (black). Variations to DSX scoring are shown in dotted lines (per-atom score) and in dashed lines (per-contact score). Abbreviations: estrogen receptor (ER), cyclin-dependent kinase 2 (CDK2), epidermal growth factor receptor (EGFr), coagulation factor Xa (FXa), angiotensin-converting enzyme (ACE), phosphodiesterase 5 (PDE5), dihydrofolate reductase (DHFR), aldose reductase (ALR2), cyclooxygenase 2 (COX2) and enoyl-ACP reductase (InhA).

3.4. Virtual screening performance broken down by protein targets

In this subsection, the results of the VS performance studies are contrasted across the protein target families in the DUD data set, and particular emphasis is given to the most interesting individual cases. **Figure 6.8** presents the mean area under the ROC curve (ROC AUC) for each docking and scoring protocol against each protein target family, along with the respective 95% confidence intervals.

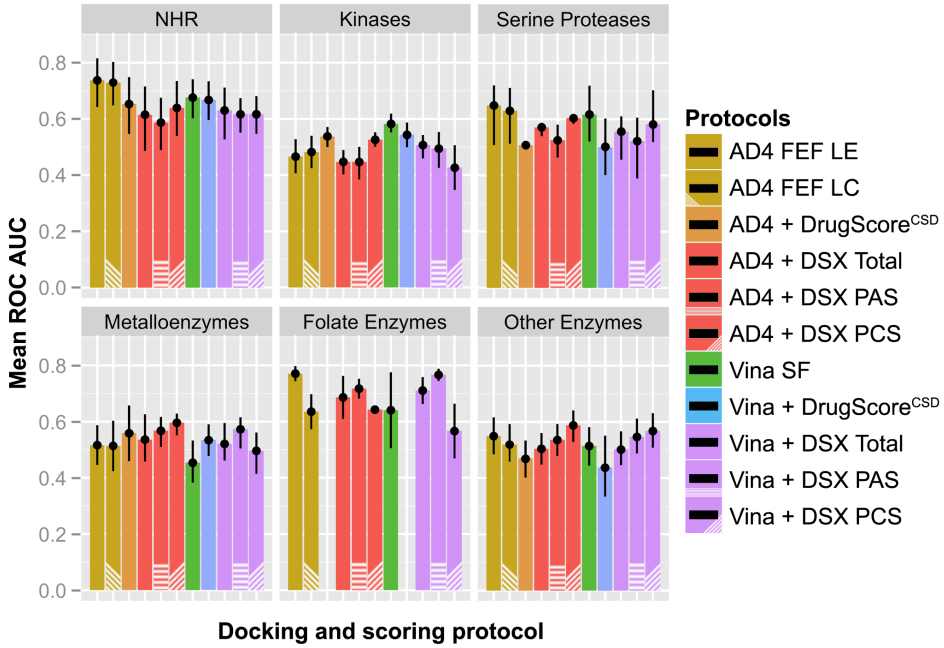


Figure 6.8. Statistical results for docking and scoring protocols using the DUD, broken down by protein family. Each bar represents the mean area under the ROC curve obtained by each docking and scoring protocol across the six subsets of protein targets in each family. Protocols are coloured as follows: AD4 Free Energy Function (yellow), AD4 + DrugScore^{CSD} (orange), AD4 + DSX (red), Vina Scoring Function (green), Vina + DrugScore^{CSD} (blue) and Vina + DSX (purple). The point ranges in black express bootstrap-based 95% confidence intervals of the mean. Abbreviations: nuclear hormone receptor (NHR), AutoDock 4 (AD4), Free Energy Function (FEF), lowest-energy pose (LE), lowest-energy pose belonging to the largest cluster (LC), per-atom score (PAS), per-contact score (PCS).

3.4.1. Nuclear Hormone Receptors

Consisting of eight protein targets, the nuclear hormone receptor (NHR) subset of the DUD yielded the best results with most docking and scoring protocols. For 6 out of the 8 individual targets, the best performing protocol encountered obtained a ROC AUC above

0.75 (see **Table 6.2** and the first 4 plots in **Figure 6.4**). Of these, three targets (ER agonist, MR and RXR alpha) yielded the highest AUC values across the entire DUD (0.85, 0.84 and 0.86, respectively) with AD4 and its Free Energy Function (FEF). The worse performances obtained for PR are likely to reflect the limited ability of the docking algorithms to predict correct ligand poses – as verified by re-docking of the respective cognate X-ray ligand.

It should be noticed, however, that poor docking predictions were also seen for PPAR gamma (using both AD4 and Vina) and RXR alpha (using AD4), while good overall enrichment and early recovery rates were obtained both for PPAR gamma, using Vina-based protocols (Vina + DrugScore^{CSD} and Vina + DSX), and for RXR alpha, using AD4 FEF (see the first two plots in **Figure 6.5**). As mentioned in subsection 3.2, this is an unexpected outcome of docking-based VS protocols. Therefore, these particular cases are subject to a deeper analysis in subsection 0.

The bar plots in **Figure 6.8** show that, in general, the docking and scoring protocols performed well amongst the NHR family, with AD4 offering slightly better results. Equally, good results were obtained in terms of overall and early recovery of distinct active chemotypes of ER agonists (10 clusters), as disclosed by an awAUC of 0.82 and awROCE values of 47.5, 79.3, 45.8 and 17.2 respectively at 0.5, 1, 2 and 5% false positive rates (see **Table 6.1** and the first plot of **Figure 6.8**).

Table 6.2. Best performing docking and scoring protocols for the nuclear hormone receptor targets in the DUD. Cases of high accuracy (RMSD ≤ 2.0 Å) and/or performance (awAUC ≥ 0.75) are highlighted in grey shade and bold, respectively.

Target	Accuracy at pose prediction: AD4 vs. Vina (RMSD in Å)	Best performing screening protocol (docking + scoring)	Overall enrichment (ROC AUC)	Early enrichment (ROCE)			
				0.5%	1%	2%	5%
AR	0.39 / 0.46	**AD4 + DrugScore^{CSD}**	**0.75**	**54.0**	**22.5**	**14.4**	**5.8**
ER agonist	0.54 / 0.61	**AD4 FEF**	**0.85**	**52.0**	**79.3**	**43.1**	**16.2**
ER antagonist	0.97 / 1.25	**AD4 FEF**	**0.76**	**39.0**	**36.8**	**36.8**	**12.4**
GR	0.58 / 0.31	AD4 FEF	0.72	21.3	12.9	10.4	4.9
MR	0.57 / 17.67	**AD4 + DSX PCS**	**0.84**	**133.0**	**88.7**	**30.3**	**14.4**
PPAR gamma	9.78 / 9.54	**Vina + DrugScore^{CSD}**	**0.83**	**123.0**	**55.5**	**31.3**	**13.7**
PR	17.85 / 18.13	AD4 FEF	0.65	24.7	15.9	12.3	6.9
RXR alpha	15.49 / 0.21	**AD4 FEF**	**0.86**	**12.5**	**30.0**	**31.8**	**14.8**

3.4.2. Protein Kinases

As shown in **Figure 6.8**, protein kinases yielded the lowest mean ROC AUC values, therefore representing the hardest challenge in the DUD for most docking and scoring protocols. On average, Vina and its built-in scoring function performed better against the kinases, with a mean ROC AUC of 0.58. This number is only slightly above the threshold of no discrimination, mirroring the difficulties in identifying efficient protocols to handle this group of targets, in general, but also the very poor results obtained with a few particular targets (HSP90 and PDGFrb; see **Table 6.3**). PDGFrb, in particular, yielded the worst VS performance results across the entire DUD data set, although it is worth noting that no experimental structure is available for this target (a theoretical model is supplied in the DUD).

Despite the fact that Vina performed poorly at cognate ligand pose prediction for CDK2, the Vina + DrugScore^{CSD} protocol performed very well in terms of early enrichment (ROCE values of 145.0, 33.2, 12.5 at 0.5, 1 and 2% false positive rates; see **Figure 6.6**). However, the protocol showed poor overall enrichment (AUC of 0.59). Amongst all protocols, Vina SF offered the highest AUC (0.65) and awAUC (0.65). Moreover, CDK2 is a case where evaluation with the arithmetic weighting procedure provided very similar results (awROCE values of 148.0, 34.0 and 12.7 at 0.5, 1 and 2% false positive rates for Vina + DrugScore^{CSD}), which can be linked to the high diversity of the CDK2 set of active molecules in the DUD (32 clusters amongst 50 active molecules).

Similar results were obtained for EGFr: although Vina SF offered the highest overall enrichment, it provided weak early recovery compared to the Vina + DrugScore^{CSD} protocol, which in turn provided worst overall enrichment. However, the EGFr target represents a case where the analogue bias problem of the DUD allows for deceptive interpretations. Indeed, the AUC value of 0.58 overshadowed Vina SF's real ability to discriminate active EGFr chemotypes (awAUC of 0.71).

Table 6.3. Best performing docking and scoring protocols for the protein kinases in the DUD. Cases of high accuracy (RMSD ≤ 2.0 Å) are highlighted in grey shade.

Target	Accuracy at pose prediction AD4 vs. Vina (RMSD in Å)	Best performing screening protocol (docking + scoring)	Overall enrichment (ROC AUC)	Early enrichment (ROCE)			
				0.5%	1%	2%	5%
CDK2	0.44 / 7.44	Vina SF	0.65	31.1	22.5	11.4	6.7
EGFr	1.49 / 5.90	Vina SF	0.58	0.8	0.9	1.4	1.5
FGFr1	5.66 / 5.64	Vina SF	0.53	0.0	0.4	0.4	1.0
HSP90[a]	7.34 / 0.76	AD4 FEF	0.58	0.0	0.0	0.0	0.0
P38 MAP	15.10 / 0.66	Vina + DrugScore^{CSD}	0.64	5.8	5.8	5.2	2.8

Target	Accuracy at pose prediction AD4 vs. Vina (RMSD in Å)	Best performing screening protocol (docking + scoring)	Overall enrichment (ROC AUC)	Early enrichment (ROCE)			
				0.5%	1%	2%	5%
PDGFrb	0.47 / 9.73	Vina SF	0.53	0.0	0.4	0.4	1.0
SRC	3.65 / 3.26	Vina SF	0.69	3.0	2.1	2.3	3.2
TK[a]	0.62 / 0.46	AD4 + DSX PCS	0.68	15.0	5.6	5.3	1.9
VEGFr2	2.81 / 1.38	Vina + DrugScore^{CSD}	0.56	123.0	32.8	12.7	4.2

[a] This result was based on AD4 scores assigned to the lowest energy conformation of the largest cluster of conformations.

On these grounds, of all tested protocols Vina + DrugScore^{CSD} was considered the most suited for the EGFr target (see **Figure 6.7** and **Table 6.1**), offering the best compromise between overall (awAUC of 0.70) and early enrichment (awROCE values of 8.0, 5.1 and 3.2 at 0.5, 1 and 2% FPR). Interestingly, the Vina + DrugScore^{CSD} also provided the best overall enrichment for p38 MAP (AUC 0.65), but particularly poor outcomes were obtained for this target with most protocols. SRC was another particularly challenging target for the protocols, with the highest overall enrichment being given by Vina SF (AUC of 0.69) at the cost of poor early enrichment of actives. Finally, the AD4 + DrugScore^{CSD} protocol offered the best overall enrichment for the VEGFr2 target, but with very poor performance at early stages of VS. In contrast and as evidenced in **Figure 6.6**, the Vina + DrugScore^{CSD} protocol showed exceptional performance at 0.5, 1 and 2% of false positive rates (ROCE of, respectively, 123.0, 32.8 and 12.7), which suggests that the combined use of both protocols should be considered.

3.4.3. Serine Proteases

Amongst the serine protease targets, FXa and thrombin are interesting cases where the use of arithmetic weighting offered different perspectives of VS performance to the non-weighted approach. The 146 FXa actives in the DUD cluster into just 19 dissimilar groups, which is suggestive of significant chemical analogy within the set.

Equally, from the 68 thrombin actives only 14 clusters of diverse chemotypes are formed. Interestingly, while only reasonable overall performance is revealed by the ROC AUC of 0.70 for FXa (Vina + DSX PCS protocol), the awAUC of 0.76 obtained with the Vina + DSX PAS protocol reflects a good ability to discriminate diverse FXa active chemotypes (see **Figure 6.8** and **Table 6.1**). In contrast, the arithmetic weighted metric discloses worse performance in terms of early enrichment. The Vina + DSX protocol offered the most consistent early enrichments, yielding equal ROCE and awROCE values (18.7, 7.9 and 5.1 respectively at 0.5, 1 and 2% FPR). Larger differences between the two metrics were observed for thrombin, both in terms of overall enrichment (ROC AUC of 0.72 versus a ROC

awAUC of 0.55, with Vina SF) and early recovery rates (ROCE values of 20.3, 14.2 and 13.3 at 0.5, 1 and 2% FPR versus awROCE values of 0, 2.3 and 2.2, respectively, with Vina SF). Therefore, it is clear that caution should be taken while choosing the most appropriate docking and scoring protocols for VS with serine proteases, in order to avoid misleading interpretations due to analogue bias in benchmark sets.

Table 6.4. Best performing docking and scoring protocols for the serine proteases in the DUD.

Target	Accuracy at pose prediction AD4 vs. Vina (RMSD in Å)	Best performing screening protocol (docking + scoring)	Overall Enrichment (ROC AUC)	Early Enrichment (ROCE)			
				0.5%	1%	2%	5%
FXa	2.52 / 1.56	Vina + DSX PCS	0.70	1.8	1.6	1.1	2.5
Thrombin	4.18 / 5.93	Vina SF	0.72	20.3	14.2	13.3	5.8
Trypsin	4.28 / 0.53	AD4 FEF	0.71	22.7	12.9	8.5	5.7

3.4.4. Metalloenzymes

Proteins containing metal ions and/or prosthetic groups represent difficult cases to handle by computational methods. In time, and as more experimental information comes to light, the integration of adequate parameters for metals and other exotic species in molecular mechanics-based force fields is becoming more common and efficient. As shown in **Figure 6.8**, in general the metalloenzymes yielded poor enrichment profiles. Re-scoring with DrugScore^{CSD} and the new DSX proved useful for this subset of the DUD (see **Table 6.5**).

The angiotensin-converting enzyme (ACE) was a particularly difficult case where little distinction between actives and decoys was achieved by most docking and scoring protocols. In fact, the difficulties with this metalloenzyme arised immediately at ligand pose prediction. As seen in section 3.1, neither AutoDock 4 nor Vina were able to identify correct conformations for the cognate ligand included in the DUD using default settings. The AD4 +DSX PCS protocol reached a ROC AUC of 0.64, a value that is above the line of no discrimination. This number is slightly deceiving due to the analogue bias problem. Using the arithmetic procedure, thus taking the 18 clusters of the ACE subset into consideration, a poor selection of structurally diverse actives is verfied (ROC awAUC of 0.60). On the other hand, the best early enrichment profile was given by the Vina + DrugScore^{CSD} protocol, with ROCE values of 20.3, 14.6 and 10.2 at 0.5, 1 and 2% FPR, and awROCE values of 21.7, 15.4 and 8.1 at the same levels and respectively, suggesting that a combination of protocols should be considered.

The AD4 + DrugScore^{CSD} protocol offered the best overall enrichment for ADA, while poor early enrichment was verified for most docking protocols. An interesting exception was the Vina + DSX PAS protocol, where ROCE values of 43.0, 21.7 and 7.6 were attained, respectively at 0.5, 1 and 2% of the ranked decoys.

Finally, the best overall enrichment for PDE5 was obtained with the AD4 + DSX protocol (AUC of 0.67), but best early enrichment was given by Vina + DSX protocol (ROCE of 19.3, 16.7 and 15.4 at 0.5, 1 and 2% FPR). Globally, these results can be regarded as satisfactory, but the superior discriminative power of the Vina + DSX protocol towards the diverse active chemotypes of PDE5 in the DUD must be highlighted as well. This protocol yielded an awAUC of 0.76 and awROCE values of 38.3, 21.9 and 16.8 at 0.5, 1 and 2% false positive rates, which are promising results.

Table 6.5. Best performing docking and scoring protocols for the metalloenzymes in the DUD. Cases of high accuracy (RMSD ≤ 2.0 Å) are highlighted in grey shade.

Target	Accuracy at pose prediction: AD4 vs. Vina (RMSD in Å)	Best performing screening protocol (docking + scoring)	Overall Enrichment (ROC AUC)	Early Enrichment (ROCE)			
				0.5%	1%	2%	5%
ACE	16.98 / 8.92	AD4 + DSX-PCS	0.64	0.0	2.2	3.2	2.5
ADA	2.07 / 7.03	AD4 + DrugScore^{CSD}	0.66	0.0	0.0	4.8	2.8
COMT	18.65 / 10.19	AD4 + DSX-PAS	0.58	45.0	45.3	11.3	6.2
PDE5	1.56 / 0.79	AD4 + DSX	0.67	13.0	9.8	5.8	4.4

3.4.5. Folate Enzymes

The folate enzymes family consists of just two members, but it is worth noting it was a group of targets for which most protocols produced exceptional results – exceptions were Vina SF and protocols employing DSX-PCS and DrugScore^{CSD} (re-scoring with DrugScore^{CSD} was not possible for folate enzymes due to unhandling of their cofactors, which participate in ligand binding). Both DHFR and GART obtained good results at cognate ligand pose prediction, early and overall enrichment with most of the docking and scoring protocols under evaluation (see **Figure 6.4** and **Table 6.6**). With a number of 201 known actives grouped into 14 diverse clusters, DHFR represented a good challenge for the protocols and the results were encouraging: the Vina + DSX protocol offered an awAUC of 0.76 along with awROCE values of 24.7, 19.8 and 14.2 at 0.5, 1 and 2% FPR, which are all indicative of good discriminative performance.

Table 6.6. Best performing docking and scoring protocols for the folate enzymes in the DUD. Cases of high accuracy (RMSD ≤ 2.0 Å) and/or performance (awAUC ≥ 0.75) are highlighted in grey shade and bold, respectively.

Target	Accuracy at pose prediction AD4 vs. Vina (RMSD in Å)	Best performing screening protocol (docking + scoring)	Overall Enrichment (ROC AUC)	Early Enrichment (ROCE)			
				0.5%	1%	2%	5%
DHFR	2.00 / 0.47	**Vina + DSX**	**0.76**	**24.7**	**19.8**	**14.2**	**6.8**
GART	2.05 / 9.49	**AD4 FEF**	**0.80**	**15.7**	**5.2**	**9.5**	**5.3**

3.4.6. Other Enzymes

Despite lacking the functional metals present in the metalloenzymes, in general this group of targets was also associated with weak enrichments for most docking and scoring protocols (see **Table 6.11**). In fact, this is by far the largest group in the DUD data set and, out of its 14 targets, only 3 yielded good or exceptional VS performances (HIVPR, PARP and PNP). Overall, the AD4 + DSX-PCS protocol offered the best results across the entire subset (mean ROC AUC of 0.59).

Even though the Vina SF offered the best overall enrichment for ALR2 (ROC AUC of 0.69), the AD4 + DSX protocol should be considered as complementary because of its exceptional ability to discriminate diverse actives at early stages of virtual screening (awROCE values of 153.0, 48.0, 16.4 and 8.4 respectively at 0.5, 1, 2 and 5 FPR). Moreover, AutoDock 4 outperformed Vina at predicting the best pose for the ALR2 X-ray ligand provided in the DUD.

COX1 is one of a set of cases where excellent docking accuracy was translated into poor/weak VS performance. While both AutoDock 4 and AutoDock Vina excelled at predicting the native X-ray pose of the cognate ligand, all docking and scoring protocols performed no better than random picks at discriminating COX1 actives from decoys. An interesting exception was found by re-scoring AD4 predictions with DSX and re-ranking results by per-contact score (ROC AUC of 0.61). In addition, this protocol yielded particularly good early enrichments (ROC values of 40.0, 32.0 and 10.7 at 0.5, 1 and 2% of the ranked decoys).

HIV reverse transcriptase is another example where, despite great docking accuracy for both AD4 and Vina, poor VS performance was obtained. The built-in FEF of AutoDock 4 offered the best overall enrichment (ROC AUC of 0.59), but very poor early enrichment. The Vina + DSX protocol, by contrast, provided reasonable enrichment in actives at early stages of VS (ROCE values of 37.5, 20.8 and 9.4 at 0.5, 1 and 2% FPR, respectively).

A final case where excellent pose predictions did not result in good overall VS performance was that of *Mycobacterium tuberculosis* enoyl-acyl carrier protein reductase (InhA). The best performing protocol found for this target, AD4 + DSX-PCS, did not go beyond an ROC AUC of 0.61, disclosing limited discrimination of actives over decoys overall. However, reasonable discrimination occurred at early intervals of the ranked data set, with ROCE values of 23.3, 19.5 and 10.9 at 0.5, 1 and 2% FPR. Another interesting finding was that even though Vina performed worst in terms of overall VS performance, it presented exceptional early enrichment (ROCE of 129.0, 42.8 and 20.8 at 0.5, 1 and 2% FPR).

Table 6.7. Best performing docking and scoring protocols for other enzymes in the DUD. Cases of high accuracy (RMSD ≤ 2.0 Å) and/or performance (awAUC ≥ 0.75) are highlighted in grey shade and bold, respectively.

Target	Accuracy at pose prediction: AD4 vs. Vina (RMSD in Å)	Best performing screening protocol (docking + scoring)	Overall Enrichment (ROC AUC)	Early Enrichment (ROCE)			
				0.5%	1%	2%	5%
AChE	5.05 / 0.64	Vina SF	0.67	3.5	5.9	3.7	4.5
ALR2	1.33 / 3.44	Vina SF	0.69	9.5	4.8	4.4	7.9
AmpC	12.90 / 3.73	AD4 + DSX-PCS	0.53	11.8	5.2	5.6	4.1
COX1	0.99 / 0.28	AD4 + DSX-PCS	0.61	40.0	32.0	10.7	3.4
COX2	20.96 / 21.71	Vina + DSX-PCS	0.59	3.5	2.8	2.4	1.9
GPB	0.62 / 0.70	Vina + DSX-PCS	0.72	0.0	0.0	1.0	2.0
HIVPR	11.16 / 2.20	**Vina + DSX-PCS**	**0.80**	**56.5**	**47.0**	**29.1**	**13.2**
HIVRT	0.84 / 0.73	AD4 FEF	0.59	0.0	2.8	2.6	1.0
HMGR	2.04 / 1.72	Vina + DSX-PCS	0.71	114.0	50.0	16.3	7.5
InhA	0.44 / 0.22	AD4 + DSX-PCS	0.61	23.3	19.5	10.9	5.6
NA	23.73 / 0.41	AD4 FEF	0.73	20.3	14.5	22.0	10.2
PARP	2.37 / 0.93	**AD4 FEF**	**0.83**	**max.**	**303.0**	**64.9**	**13.0**
PNP[a]	15.25 / 20.63	**AD4 + DSX-PCS**	**0.79**	**160.0**	**50.0**	**29.1**	**14.4**
SAHH[a]	0.36 / 20.13	AD4 + DSX-PAS	0.61	0.0	3.2	3.2	2.4

[a] These results were based on AD4 scores assigned to the lowest-energy conformation of the largest cluster of ligand poses, while the other results were calculated as in previous tables using the lowest-energy poses.

To contrast the former examples, neuraminidase (NA) was another peculiar case where, even though AutoDock 4 failed at predicting the cognate ligand's native pose, AD4 FEF outperformed all other protocols in terms of VS performance. In fact, all other protocols, without exception, perform worse than random against NA. Thus, because NA represents yet another example of an unexpected outcome of a VS experiment, it is analysed in depth in the next subsection.

3.5. In-depth analysis of *unexpected* cases

Throughout the analysis of the results of our VS experiments, we encountered a few cases where reasonable or good discrimination of active molecules over inactive ones did not seem to depend on the ability to discriminate correct ligand poses within the target's binding site. Simply put, these were cases where the VS performance of the employed protocol was not linked to the docking accuracy (or pose fidelity) of the underlying docking algorithm. As mentioned before, this is an unexpected (and worrisome) outcome of a docking-based VS experiment, because the ability of one given VS protocol to discriminate actives over inactives should be intimately linked to its ability to identify the most favourable interactions between the active molecules and target's active/binding site. We selected four of the most important examples to carry out a more exhaustive analysis of ligand and receptor properties that could be linked to such unexpected behaviour; namely:

- peroxisome proliferator activated receptor γ (PPARg) – with an RMSD of 9.54 Å for the lowest-energy (E) docked pose of the cognate X-ray ligand (generated by Vina) and a ROC AUC of 0.83 with the Vina + DrugScore^{CSD} protocol;
- retinoic X receptor α (RXRa) – with an RMSD of 15.49 Å for the lowest-E docked pose of the cognate X-ray ligand (by AD4) and a ROC AUC of 0.86 with AD4 FEF;
- neuraminidase (NA) – with an RMSD of 23.73 Å for the lowest-E docked pose of the cognate X-ray ligand (by AD4) and a ROC AUC of 0.73 with AD4 FEF;
- purine nucleoside phosphorylase (PNP) – with an RMSD of 15.25 Å for the lowest-E docked pose of the cognate X-ray ligand (by AD4) and a ROC AUC of 0.79 with the AD4 +DSX PCS protocol.

To contrast these examples with cases we believe fall in the category of *expected* outcomes of VS, we selected four examples where the good VS performance (ROC AUC ≥ 0.75) provided by a VS protocol followed from good docking accuracy (RMSD < 2.5 Å) provided by the respective docking program: androgen receptor (AR), estrogen receptor agonist (ER agonist), mineralocorticoid receptor (MR) and glycinamide ribonucleotide transformylase (GART).

Table 6.8 reports the number of heavy atoms, providing a depiction of the size of the compounds, and the number of rotatable bonds, providing a picture of the compounds' flexibility, for the cognate (X-ray) ligand in complex with the target receptor, and for the active and decoy compounds in the respective data set (median values). For each group of cases described above, overall median values are given to ease the direct comparison of the properties between the two groups.

Table 6.8. Ligand-centric properties of compounds in the DUD data set associated with both *unexpected* and *expected* VS behaviour.

Target	PDB id	Heavy atom count			Rotatable bond count		
		Cognate ligand	Actives (median)	Decoys (median)	Cognate	Actives (median)	Decoys (median)
Poor docking accuracy + Good VS performance							
PPARg	1fm9	41	41	37	12	12	9
RXRa	1mvc	28	28	28	3	5	7
NA	1a4g	23	21	22	6	6	5
PNP	1b8o	19	19	20	2	3	3
Median		**25.5**	**24.5**	**25**	**4.5**	**5.5**	**6**
Good docking accuracy + Good VS performance							
AR	1z95	21	22	21	0	1	3
ER agonist	1l2i	24	21	20	2	1	3
MR	2aa2	26	26	25	3	2	3
GART	1c2t	34	31	34	10	9	9
Median		**25**	**24**	**23**	**2.5**	**1.5**	**3**

From a ligand-centric viewpoint, the only significant difference that could be detected between the two groups of case studies was in the number of rotatable bonds of the cognate ligand, of the actives and of the decoy compounds. The most pronounced difference found was in the number rotatable bonds of the actives, with a median value of 5.5 for the first group (the *unexpected* cases) versus a median value of 1.5 for the second group (the *expected* cases). Ligand flexibility is a well-known challenge posed to docking algorithms: typically, difficulties increase with the number of flexible bonds in the compounds to be docked. In principle, the higher flexibility of the cognate ligands of the first group could justify the higher RMSD values obtained during the re-docking procedures. Indeed, by looking at the lowest-energy docked poses generated for the cognate ligand of PPARg (12 rotatable bonds) by AD4 and Vina, one can realise that both programs are able to find the main binding site with precision, yet suggesting a reverse binding mode for the ligand (recall **Figure 6.2**). In both cases, this results in RMSD values above 9 Å. However, it can be argued that high flexibility is also found among the respective benchmarking set (actives and decoys), suggesting that similar difficulties were encountered by the docking programs while docking the benchmarking compounds into the binding site of PPARg. Moreover, looking at other examples in the first group individually, one can identify compounds holding low flexibility (e.g. PNP), even as low as that of the compounds comprised in the second group. These counter-examples show that ligand flexibility alone cannot explain all cases showing unexpected behaviour.

We went further to collect a number of receptor-related characteristics that could help us shading light over this subject. **Table 6.9** reviews features like the X-ray resolution at

which each structure was determined, the size of the docking simulation box, and summarizes the results of a set of analyses performed using two alternative binding site prediction algorithms, Q-SiteFinder [260] and DoGSite [659]. These algorithms take both geometric and energetic criteria into consideration (based on the grid approach explained in subsection 1.4.4 of the introductory chapter) to detect and characterise potential binding sites on protein structures. Besides analysing the predicted binding sites (and subpockets) visually, we looked at measures like the volume and the depth of the predicted sites.

Table 6.9. Collection of receptor-related features retrieved from the DUD data set or determined by molecular modelling techniques for DUD examples associated with both unexpected and expected VS behaviour.

Target	X-ray resolution (Å)	Docking simulation box size (Å^3)	Q-SiteFinder analysis		DoGSite analysis	
			Pocket volume (Å^3)	Precision[(a)] (%)	Pocket volume (Å^3)	Depth[(b)] (Å)
Poor docking accuracy + Good VS performance						
PPARg	2.10	83333	804	66	1321	8.75
RXRa	1.90	82288	561	88.1	944	24.43
NA	2.20	85077	196	71.2	184	13.2
PNP	1.50	80190	476	73.7	645	19.24
Median	**2.00**	**82811**	**518.5**	**72.5**	**795**	**16.22**
Good docking accuracy + Good VS performance						
AR	1.80	82240	556	91.5	632	2.33
ER agonist	1.95	82288	485	87.3	509	1.65
MR	1.95	82924	472	94.4	497	0.89
GART	2.10	87814	589	79.4	505	16.33
Median	**1.95**	**82606**	**520.5**	**89.4**	**821**	**1.99**

[(a)] In Q-SiteFinder, *precision* is a measurement of how well the predicted binding site maps onto the ligand coordinates. It is defined as the percentage of (methyl) probe sites (as the ones explained in Chapters 2 and 3) defined by a single cluster of probes that are within 1.6 Å of an atom of a particular ligand When a single ligand gives a success in two separate sites, only the higher ranking site is counted, since they are part of the same binding site. [(b)] In DoGSite, *depth* corresponds to the maximal distance between a grid point located within the solvent exposed shell and the deepest grid point of the buried shell. Thus, the depth is described by the maximal separation between the solvent and the buried part of the binding site.

Q-SiteFinder and DoGSite provided reasonably distinct results in terms of pocket volumes, with the former presenting more conservative (and precise) volumes around the structure of the bound ligands; binding site volumes predicted by DoGSite were, in general, larger. However, the two algorithms did agree in the comparison of the two groups of cases under study, revealing little or no differences between the two. Interestingly, according to the DoGSite analysis, all binding sites belonging to targets yielding unexpected VS outcomes are deeper (i.e. more buried), whereas most binding sites belonging to targets yielding expected VS outcomes are more superficial. The only exception in the

second group is GART, whose main binding site has a predicted depth of 16.33 Å, a value that is comparable to the median binding site depth of the first group (16.22 Å).

These results led us to hypothesise that, besides ligand flexibility, the high depth of the main binding sites could represent an additional hurdle (and source of bias) to the docking algorithms, driving the docking of cognate ligands and benchmarking set compounds to alternative (and more accessible) receptor sites (see **Figure 6.9**).

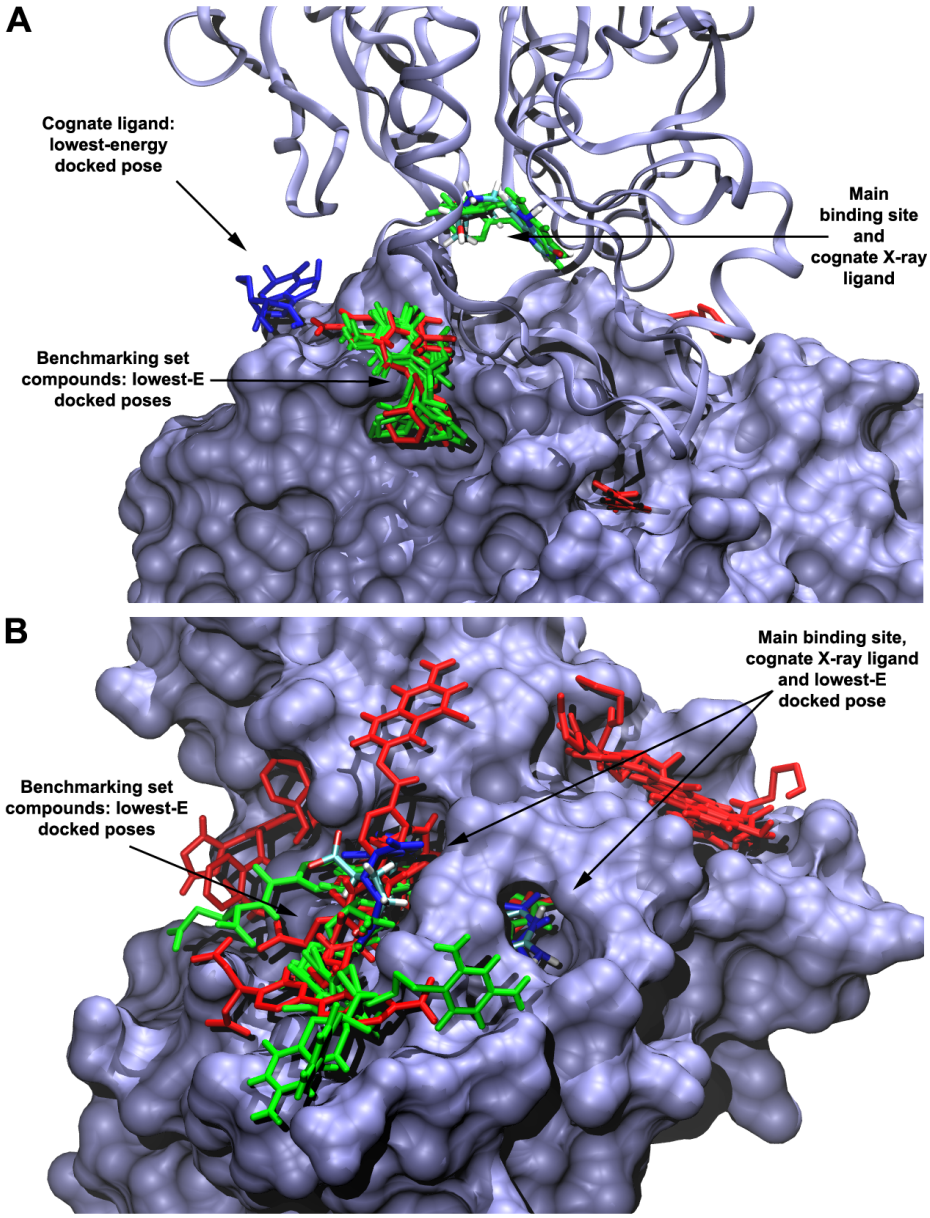


Figure 6.9. Representation of the X-ray structure of two DUD targets providing biased VS outcomes. Shown are the native (X-ray) pose of the respective cognate ligand (coloured by element) and the lowest-energy poses generated by docking for that same cognate ligand (blue) and for randomly selected examples of actives (green) and decoys (red). (A) pur-

ine nucleoside phosphorylase (PNP) and (B) glycinamide ribonucleotide transformylase (GART).

Figure 6.9 provides a visual depiction of the docking results obtained for two case studies, one belonging to the first group (*unexpected* VS outcomes) and the other belonging to the second group (*expected* VS outcomes). **Figure 6.9A** illustrates the results for target PNP, where it can be seen that AutoDock Vina docks the cognate ligand and most benchmarking set compounds into alternative and more superficial binding sites. Considering that, in principle, the decoy molecules lack the chemical topology required for interaction with the main binding site, this result might be expected for *inactive* molecules. However, this principle does not hold for molecules that are known to bind to the main binding site of the target in order to produce its activity (i.e. the active compounds). This picture clearly shows that, whatever discrimination between PNP actives and decoys may have occurred, it took place away from the main binding site of PNP.

Equally, most of the benchmarking set compounds of GART (shown by random choice in **Figure 6.9B**) are docked into alternative and superficial receptor sites. To contrast with the previous case, however, AD4 correctly binds the cognate ligand into the main binding site of the protein (RMSD of 2.05 Å). Still, these results suggest that the good overall VS performance provided by AD4 FEF for GART may in fact derive not from a good docking accuracy but from the scoring of ligand conformations docked away from the main binding site of the target. Furthermore, these results confirm that the DUD data set does not provide enough experimental information to allow for meaningful evaluations of docking accuracy. A reasonable number of crystal structures of ligand-bound conformations of each target would be required in order to conduct the appropriate re-docking studies and thus collect enough statistical evidence on the pose fidelity of the algorithms.

In order to quantify the problem of ligand docking into alternative receptor sites, we computed the Euclidean distance between the centre of mass of all lowest-energy docked poses and the centre of mass of the cognate X-ray ligand (determined from the experimental coordinates provided in the DUD). **Table 6.10** reports the median values for the centre-of-mass distances between the active and decoys compounds of each benchmarking set under evaluation and its respective cognate X-ray ligand. Globally, the values of the first group are higher than those of the second group (both for the actives and the decoys). However, inspecting each case individually, it can be found that the centre-of-mass distances for targets RXRa and NA (belonging to the first group) are comparable to the global median value of the second group. This suggests that, even though associated with high RMSDs (and thus low docking accuracy), the good VS performances provided by AD4 FEF for these two targets result from the scoring of active compounds correctly docked into the main binding site of the respective target. Conversely, the table shows

that most actives compounds of GART are docked into alternative receptor sites, and that the good VS performance provided by AD4 FEF results from the scoring of wrong ligand conformations.

Table 6.10. Median Euclidean distances computed between the centre-of-mass of the benchmarking set compounds and the centre-of-mass of the respective cognate X-ray ligand of the selected DUD targets.

Target	Centre-of-mass distances between docked poses and cognate X-ray ligand (Å)	
	Actives (median)	Decoys (median)
Poor docking accuracy + Good VS performance		
PPARg	6.22	8.70
RXRa	1.15	16.14
NA	1.36	16.70
PNP	12.53	12.69
Median	**3.79**	**14.41**
Good docking accuracy + Good VS performance		
AR	1.13	11.92
ER agonist	0.76	7.73
MR	1.08	14.93
GART	10.74	6.51
Median	**1.10**	**9.82**

The results presented in this subsection clearly indicate that the unexpected VS outcomes obtained for a number targets in the DUD data set result either from a misleading evaluation of docking accuracy based on little evidence (particularly when the only cognate X-ray ligand provided with a given benchmarking set is fairly flexible) or from the favourable scoring of active compounds in alternative (and more accessible) receptor binding sites, thus providing an artificial notion of enrichment.

3.6. Virtual screening performance across the entire DUD data set

In this section, we provide an analysis of the virtual screening performance of the docking and scoring protocols over the entire DUD data set. Despite the efforts that went into the creation of the DUD, it retains important gaps. These include the dominance of enzymes among the target classes. Of the 40 targets in the data set, 30 are enzymes. The inclusion of additional (and more diverse) binding sites to the DUD is an obvious way to cover more "target space". A comprehensive coverage of every possible binding site and pharmaceutical target classes is, however, unworkable. Such a benchmark would be un-

wieldy and unrealistic. First, no crystal structure is available for perhaps half of all known drug targets. Second, for most ligand binding sites in the PDB, few if any ligands are known that could be used as controls. With the above in mind, while the analysis presented in this section may be regarded as an expected exercise at this point, great care should be taken in order to prevent misleading extrapolations to be made and misplaced conclusions to be taken.

Table 6.11 reports the mean ROC AUC and mean ROCEs across the DUD benchmarking sets. The statistics were based on the 40 DUD targets except for protocols employing DrugScore^{CSD}, with which only 30 targets could be re-scored. This was due to the limited handling of proteins containing certain metals and/or prosthetic groups by DrugScore^{CSD}. Since a significant variance was found for the mean ROC AUC of these relatively small-sized samples, 95% confidence intervals were determined using a re-sampling (bootstrapping) method. In general, the bootstrap-based confidence intervals were narrower than parametrically estimated confidence intervals, thus facilitating the comparison of the mean ROC AUC values.

Table 6.11. Statistical results for the docking and scoring protocols using the DUD data sets, including mean AUCs and median/mean ROCEs.

Docking and scoring protocol		Mean ROC AUC[d]	95% CI[e]	Median/mean ROCE[f]			
				0.5%	1%	2%	5%
AD4 FEF[a]	LE	**0.58**	**0.52–0.61**	**4.9 / 13.2**	**4.4 / 15.6**	**4.3 / 8.5**	**3.5 / 4.4**
	LC	**0.57**	**0.53–0.61**	2.3 / 14.9	3.5 / 16.7	3.2 / 7.6	2.0 / 3.8
AD4 + DrugScore^{CSD} [b]		0.55	0.50–0.59	4.5 / 9.7	3.1 / 16.2	2.6 / 5.6	3.0 / 7.8
AD4 + DSX[c]	total	0.53	0.49–0.57	**4.9 / 13.8**	3.1 / 8.0	2.5 / 5.1	2.3 / 3.1
	PAS	0.54	0.50–0.57	**4.8 / 15.8**	3.4 / 11.4	2.4 / 4.5	2.1 / 3.1
	PCS	**0.59**	**0.56–0.62**	3.5 / 15.7	3.2 / 9.7	2.3 / 5.9	2.0 / 3.0
Vina SF		**0.57**	**0.53–0.61**	**6.5 / 22.89**	**4.7 / 21.5**	3.4 / 8.0	2.9 / 3.9
Vina + DrugScore^{CSD} [b]		0.54	0.50-0.59	**17.0 / 38.7**	**8.0 / 14.9**	**5.6 / 9.7**	**4.5 / 8.8**
Vina + DSX[c]	total	0.54	0.51–0.58	**16.2 / 26.5**	**8.8 / 16.1**	**6.7 / 9.1**	**3.8 / 4.3**
	PAS	0.56	0.52–0.60	2.1 / 16.8	3.4 / 9.3	2.6 / 6.1	2.0 / 3.2
	PCS	0.54	0.50-0.58	0.0 / 10.2	1.0 / 5.8	1.0 / 3.8	1.4 / 2.3

[a] For AutoDock 4 free energy function, both ranking by lowest energy conformation scores (LE) and by largest cluster conformation scores (LC) are shown. [b] Results for DrugScore^{CSD} were based on 30 DUD targets, due to failure in handling metals/co-factors. [c] For DSX, besides the total score, ranking by per-atom scores (PAS) and per-contact scores (PCS) are reported. [d] A mean AUC of 0.5 represents random performance. [e] Bootstrap-based 95% confidence intervals. [f] Median and mean ROC Enrichments (ROCE) were determined for four early false positive rates.

Pearson correlation coefficients and *p*-values for the comparisons between mean AUCs are shown in **Table 6.12**. These correlations only provide a measure of whether two docking and scoring protocols are likely to offer similar performance against a certain target, not whether the same active molecules are likely to be recovered. As expected, correla-

tion was small, in general, arising mostly amongst protocols employing the same docking engine. For example, while weak correlation is found between AD4 FEF and Vina's Scoring Function (0.52), good correlation is found between AD4 + DrugScore^{CSD} and Vina + DrugScore^{CSD} (0.76), and even higher between AD4 + DSX and Vina + DSX (0.83). This result indicates that the overall performance of the protocols is more influenced by the performance of the scoring functions than the performance of docking itself.

Furthermore, limited statistical significance was found for the difference of the means across the entire DUD data set. While all mean AUC values were above 0.5, meaning that all tested docking and scoring protocols tended to select active molecules over decoys in a non-random way across the entire set of targets, this result may also suggest that no particular protocol significantly outperforms another across a wide range of different targets. It can be denoted, nonetheless, that AutoDock Vina and its native scoring function provide the most balanced results in terms of overall performance (mean AUC of 0.57) and early enrichment at 0.5 and 1% false positive rates (median ROCE of 6.5 and 4.7, respectively). Still, as off-putting as these results may seem, they emphasize the importance of a careful and individual evaluation of customized (or *tailor-made*) protocols for a given target. They also disclose the importance of gathering enough statistical evidence before drawing any broad conclusions on the applicability of one particular protocol to multiple targets.

Table 6.12. Global performance comparison between pairs of VS protocols, evaluated by the mean ROC AUCs for the DUD data set [a].

		AD4 FEF		**AD4 + DrugScore CSD**	**AD4 + DSX**			**Vina SF**	**Vina + DrugScore CSD**	**Vina + DSX**		
		LE	LC		total	PAS	PCS			total	PAS	PCS
AD4 FEF	LE		**0.95**	0.52	0.62	0.58	0.54	0.52	0.29	0.45	0.41	0.23
	LC	0.026		0.62	0.66	0.61	0.59	0.61	0.39	0.50	0.43	0.24
AD4 + DrugScore^{CSD}		0.100	0.209		0.62	0.50	0.68	0.58	0.76	0.56	0.37	0.48
AD4 + DSX	total	**0.011**	0.054	0.266		**0.86**	0.72	0.68	0.42	**0.83**	0.63	0.46
	PAS	**0.029**	0.128	0.289	0.537		0.73	0.47	0.32	0.67	0.76	0.43
	PCS	0.795	0.223	0.212	**1.72E-4**	**3.77E-4**		0.60	0.32	0.60	0.60	0.52
Vina SF		0.545	0.845	0.147	**0.020**	0.118	0.306		0.58	0.75	0.46	0.20
Vina + DrugScore^{CSD}		0.148	0.267	0.854	0.500	0.491	0.335	0.121		0.56	0.45	0.51
Vina + DSX	total	0.104	0.318	0.825	0.225	0.627	**0.009**	0.080	0.995		0.78	0.43
	PAS	0.343	0.799	0.874	**0.087**	**0.083**	0.080	0.647	0.972	0.219		0.45
	PCS	0.124	0.318	0.084	0.693	0.926	**0.011**	0.243	0.248	0.791	0.313	

[a] The upper triangle comprises the Pearson correlation coefficients and the lower triangle comprises the *p*-values for paired *t*-test. High Pearson correlations (>0.80) are highlighted in bold. Statistically significant (95%) *p*-values are also in bold.

Generally, re-scoring procedures did not produce improvement to the overall performance of AutoDock built-in scoring functions. The only exception was the per-contact scores (PCS) of DSX in the re-rank AutoDock 4 docking results. In fact, the AD4 + DSX-PCS protocol reaches the highest mean ROC AUC across the entire DUD data set (0.59). In contrast, re-scoring with both DrugScore^{CSD} and DSX produced significant increases in early recovery rates of Vina docking results at all levels of false positive rate reported (median ROCE of 17.0, 8.0, 5.6 and 4.5 for DrugScore^{CSD}, respectively at 0.5, 1, 2 and 5% FPR, and 16.2, 8.8, 6.7 and 3.8 for DSX). Taken together, these results highlight the critical importance of combining multiple docking and scoring protocols in the pursuit of a good balance between overall and early enrichment, and thus to achieve consensus towards the best possible outcomes of docking-based virtual screening.

4. Conclusions

In this chapter we tested the application of several docking and scoring protocols to a set of pharmaceutically relevant targets. Even though the detailed benchmark of docking and scoring solutions against all target systems of interest falls outside the scope of this project, we regard the testing of our docking-based VS protocols against the diversity of targets deposited in the DUD as a neat academic exercise to support the efforts of protocol porting onto a large volunteer computing platform. Taken globally, the results of our analyses show that the protocols are reasonably effective, but highly inconsistent, tending to perform well on a given system and weakly on the next.

We found no direct association between docking accuracy (or pose fidelity) and performance (or enrichment) in virtual screening. In fact, our results can be grouped in four possible scenarios (or outcomes): (*i*) the correct prediction of the cognate ligand binding mode was followed by acceptable or good enrichment rates; (*ii*) the cognate ligand was correctly docked, yet the VS performance was poor; (*iii*) the cognate ligand was poorly docked and the enrichment rates were low; and (*iv*) an acceptable or even good VS performance was not linked to the correct prediction of the cognate ligand binding mode. While outcomes (*i)* and (*iii*) are somewhat expectable, outcome (*ii*) illustrates failures of scoring functions to discriminate active from inactive molecules, despite correct ligand poses being found – a frustrating scenario from the viewpoint of structure-based virtual screening (SBVS), which reflects the limited ability to correctly predict binding affinites of docking scoring functions. Cases falling under outcome (*iv*) are no less inconvenient

and particularly hard to explain. The statistical evidence on the docking accuracy of AD4 and Vina against the DUD targets is not as high as one would like. Indeed, it would require additional structural information from multiple target-ligand complexes and exhaustive re-docking studies in order to draw more rigorous conclusions. Such discrepancies between docking and ranking accuracies have been previously witnessed and reported [89,419]. The five targets falling under outcome (*iv*) belong to three different families and are largely unrelated. Indeed, our results show that these unexpected outcomes are linked to a misleading evaluation of docking accuracy based on little evidence, especially when the structures of cognate ligands are scarce and/or very flexible, or from the favourable scoring of active compounds in alternative (and more accessible) receptor binding sites. In fact, we believe that the latter scenario represents an additional source of artificial enrichment, which is clearly linked to the limitations of scoring functions in current use, and propose the exploitation of complementary measures (such as centre-of-mass distances between the docked benchmarking set compounds and the respective cognate X-ray ligands) when evaluating the performance of docking-based VS protocols.

Vina and its native scoring function provided the best overall docking accuracy against a multitude of targets. While, in general, the use of alternative scoring functions was proven unprofitable towards the discrimination of the top ligand poses generated by AD4, the use of protocols based on re-scoring of the predicted complexes with either DrugScore^{CSD} or DSX offered important improvements in terms of the ability to select active molecules of certain targets. This was particularly visible/relevant in the case of enzyme targets. It is worth emphasising that, in general, all tested docking and scoring protocols were able to recognize actives better than random selection. Hence, in this chapter, a best-performing docking and scoring screening protocol was highlighted for each target of pharmaceutical interest in the DUD. Nevertheless, it is also worth noting that more elaborate scoring functions or ranking approaches should be pursuit for particularly hard-to-handle targets: the androgen receptor (AR), adenosine deaminase (ADA), the cyclooxygenase enzymes (COX-1 and COX-2), the enoyl ACP reductase (InhA), the P38 mitogen-activated protein kinase, the human heat shock protein 90 kinase (HSP90) or the platelet derived growth factor receptor kinase (PDGFrb). For example, the use of machine learning methods (such as support vector machines, SVMs) trained by associating sets of individual energy terms retrieved from docking scoring functions with the known binding affinity of each compound, has shown very promising results in the case of *Mycobacterium tuberculosis* InhA and other DUD targets [660].

Despite no significant statistical differences were found in this study to support the idea of one particular VS protocol outperforming all others, Vina and its native scoring function provided the most balanced results in terms of both docking accuracy and VS performance against the set of targets deposited in the DUD.

Chapter 7

Virtual high-throughput screening and the discovery of new TTR amyloid inhibitors: from hits to leads

"It's not how hard you hit. It's how hard you get hit…
and keep moving forward…"

[Randy Pausch, in *The Last Lecture*]

1. Introduction

Given the undesirable features associated with many of the TTR amyloid inhibitors identified to date, the screening of new lead compounds with diverse properties is a critical goal. At this stage, we are particularly interested in identifying potential inhibitors with adequate solubility profiles, low propensity to aggregate in solution and reduced halide fraction. In this chapter, we made use of the knowledge collected throughout this PhD project and engaged in an unprecedented quest to discover novel, functional and safe TTR amyloid inhibitors.

As disclosed throughout the previous chapters of this thesis, the use of virtual screening (VS) has clear advantages over high-throughput screening (HTS). Indeed, VS can access far more chemistry, much faster, and at a much lower cost. The top scoring compounds from VS can then be tested rapidly, and although only a few of them are likely to actually bind the target, even a couple of novel hits can be extremely useful. However, while certain VS techniques, such as docking and scoring, may be useful for the VS of a set of compounds against a single target, large scale campaigns against multiple targets are often hindered by their high computational complexity or lack of practical methodologies to accurately estimate binding affinities.

Putting into perspective some of the ideas presented in the introductory chapter on the importance of drug discovery in the academia, we have been compelled to explore a platform that could foster the involvement of academic researchers in projects to leverage drug discovery, simultaneously capturing the interest and participation of volunteer citizens and society in general. We have felt fascinated by the concept and the outcomes of volunteer computing projects such as the Folding@home project [661,662], the Rosetta@home project [663], or the FightAIDS@home project [664] implemented at the World Community Grid. Indeed, amongst several other feats, the use of large computer grids has enabled a deeper understanding of the protein folding problem, aided the design of therapeutic peptides, and allowed the study of mutational effects on protein-ligand binding. With the rapidly increasing availability of protein structures, the access to such computational resources, accompanied by the development of multi-target efficient VS protocols is becoming even more important.

In this final chapter, we present the results of a large-scale campaign comprising a total of 22 VS protocols on a tailored library of approximately 2.3 million drug-like compounds described in Chapter 4. We used the best-performing protocols against TTR (identified by grey shading across multiple tables in Chapter 5 and summarized in **Table 7.1**), from the

ones employing fast searches across the two-dimensional and three-dimensional similarity space, all the way to methods using the knowledge of the TTR receptor structure to perform complex searches in protein-ligand conformational space. We went further to implement the most computationally-demanding VS protocols on a large volunteer computing network, which are now available to academic researchers requiring high processing capabilities.

Finally, in this chapter the top-scoring compounds retrieved by each VS protocol are evaluated in terms of their molecular properties and compliance with 3D pharmacophore models devised to differentiate the distinct binding modes observed for some of the known binders on each TTR binding site. We further reveal the results of a biochemical assay designed to evaluate the amyloid aggregation inhibitory activity of the selected virtual hits. Forty-seven commercially available candidates were purchased from their respective chemical suppliers. Of these, 38 have been experimentally tested for inhibition of TTR fibril formation. Thirty-one compounds showed inhibitory activity to some extent, of which five revealed to be good inhibitors, reducing amyloid fibril formation to less than 40% *in vitro* at a protein-ligand concentration *ratio* of 1:2.

2. Materials and methods

Twenty-two virtual screening protocols presented in Chapters 5 and 6 were implemented across different hardware resources in order to carry out multiple virtual screening campaigns using a screening set of approximately 2.3 million compounds. These 22 best-performing VS protocols are described in **Table 7.1**, along with the respective type of hardware that was used to run them.

Table 7.1. Description of the 22 best-performing protocols used for large-scale VS campaigns against TTR amyloid.

Virtual Screening protocol	Ligand or Receptor template	Hardware resources
2Dsim CHF Euclidean distance	Phenox	Macintosh laptop
2Dsim CHF Euclidean distance	PCB18	Macintosh laptop
2Dsim CHF Euclidean distance	Thyroxine (T_4)	Macintosh laptop
2Dsim CHF Tanimoto index	PCB18	Macintosh laptop
2Dsim PF Tanimoto index	PCB18	Macintosh laptop
2Dsim UNITY Tanimoto index	Phenox	Linux server

Virtual Screening protocol	Ligand or Receptor template	Hardware resources
2Dsim UNITY Tanimoto index	PCB18	Linux server
ROCS Combo	Phenox	Parallel Virtual Machine
ROCS Combo	PCB18	Parallel Virtual Machine
ROCS ScaledColor	Phenox	Parallel Virtual Machine
ROCS Scaled Color	Thyroxine (T_4)	Parallel Virtual Machine
EON ETcombo	Phenox	Parallel Virtual Machine
LigMatch 2D+3D	Multiple templates	HPC cluster
2Dsim CHF Euclidean distance	DBF47-OH (concatamer)	Macintosh laptop
2Dsim CHF Euclidean distance	DBF47-COOH (concatamer)	Macintosh laptop
EON ETcombo	Phenox-OH (concatamer)	Parallel Virtual Machine
EON ETcombo	Phenox-COOH (concatamer)	Parallel Virtual Machine
FRED Chemgauss3	TTR structure 2g9k (PDB id)	Parallel Virtual Machine
FRED Chemscore	TTR structure 2g9k (PDB id)	Parallel Virtual Machine
AD4 + DrugScore^{CSD}	TTR structure 1bm7 (PDB id)	Volunteer computing platform
AD4 + DrugScore^{CSD}	TTR structure 2g5u (PDB id)	Volunteer computing platform
AD4 + DrugScore^{CSD}	TTR structure *1bm7opt*	Volunteer computing platform

2.1. Computer resources

In this subsection, emphasis is given to the setting up of hardware resources used to carry out several computational screening campaigns against TTR amyloid. Subsection 2.1.5, in particular, provides a detailed description of the strongest efforts in terms of the implementation of new resources employed throughout this project.

2.1.1. Virtual screening on a Macintosh laptop

All virtual screening runs employing ChemAxon's two-dimensional fingerprints and similarity searches were performed on a Macintosh laptop (with an Intel Core 2 Duo CPU at 2.4 GHz) running Mac OS X Version 10.5.8 (Leopard). Each run took approximately 2.5

hours to complete on a single CPU core, which illustrates not only the swiftness of 2D similarity searches but also the ever-increasing power of computer hardware.

2.1.2. Virtual screening on a Linux server

All UNITY 2D-searches were performed on a Linux server running OpenSUSE 10.1 where the SYBYL package was installed. The searches were run with for 4 CPU cores and completed in approximately 30 minutes (per VS protocol).

2.1.3. Virtual screening on a Parallel Virtual Machine

ROCS, EON and FRED calculations can be efficiently split up over computer networks taking advantage of a so-called *parallel virtual machine* (or PVM). PVM is a freely available library for running processes on more than one processor on one or more machines (the source code is freely available at:
http://www.csm.ornl.gov/pvm/pvm_home.html).
PVM jobs involve some network traffic, sending multi-conformer molecules generated by OMEGA from the master node computer to the slaves and retrieving results back to the master. The single ROCS slave jobs are relatively fast compared to the network I/O time, so for scaling beyond 32 CPUs, additional considerations need to be taken during the setting up of the jobs.

A PVM was established in our lab involving one Linux server with 8 Xeon CPU cores and two quad-core machines, thus enrolling a total of 16 CPU cores on 3 computers. Openeye's toolkit was used to convert regular multi-conformer MOL2 files to rotor-offset-compressed OEB files, which significantly reduced the I/O on the master and allowed ROCS, EON and FRED to "scale" to a larger number of CPUs. An additional variable for PVM scaling (pvmpass) was used to modify the number of molecules sent to each slave. The default setting was changed from 5 to 7 based on basic scalability tests.

2.1.4. Virtual screening on HPC clusters

All LigMatch calculations were run on the SeARCH cluster at University of Minho, in Portugal. The runs used up to 24 CPU cores (Intel Xeon E5130) across 6 computing nodes, employing a *task-farming* scheme (often referred to as *false parallel*). Overall, each run took approximately 25 days to complete. The high CPU consumption is linked with the relatively low speed of the program GH8, part of LigMatch's technology, which ap-

proximately aligns a modest number of 130 conformers per minute on a single CPU core (compared to the 600-800 conformers per second attained by ROCS).

All top-1000 virtual hits retrieved by each VS protocol were docked to TTR receptors (1bm7, 2g5u and *1bm7opt*) using AutoDock 4 on the Milipeia cluster at University of Coimbra, in Portugal. The runs used up to 16 CPU cores (AMD Opteron 275) across 4 computing nodes (Sun Fire X4100), employing a *task-farming* scheme. Each top-1000 series took approximately 70 hours to dock, and a total of approximately 1500 hours of computation were used.

2.1.5. Virtual screening on a Volunteer Computing platform

Ibercivis is a volunteer computing platform launched in Spain with the goal of serving the computing needs of multiple scientific research groups in the Iberian Peninsula and bringing citizens worldwide to actively collaborate in the Science making process (www.ibercivis.net, accessed 01/09/2011). Currently, the platform hosts ten scientific projects coming from different countries and in various fields of research. It makes use of the Berkeley Open Infrastructure for Network Computing (BOINC) middleware (http://boinc.berkeley.edu/), whose framework operates under the client-server architecture. A client program runs on the volunteer's computer under client-predefined settings. The local computer contacts with any of the project servers it is registered with to retrieve a processing job (workunit). Once the processing of the downloaded workunit is concluded, the client contacts the project server to upload the result and requests a new job from the same or an alternative server.

During this PhD project, our docking-based VS protocols were ported onto Ibercivis under the subproject (codename) AMILOIDE. **Figure 7.1** shows an adaption of the general BOINC scheme, illustrating how the protein-ligand docking and re-scoring pipelines integrate in the architecture underlying Ibercivis. At the project back-end, the entire workflow is managed by a set of wrapper scripts and daemons, allowing researchers to easily choose which protocols they wish to assign to a particular job queue. Given the computational costs associated to each task, only the protein-ligand docking procedure is executed at the participant's computer. The preparation of receptor and ligand structures, including the computation of 3D affinity maps for all possible ligand atom types, is done in the server complex. Equally, the re-scoring pipeline takes place back in the server once results are retrieved. This pipeline is only available to researchers holding valid licenses for DrugScore^{CSD} and/or DSX.

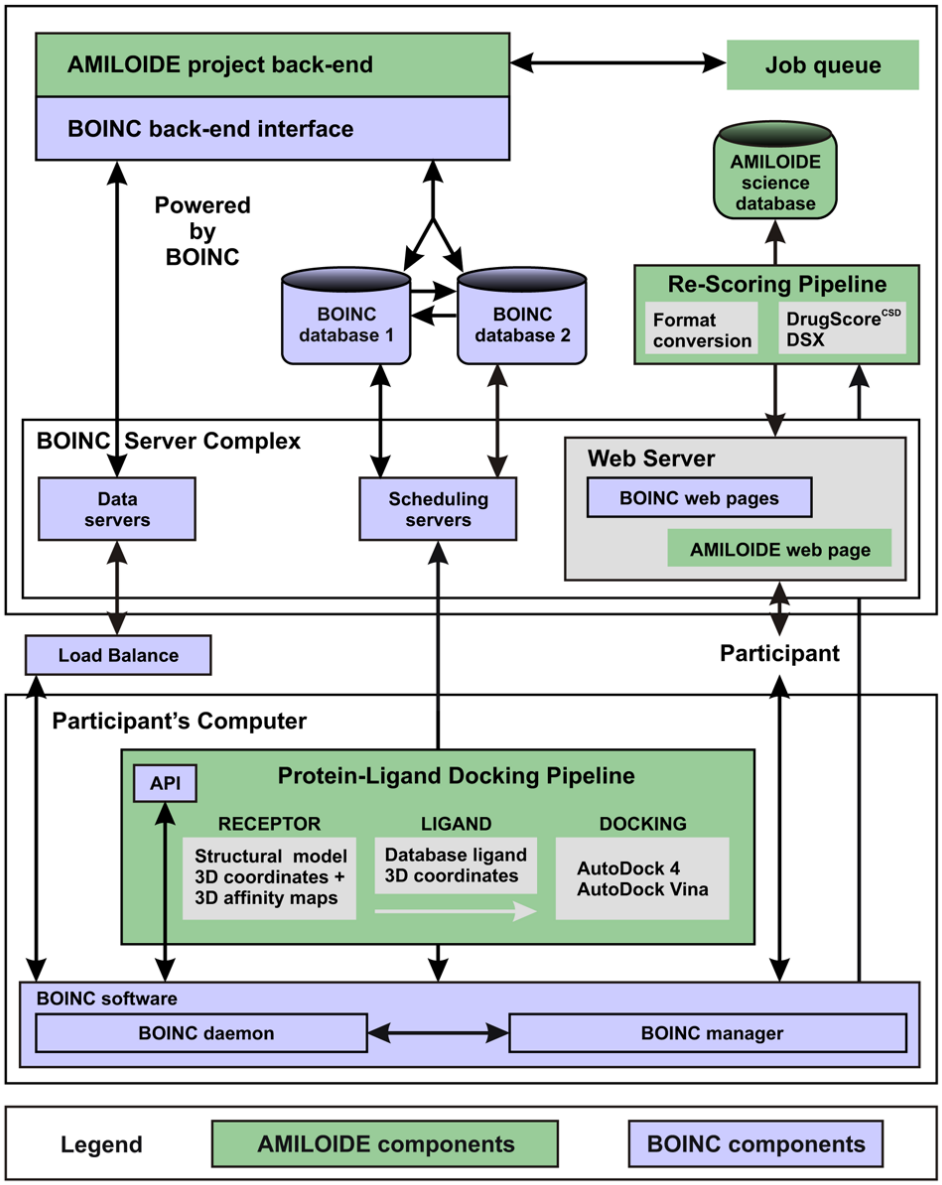


Figure 7.1. Integration of the AMILOIDE project in the BOINC system architecture underlying Ibercivis. Objects in green are part of the hosted project, whereas objects in blue are part of the software suite comprising the BOINC system. At the participant's level, a BOINC work unit is received containing one receptor structure with respective 3D maps of atomic affinity potentials and one ligand from the virtual library of compounds. Then, AutoDock 4 and/or AutoDock Vina perform docking. Back at the BOINC server, results retrieved are validated and submitted to re-scoring with DrugScore^{CSD} and/or DSX.

The implementation of the AMILOIDE subproject motivated the expansion of the Ibercivis platform from Spain, where it was originally conceived under the name Zivis, to Portuguese territory. Ibercivis has been developed by researchers at the Institute of Biocomputation and Physics of Complex Systems (BIFI), in Spain. The expansion to Portugal was carried out by our group at the Centre for Neuroscience and Cell Biology (CNC), in Portugal, and comprised two branches of development. One branch involved the expan-

sion of the Ibercivis backbone, providing a new scheduler (i.e. the program that coordinates the work that is issued, making the best use of the available citizen computers) and a Portuguese version of the corporate website, where citizens can retrieve information about the projects, subscribe to the network and consult diverse statistics. The second branch involved the expansion of the Ibercivis application development, with the training of a local support group and installation of a server for application hosting. Two dedicated blade servers were installed at the Foundation for National Scientific Computing (FCCN), in Portugal, featuring 4 Xeon E7430 CPUs and 16 GB of RAM. Currently, the total storage capacity available to the scientific applications is 10 terabytes.

The new BOINC system architecture underlying Ibercivis features two main hosts, one in Portugal and one in Spain, that contain the feeder and the scheduler module, for sending tasks to BOINC clients and for registering new ones. The servers also contain the scientific subprojects' directories, hosting either the applications' signatures or the applications themselves. Auxiliary hosts contain additional functionalities: a few hosts in BIFI (Zaragoza, Spain) and CSIC (Madrid, Spain) provide BOINC database redundancy; a host at Rediris (Madrid, Spain) provides redirection proxies and auxiliary services, such as database-purge and file deletion control. Templates and code for generating BOINC workunits reside in research group hosts or in common facility hosts. The templates and the transitioner daemon to generate results reside in the main host in Spain, and a redundancy database is to be installed in the main host in Portugal. A common facility that receives results and runs validation and assimilation is provided by CETA-CIEMAT (Madrid, Spain) and *ibercivis04* at FCCN (Lisbon, Portugal). Academic groups associated with each subproject can volunteer new BOINC hosts, or they can use generic ones provided by CETA-CIEMAT, FCCN and BIFI. Finally, a load balancer allows redirects towards alternate hosts, in case of failure of upload/download services.

2.2. Biological evaluation

Under the aim of the discovery of novel TTR stabilisers and the validation of our VS protocols, we implemented an *in vitro* turbidity assay wherein partial unfolding and aggregation of recombinant TTR is promoted by acidification. Samples of purified WT-TTR (at a concentration of 3.6 μM) were mixed with each of the compounds under evaluation (7.2 μM), incubated for 30 minutes at 25 °C, and then acidified to *p*H 4.4 to induce amyloidogenesis. The samples were then incubated at 37 °C for 72 hours and turbidity measurements were taken at 350 and 400 nm using a UV-visible spectrometer, and at 0, 24, 48 and 72 hours time points. Positive and negative control assays were performed in the presenc of flufenamic acid and in the absence of ligand, respectively.

3. Results and discussion

In this section, we present and discuss the most relevant outcomes of several VS campaigns aimed at identifying new TTR amyloid inhibitors. Twenty-two VS protocols based on 2D similarity, 3D (shape and chemical) similarity, 3D electrostatic similarity searches, and docking and scoring, were exploited under four distinct computational environments, from laptop and desktop computers, through a parallel virtual machine and an HPC cluster, all the way to a volunteer computing plaform.

Of all types of methods, 2D similarity-search based methods were by far the fastest, taking only a couple of hours to screen the entire library of 2.3 million compounds on a laptop computer. 3D similarity search protocols based on OpenEye's software were also reasonably fast, taking approximately 2.6 hours to screen the library on 16 CPU cores (enrolled on the PVM). LigMatch calculations were far more demanding, taking approximately one month (overall) to complete on an HPC cluster, using up to 24 CPU cores.

As expected, the most demanding calculations were those emplyoing VS protocols based on AutoDock 4, which prompted us to use the volunteer computing platform Ibercivis. Under the scope of AMILOIDE, we carried out three independent virtual screening campaigns using the three TTR structures identified in Chapter 2 (1bm7, 2g5u and *1bm7opt*) and the AD4 + DrugScore^{CSD} protocol. Overall, the three campaigns took approximately 15 months to complete, consumed approximately 8 million hours of BOINC time, which in turn corresponds to approximately 3.3 million hours of (continuous) CPU time. It should be noted that this number is almost twice the estimated CPU time for the three campaigns, which is due to the fact that not all participants' computers use 64-bit CPUs. In fact, we verified that, on average, 32-bit CPUs take nearly three times as much as 64-bit CPUs to complete a docking run with AutoDock 4. AMILOIDE workunits were processed across almost 70,000 different CPU cores, running on an average of 10,000 CPU cores per day. Approximately 95 million BOINC credits were granted to Ibercivis participants throughout the course of the first 1.5 years of activity. Of all results retrieved and validated, 94% corresponded to successfully accomplished docking runs. All missing/aborted workunits were re-submitted until the full completion of the three screening campaigns.

Inopportunely, however, none of the three VS campaigns running on Ibercivis was completed before the analyses presented in this chapter were performed. Therefore, of the 22 executed VS protocols, only 19 could be included in the following subsections at this stage.

3.1. Inspection of virtual screening hits

All top-1000 VS hits retrived by each VS protocol were compiled in SDF files and tabled in linked spreadshits, where, for every entry retrieved by each protocol, all available scores and similarity measures were stored. These spreadshits were studied in order to identify trends and consensus amongst the different VS protocols. Throughout the following sub-sections we present some of the main insights derived from the inspection of the top-1000 and the top-100 VS hits.

3.1.1. Molecular properties

Figure 7.2 summarizes predicted physicochemical properties for the top-100 hits retrieved by each of the best-performing VS protocols. More detailed information is given in **Table F.1** of the Appendix. It is worth noting that despite the use of template queries holding heavy halogen atoms, the median values for molecular weight across all sets of hits is within the range of lead-likeness, typically defined between 150 and 440 Daltons. This may relate back to the reduced halide fraction of our tailored screening set, translated into low halide fractions amongst the top hits (null median halide fraction for most VS protocols). However, there are at least two noticeable exceptions to this trend (VS protocols *2Dsim CHF Tanimoto PCB18* and *2Dsim UNITY Tanimoto PCB18*; see **Figure 7.2**) and a few outliers amongst the VS hits with near-maximum halide fraction allowance (e.g. *ROCS Scaled Color T_4*; see **Figure 7.2**), which shows that the chosen VS protocols can themselves suggest molecules devoid of undesirable halogens. Equally, the predicted octanol/water partition coefficients (XlogP) and solubility profiles of the selected hits are by far more favourable than those of the known actives. Eighteen out of the 26 known actives (69%) fall in the insoluble, poorly soluble or moderately soluble categories, whereas only 234 out of 1900 top-100 hits (12%) fit in these categories. Moreover, 31% of the known actives are predicted aggregators, whereas only 0.7% of the retrieved hits fail the aggregator filter. Interestingly, VS protocols employing T_4 as a template query (namely VS protocols *2D CHF Euclidean T_4*, *ROCS Scaled Color T_4* and *LigMatch 2D+3D*) retrieve the most soluble series of candidates.

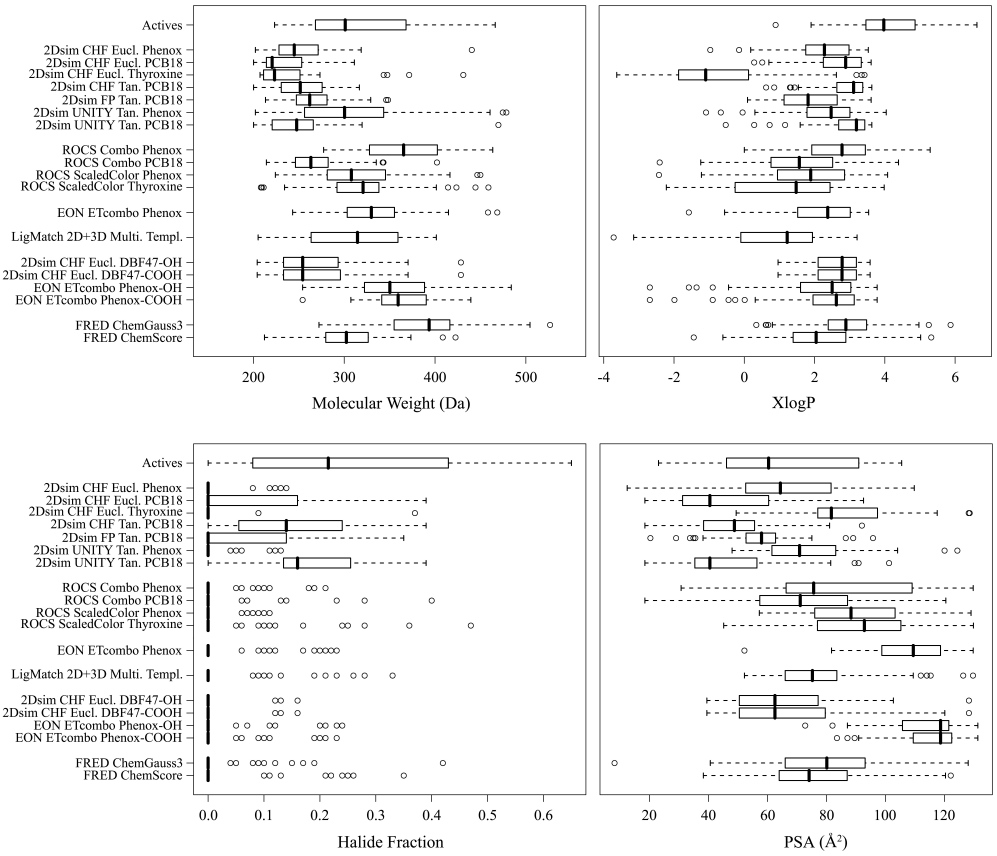


Figure 7.2. Box plots to compare four different physicochemical properties across the top-100 hit compounds retrieved by each VS protocol (box region providing a graphical view of the median and quartiles of the distribution; dotted lines extending to maximum and minimum values of the data set): molecular weight (MW), predicted octanol/water partition coefficient (XlogP), halide fraction and polar surface area (PSA)**.** The properties of TTR actives (compounds 1–23 in **Figure 4.2** of Chapter 4) are given in the first row for comparison purposes**.**

In comparison with the known TTR actives, the top ranked compounds generally present a higher number of hydrogen-bond acceptor and donor groups, which in principle could establish additional points of interaction within the TTR receptor sites and increase binding specificity. This would favour the enthalpic contributions to ligand binding, which are beneficial in terms of specificity. Moreover, the median values for the sum of formal charges are mostly negative amongst the series of VS hits. Series selected by protocols using EON, in particular, present median values of minus 2 (-2.0). This may also be beneficial given the presence of two positively charged Lysine residues at the entrance of the TTR binding sites. Not so favourable results are found in terms of ligand flexibility, as the median values for the number of rotatable bonds are higher for the series of hits compared to the known active molecules. It is likely this will represent an entropic penalty to ligand binding, and therefore extra attention must be paid to this aspect during the selection of hits for biological evaluation and eventual lead optimisation stages.

From the pharmacological viewpoint, the targeting of TTR represents an uncharacteristic challenge, since the blood plasma constitutes not only a vessel for drug distribution but also the actual target compartment itself. Therefore, an important aspect to take into consideration during the selection/design of inhibitors is their ability to cross membranes and spread through compartments where they are likely to cause adverse effects. The polar surface area (PSA) is a descriptor that has shown good inverse correlation with passive transport across biological membranes. A PSA of 60 $Å^2$ or less is associated with the ability to permeate the blood-brain barrier, whereas 140 $Å^2$ is the higher limit for drug absorption. Kelder et al. suggested a "window of opportunity" for designing non-CNS penetrating and orally absorbed compounds, by maintaining PSA values between 70 and 120 $Å^2$ [665]. The median PSA value for the known TTR actives is 60 $Å^2$. In contrast, most of our VS hits present a median PSA value higher than 70 $Å^2$. The exceptions are mostly comprised of the series selected by 2D similarity searches using PCB18 as template. At this point it should be noted that although PCB18 is a highly potent inhibitor of TTR amyloid, it is part of a family of compounds (PCBs) that has been connected with bioaccumulation, rodent and human toxicity (see reference [219] and references therein). These features can be linked with the high hydrophobicity of the compounds or their low PSA. Even though PCB18 has proved to be a successful template in terms of VS performance (see Chapter 5), a critical analysis of hits retrieved by VS protocols employing this template must be undertaken.

3.1.2. Complementarity between VS protocols

To ascertain the complementarity between protocols we examined the degree of overlap amongst the top-1000 hits retrieved by each protocol (a heat map is provided in **Figure 7.3**). As anticipated, given the large size of the screening set and the diversity of methods employed, a high level of complementarity is found across the VS protocols, with more than 80% of the hits being unique. As also expected, the patterns of overlap are consistent with commonalities in the underlying VS protocols. Indeed, protocols employing the same template query and/or similarity metric consistently show some overlap. The highest degree of overlap is observed amongst 2D similarity search methods employing the same similarity metric. However, 2D Pharmacophore Fingerprints represent an interesting outlier, showing a pattern of overlap similar to that of ROCS combo score, LigMatch and docking with FRED (i.e. no significant overlap with any other method). These results prove that the use of 2D fingerprints alone is associated with a more conservative selection, biased towards the chosen template. The increase in complexity of the methods, by inclusion of additional information such as molecular shape, electrostatic properties,

multiple templates or receptor structure data, yields a more diverse selection of compounds, which is a sought outcome in a VS campaign.

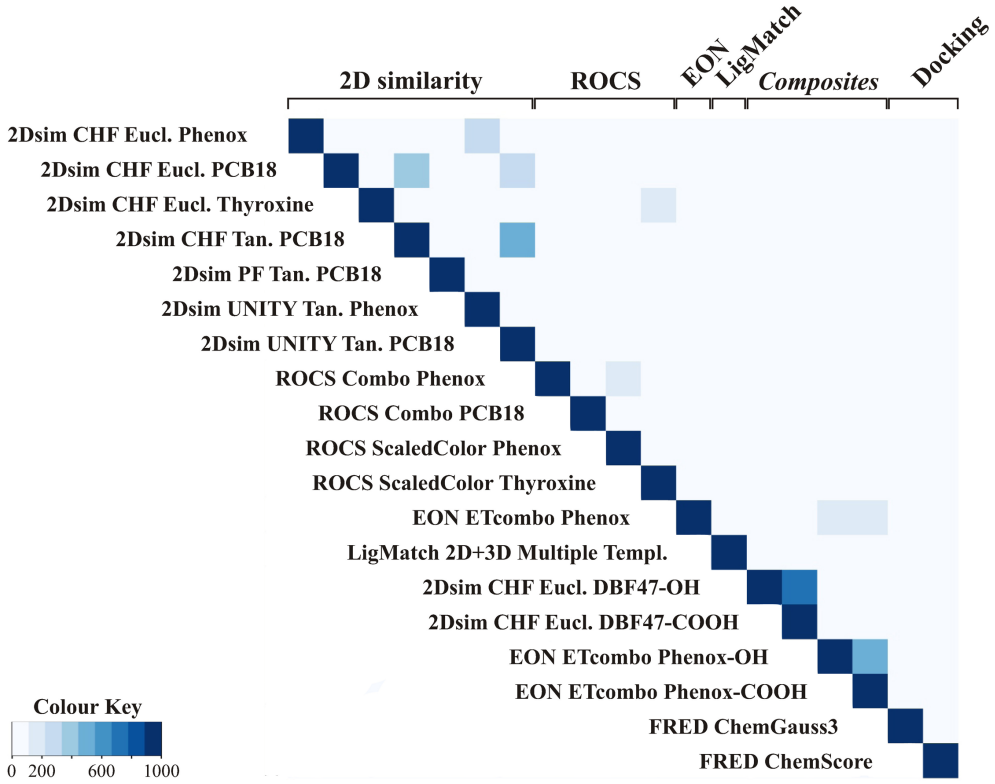


Figure 7.3. A heat map for overlapping entries amongst the top-1000 hits retrieved from the screening set by each VS method. As shown in the colour key, darker blues represent strong overlap, whereas lighter blues represent poor overlap.

3.1.3. Docking and scoring of top-100 virtual hits

All series of top-100 VS hits were docked into the binding sites of TTR with AD4, along with all active, composite and decoy molecules. The resulting complexes were visually inspected and re-scored with DrugScore^{CSD}. Detailed results are given in **Table F.2** of the Appendix. In spite of the limited accuracy of scoring functions at predicting reported binding affinities, DrugScore^{CSD} was able to distinguish reasonably well the known TTR actives from the decoy molecules (as seen in the enrichment studies, Chapter 5). In general, interactions established with the known actives on binding site *AC* are assigned with better scores than interactions within site *BD*. Although no consistent structural differences were found between the two binding sites across multiple complexes, the number of hydrogen bonds established with the TTR receptor is consistently inferior on site *BD* for all series of VS hits. The median score attained for the active compounds on binding site *AC* is the highest of all series. This does not apply to binding site *BD*, however, where the median scores favour some of the series of VS hits, namely those retrieved using the

2D UNITY Phenox, *LigMatch 2D+3D*, *2D CHF DBF47COOH* and *FRED Chemscore* protocols. Better combinations of median scores on the two TTR binding sites are in fact found for many VS hit series, which might be interpreted as a positive indication of potential candidates with favourable binding cooperativity.

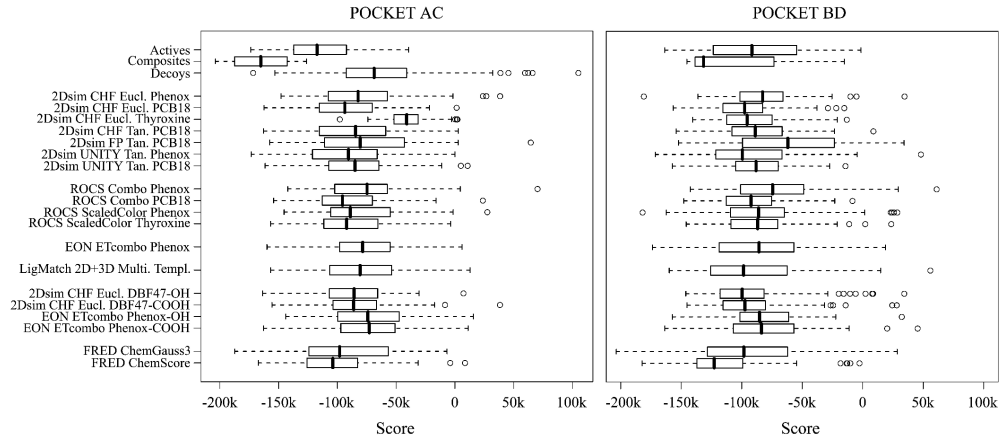


Figure 7.4. Box plots to compare the scores assigned to the top-100 hit compounds retrieved by each VS protocol. All top-100 virtual hits were docked into both TTR binding pockets and re-scored with DrugScore^{CSD}. For comparison purposes, the scores assigned to TTR actives (compounds 1–23 in **Figure 4.2**, Chapter 4), composite ligands and decoys of the benchmarking set are displayed in the first three rows.

3.1.4. Fitness of the virtual hits for TTR pharmacophore models

The fitness of the top-100 VS hits for the pharmacophore queries devised for both TTR pockets in Chapter 3 was assessed using UNITY 3D searches within SYBYL. Detailed results are provided in **Table F.2** of the Appendix. Unsurprisingly, VS protocols employing phenox as template were able to identify ligands that fit the more specific QRY AC1 (see **Figure 3.11** of Chapter 3). On the other hand, protocols using the PCB18 template performed better at retrieving ligands that fit the more general (and conceivably more "promiscuous") pharmacophore models of TTR: mainly QRY AC2 (**Figure 3.12**), and also QRY BD1 (**Figure 3.14**) and QRY BD2 (**Figure 3.15**). While the latter models may be useful to identify a higher number of TTR binders (lead-like compounds), QRY AC1 seems to show promise for identifying more specific inhibitors (with less room for optimisation).

It is interesting to note that, of the 31 compounds showing experimental inhibitory activity of TTR-amyloid formation, 23 fit at least one of the 3D pharmacophore hypotheses. The same applies for 4 out of the 5 excellent inhibitors discovered. Strikingly, amongst the top-100 VS hits, only 4 compounds passed the combination of queries AC1/BD1. Amongst these 4 compounds is the strongest amyloid inhibitor discovered. Unluckily, the remaining 3 hits could not be tested experimentally due to depletion of the respective

vendors' shelves. Nevertheless, this combination of hypotheses will be further used for post-processing hits retrieved by the large-scale VS campaigns run on Ibercivis.

3.2. Selection of hits for experimental evaluation – the "educated guess"

The final selection of compounds to proceed for experimental evaluation was carried out by visual inspection of the docked (lowest-energy) pose of approximately 1000 virtual hits retrieved by each VS protocol. The visual analysis of the VS hits was closely accompanied by direct inspection of their molecular properties – namely, their MW, XlogP, halide fraction, and PSA (as plotted in **Figure 7.2** for the top-100 hits of each VS protocol), and other measures such as their predicted affinities (given by DrugScore^{CSD}), number of hydrogen bonds established within TTR binding sites and fitness for the pharmacophore queries developed in Chapter 3 (as tabled in **Table F.2** of the Appendix). Approximately 80 compounds were marked as "promising", according to their properties/scores and sensible interactions established within TTR binding sites. Of these, 47 compounds were available in shelve from different chemical suppliers and were purchased to be experimentally tested.

3.3. Experimental evaluation

Experimental testing of the activity of the selected screening hits was assessed through a biochemical assay based on promoting TTR fibril formation through acidification of the test solution. Formation of amyloid fibrils in the presence or absence of the compounds under testing is followed by turbidity measurements taken at different points in time. The values presented in **Table 7.2** were determined after 72 hours of incubation in the presence of each of the 38 VS hits so far tested.

Of the 38 compounds experimentally tested, 31 compounds show inhibitory activity to some extent, disclosing ability to bind and stabilise TTR. Five compounds revealed good inhibitory activity, allowing less than 40% fibril formation *in vitro* at protein-ligand stoichiometry 1:2 (compounds **5**, **6**, **9**, **12** and **14** in **Table 7.2**). According to the classification used by Petrassi et al. [222], one compound, in particular, proved to be an *excellent* inhibitor, allowing less than 10% fibril formation under the same test conditions. **Figure 7.5** shows an endpoint plot for fibril formation in the presence of ten of the VS hits tested, illustrating a gradient of inhibition.

Table 7.2. Percentage of fibril formation after 72 hours of incubation with each of 38 VS hits tested. The values represent an average of 3 replicates.

Compound	% fibril formation (at 72 hours)
1	94.1
2	79.4
3	42.9
4	49.3
5	**3.1**
6	**18.4**
7	71.0
8	52.7
9	**39.7**
10	47.6
11	101.0
12	**15.5**
13	78.0
14	**36.7**
15	54.7
16	99.0
17	97.0
18	91.8
19	86.8
20	90.0
21	69.7
22	95.7
23	90.6
24	61.2
25	94.1
26	94.2
27	97.7
28	91.6
29	94.5
30	102.2
31	77.7
32	103.7
33	103.4
34	104.1
35	104.9
36	97.7
37	92.7
38	102.0

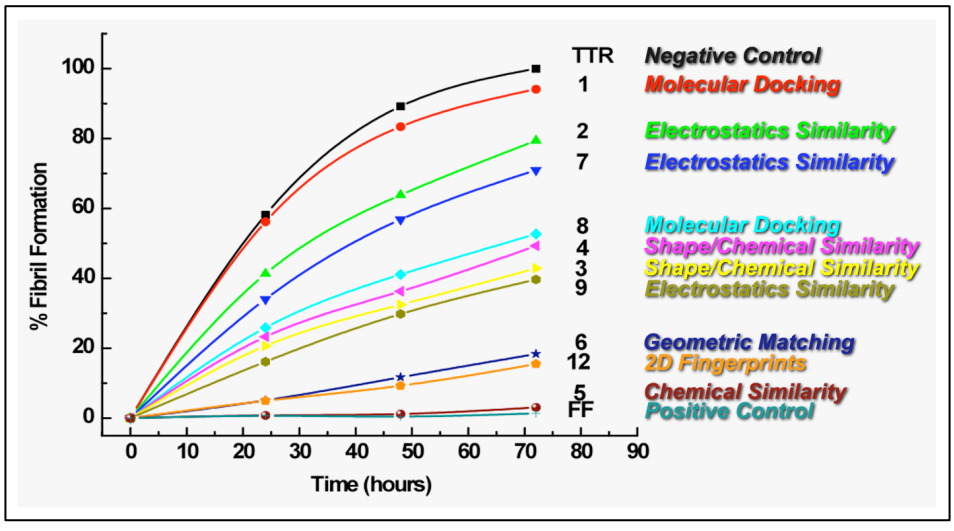


Figure 7.5. Percentage of in vitro fibril formation by transthyretin measured by turbidimetry at 24, 48 and 72 hours (at *p*H 4.4 and 37 °C). The numbers correspond to virtual hits prioritized for experimental evaluation. On the right, virtual screening techniques employed to select each of the hits are identified. The negative control corresponds to incubation of TTR in the absence of compounds, whereas the positive control corresponds to incubation of TTR in the presence of flufenamic acid, a known and potent inhibitor.

3.4. Evaluation of VS methods

The family of VS methods through which each compound was identified is also disclosed in **Figure 7.5**. In the figure, molecular docking refers to VS protocols based on FRED; shape/chemical similarity refers to the Combo Score of ROCS, whereas chemical similarity, alone, corresponds to the ROCS' ScaledColor score; electrostatic similarity alludes to the ET_combo score of EON; "geometric matching" corresponds to the LigMatch method; 2D fingerprints represents 2D similarity searches performed with ScreenMD using chemical hashed fingerprints. In increasing level of inhibitory activity, compound **1** was selected via molecular docking with *FRED*, employing the *ChemGauss3* scoring function; compounds **2** and **7** were identified via *EON*, both using the *ET_combo* ranking function, and respectively the "*Phenox-OH*" concatemer and *Phenox* as templates; compound **8** was selected with the *FRED ChemScore* protocol; both compounds **4** and **3** were discovered using *ROCS' Combo score* with *Phenox* as template, and compound **3** was further prioritized by *EON* using the "*Phenox-OH*" concatamer as template. Focusing on the best inhibitors discovered (those that were able to inhibit at least 60% of TTR fibril formation), compound **9** was identified via the *EON ET_combo Phenox* protocol; compound **14** (not shown in **Figure 7.5**) was identified via molecular docking with *FRED*, using the *Chemscore* scoring function; compounds **6** and **12** were discovered using the *2D CHF (ED) Phenox* protocol, and compound **6** was further prioritized by 3D geometric hashing with *LigMatch*. Finally, the strongest inhibitor discovered throughout this project, compound **5**, was identified using the *ROCS ScaledColor* T_4 VS protocol.

Besides linking the experimental activity of the most promising inhibitors to the underlying (best-performing) VS protocols that allowed their selection, it would also be interesting to carry out a retrospective analysis to assess how that linking mirrors the expectations raised for each protocol (Chapter 5). Nearly all of the 38 compounds that were experimentally tested were discovered within the range of the top-100 VS hits. This range roughly represents 0.005% of our screening set. Looking back to the clues provided by the ROCE values at 0.5 % of the ranked benchmarking set, it is interesting to see that the *ROCS ComboScore/ScaledColor Phenox* protocols and the *EON ET_combo Phenox* protocol provided the highest (Maximum) enrichment, followed the *ROCS ComboScore PCB18* protocol and the *ROCS ScaledColor Thyroxine* protocol. However, it is also worth noting that amongst all ROCS-based protocols assessed in this project, the *ROCS ScaledColor Thyroxine* protocol is the one showing the highest area under the (enrichment) curve (AUC). Equally, the *2D CHF (ED) Phenox* protocol present one of the highest AUC of all 2D-based VS protocols. Taken together, these results lead us to believe that, besides the early enrichment values, the AUC measure of (overall) performance can also provide good indications about the outcomes of virtual screening methods.

Clearly, an accurate and "fair" evaluation/validation of the VS protocols would require an unbiased, "non-educated", selection of compounds to be experimentally tested; a pre-defined number of compounds (for instance, 5 compounds) selected within a pre-defined range (for instance, the top-50 entries) should be purchased and tested, so that "experimental" enrichment factors could be determined and contrasted with those derived from our prospective studies (Chapter 5). This example alone would, however, require the testing of at least 110 compounds retrieved by the 22 VS protocols. Time and money are important constraints in every project and, unfortunately, we had to focus our efforts in the selection of compounds holding features that represent clear indications of their ability to bind, and thus stabilise, TTR tetramers. Hopefully, a more systematic validation of our VS protocols will be pursuit in the near future.

4. Conclusions

In this chapter, 22 VS protocols were implemented and explored on diverse computing plaftforms with the objective of discoverying novel TTR amyloid inhibitors. Along with 3D pharmacophore models of TTR, docking and scoring methods were used to study and prioritize the most promising candidates found amongst the top VS hits retrieved by fast ligand-based similarity search protocols and by a fast rigid docking algorithm. By analysis of the median and maximum scores obtained with the DrugScore^{CSD} function, we detected balanced profiles of estimated binding affinities on the two binding sites of TTR. The top VS hits retrieved by 2D similarity methods showed the highest compliance with the pharmacophore models devised for TTR. However, their structural similarity to the employed templates is obvious, disclosing limited ability to suggest new chemical entities (scaffold hopping). 3D similarity search methods, by contrast, provided a good compromise between retaining the essential features for TTR binding and scaffold hopping. When compared with the high-scoring compounds retrieved by most ligand-based methods, a larger percentage of *potential* false positives (compounds with inappropriate molecular shape, size or chemical topology) was found amongst the hits retrieved by docking-based methods.

Forty-seven commercially available candidates were purchased from their respective chemical suppliers. Of these, 38 have been experimentally tested for inhibition of TTR fibril formation. Thirty-one compounds showed inhibitory activity to some extent, of which five revealed to be good inhibitors, reducing amyloid formation to less than 40% *in vitro* at a protein-ligand concentration ratio of 1:2. Amongst the discovered amyloid inhibitors, the third strongest inhibitor was selected by LigMatch, a new method developed in our group at the University of Leeds that employs multiple templates. The second

strongest amyloid inhibitor was discovered through 2D similarity searches, showing some resemblance with the used template molecule. The strongest inhibitor discovered in this project (97% inhibition) was identified through a protocol based on ROCS and its ScaledColor scoring function.

In this chapter, my self-criticism goes to the handling of the large amounts of data generated through virtual screening, which could have been done more efficiently through the construction of a relational database to store VS data along with molecular property information. In turn, this could have opened doors to more profitful mining process, thus allowing for a less "educated" but more systematic selection of candidates for experimental testing. In fact, despite all our efforts throughout this project, I personally believe that what we have done in this chapter was an experimental evaluation of compounds selected with the help of virtual screening methods, rather than a systematic validation of the virtual screening methods that have been studied and implemented throughout this project. Although this may be regarded as the weakest point of my thesis, it is important to realize that a full, systematic and experimental validation of our virtual screening protocols would have required a significant investment that was, at the time, beyond the available budget.

To sum up, by combining the best-performing virtual screening protocols identified throughout this project, a subset of small molecules was retrieved from a tailored library of 2.3 million commercially available compounds. These molecules show distinctive features compared to known TTR stabilisers, holding better physicochemical properties and combined affinities for TTR binding sites. Of critical relevance to this work, from the Medicinal Chemistry viewpoint, is the exceptional solubility, appropriate polar surface areas and low halogen fraction characterising the series of compounds identified by virtual screening. Thirty-eight compounds were tested *in vitro* and five lead compounds were identified.

Further experimental evaluation of the lead compounds discovered during this project is currently being conducted using isothermal titration calorimetry (ITC). This method allows for a precise quantification of the affinity of association between each of the discovered amyloid inhibitors and TTR. Moreover, the ITC experiments will help dissecting the thermodynamical signature of this association, which will in turn open new avenues to structure-activity relationships (SAR) studies. Coupled with organic/enzymatic synthesis of analogs and derivatives, these studies will pave the way to the process of lead optimisation.

5. Conflicts of interest statement

It must be stated that the structure of the active compounds discovered in this project cannot yet be revealed due to questions of intellectual property. Indeed, the most promising candidates found throughout this project are currently under evaluation for patentability, which completely hinders our keen interest in sharing their chemical formulae.

6. Acknowledgments

My first acknowledgement goes to Catarina S. H. Jesus for her extraordinary work with the experimental evaluation of the virtual screening hits. Without her technical skills and resilience before the mind-numbing adversities of the wet lab, the real value of the scientific findings presented in this chapter could not be demonstrated. I also thank the Ibercivis development staff, in particular Alfonso Tarancon, Alejandro Rivero, Cândida Silva, Darío Ferrer, Francisco Sanz García and Javier Latorre Romero for their helpful discussions and technical support throughout the deployment and maintenance of the AMILOIDE subproject. Acknowledgements are also due to the Centre for Computational Physics, Departamento de Física, Universidade de Coimbra, Coimbra, Portugal, and the Computer Science and Technology Centre, Departamento de Informática, Universidade do Minho, Braga, Portugal, for the HPC resources provided. Last but not least, I thank Ibercivis users around the World for their inestimable contribution to this work.

Epilogue

Conclusions and perspectives

"Wisdom is better than wit, and in the long run will certainly have the laugh on her side."

[Jane Austen]

1. Conclusions

The findings reported in this thesis highlight the important role of computer-aided drug design as a rational approach to the drug discovery process. Coupled with adequate experimental validation, the virtual screening methods employed in this project helped revealing the face of promising new lead compounds, which can be regarded as potentially new generations of anti-amyloid drugs should further chemical optimisation be carried out. Yet, more important than reaching an end (albeit rewarding) is what can be learnt throughout the process of reaching that end.

The focus of the project behind this thesis was the criterious evaluation of virtual screening approaches towards the identification of new compounds capable of inhibiting transthyretin (TTR) amyloid formation via binding and stabilisation of the native state protein. First, the principles of TTR-ligand interactions were studied by making use of the richness of TTR crystallographic structures. Several TTR complexes were assessed for structural quality and reliability against molecular docking algorithms. Five docking programs were tested for their ability to handle the characteristics of TTR binding sites, and the best-performing algorithms were selected for the assembly of structure-based virtual screening protocols. Moreover, two sets of pharmacophore models characterising the determinants of ligand binding to each TTR binding site were developed on the basis of the distinct positioning of known ligands within each site. With the aim of discovering new amyloid inhibitors amongst a virtual library containing more than 2 million compounds, several virtual screening protocols were assembled employing techniques that range from two-dimensional (2D) fingerprints and three-dimensional (3D) shape and chemical descriptors, through mixed 2D/3D approaches, all the way to docking and scoring. Using a benchmarking set comprised of the most important TTR stabilisers discovered to date and carefully selected decoy molecules, the performance of the different protocols and models at discriminating active molecules from inactive ones was assessed using complementary metrics based on the concept of enrichment.

Focusing on the enrichment profiles obtained, ligand-based methods performed better than protein-based methods (i.e., docking of compounds into TTR structures followed by re-scoring of the predicted complexes). Two-dimensional similarity search methods, in particular, offered the best discrimination between the known TTR actives and the decoy molecules. Given the reasonable structural diversity found amongst the known actives, these methods provide a good balance between search specificity, a key aspect for target selectivity, and search flexibility, which is necessary for eliminating the dependency on the used template molecules. In fact, the top virtual screening hits returned by these

methods show the highest compliance with the pharmacophore models devised for TTR. Nevertheless, their structural similarity to the employed templates is obvious, disclosing limited scaffold hopping capabilities as shortcoming of 2D similarity search methods. While yielding lower enrichments, 3D similarity search methods based on shape and chemical complementarity were able to suggest novel scaffolds, with noticeable added value when electrostatic similarity was included as a descriptor. Yet, because these methods use one single bioactive template they often overlooked strong TTR stabilisers with dissimilar structure. The use of concatamers to address the problem of representativeness associated to single templates brought improvement to the performance of similarity search methods, and even more balanced outcomes were attained with LigMatch, a method that can make use of multiple templates on a single virtual screening run, while combining 3D geometric hashing with a 2D pre-selection process.

Efforts made in the field of docking and scoring are noteworthy, but our results highlight important limitations of scoring functions in common use, especially when handling non-specific binders holding mostly hydrophobic binding sites, such as TTR. However, we have successfully set up a docking and scoring protocol, respectively based on AutoDock 4 and DrugScore^{CSD}, that was able to provide encouraging enrichment profiles for TTR and negligible bias towards the molecular weight of the ligands. A set of similar protocols was tested against 40 targets of pharmaceutical relevance. For each target, at least one docking and scoring protocol performing considerably better than random selection was identified. These protocols have been implemented in a large volunteer computing platform and are now available to be used by academic researchers free-of-charge.

By combining the best-performing virtual screening protocols against TTR, a subset of molecules was retrieved from a tailored library of 2.3 million commercially available compounds. This subset of approximately 2000 small molecules was analysed and showed distinctive features when compared to the known TTR stabilisers. According to our predictions, these molecules display better profiles of physicochemical properties such as good solubility, appropriate polar surface area and low halide fraction. Moreover, by analysis of the median and maximum score values assigned by DrugScore^{CSD} to these subset of hits, we detected balanced profiles for the estimated binding affinities.

Amongst the top virtual screening hits retrieved by the fast ligand-based similarity search protocols and computationally-expensive docking-based VS protocols, 47 compounds were purchased to be experimentally tested for inhibition of TTR amyloid formation, after selection by visual inspection of low-energy conformations docked into TTR binding sites and analysis of compliance with the two sets of pharmacophore models developed. Of these, 38 have been experimentally tested for inhibition of TTR fibril formation. Thirty-one compounds showed inhibitory activity to some extent, of which five re-

vealed to be good inhibitors, reducing amyloid fibril formation to less than 40% *in vitro* at a protein-ligand concentration *ratio* of 1:2. Of particular importance to our goals are the properties that characterise the compounds prioritized throughout our virtual screening campaigns, such as their exceptional solubility and low halogen content (which are critical to avoid problems like bioaccumulation and toxicity), and appropriate polar surface areas (meaning that the selected compounds are unlikely to cross the blood-brain barrier and thus cause adverse effects in compartments where their activity is not required).

To conclude, we emphasize the importance of a thorough validation of virtual screening methods as an essential step to any virtual screening campaign against targets of biological relevance. It should be noted that the studies presented in this thesis cannot provide an exhaustive comparison of all virtual screening techniques currently available, and different (or even better) results could have been obtained using other methods. The increase in complexity of the virtual screening protocols by inclusion and combination of additional layers of information, such as molecular shape, electrostatic properties, the use of composite templates or multiple templates, and receptor information in molecular docking, may be accompanied by decreases in performance, but rewarded by a higher likelihood of retrieval of novel scaffolds with distinctive physicochemical properties.

Lessons learned in this project can be broadly employed towards the identification of novel and safer amyloid inhibitors.

2. Virtual screening without borders or boundaries: promoting lead discovery in academia with the help of volunteer citizens

The design of this project was originally "biased" towards the goal of discoverying new amyloid inhibitors, focusing on transthyretin as a case study. However, the fact that some of the best results for this particular target were grounded on academic-free software tools paved the way to the development of a computational screening resource, which is now available to the academic researchers. Indeed, the AMILOIDE subproject at Ibercivis was meant to be a seed for further *in silico* lead discovery campaigns to be carried out at the academia. With this in mind, the potential application of currently implemented docking-based protocols against a variety of targets of pharmaceutical interest is well illustrated in this thesis, and several improvements are underway.

A critical aspect of concern to us is the efficient handling and management of the large amounts of data generated by high-throughput docking (HTD) and other virtual screening approaches. At the moment, the integration of a system for the efficient submission and monitoring of HTD jobs and subsequent data management is being studied. Open source solutions based on structured-query language database software have already been designed for VS purposes and represent a promising starting point for new developments of docking and scoring on Ibercivis (see references [666,667]).

Moreover, new ways to facilitate the access of researchers to the distributed virtual screening resource are now envisaged. In 2009, Irwin et al. presented DOCK Blaster, a web-based docking system for the screening of large libraries in a fully automated and unsupervised manner [668]. DOCK Blaster is a remarkable example of a sensible and efficient integration of extremely useful software and hardware resources. Similar features can soon be made available at Ibercivis as a follow up of the work presented in this thesis. Surely, a web-based "on-demand" HTD service will represent a pioneer concept in the volunteer computing realm, as it will also offer investigators an unprecedented resource for large-scale virtual screening.

As with many things in life, isolated Science means no Science at all. Science communication should, therefore, be regarded as critical component of the work of any researcher. With this in mind, the Ibercivis project has also proven to be an excellent resource to take scientific findings to the volunteer citizens and to the population in general. Over the last years, a large amount of educational contents have been produced and scattered through the Ibercivis communication channels, which as been utterly rewarding. Science dissemination will be further kept across the media (press, radio, television, with particular focus on the Internet). The website is managing all the information and contents developed for the project, such as the information required to participate in the project, personal statistics and global information, teaching units for secondary schools, leaflets and podcasts (and videocasts) of the scientific research being carried out.

The AMILOIDE subproject marked the onset of the Portuguese contribution on Ibercivis, a notable achievement to the distributed computing realm in the Iberian Peninsula. So far, the acceptance of the project by the general public has been quite positive and encouraging. It is our hope that Portuguese citizens will take more and more interest in the process of scientific discovery and feel motivated to contribute to it by donating CPU time on the network. The growth of Ibercivis will open exciting new avenues to the AMILOIDE project, like the study of mutation effects on the binding of drugs to different proteins involved in amyloid disorders, and the implementation of more complex methods to estimate binding affinities more accurately. But it will also foster the spawn of new scien-

tific research projects that have been hindered due to limitations of computing resources and/or processing power.

3. Perspectives

Democratisation of resources on which virtual screening depends is essential to bring the technique to a broad community. The Protein Data Bank (PDB) and MODBASE already provide thousands of receptor structures determined experimentally or by computer-based modelling, respectively. Resources like ChEMBL, DrugBank, BindingDB or Binding MOAD make available activity data of millions of small organic molecules. Several groups and software developers offer many molecular modelling applications free-of-charge to academics – including docking programs, similarity search algorithms and, more recently, pharmacophore modelling and search tools. Even though some of these programs require expert knowledge or great learning efforts, web-accessible resources ease the accessibility of virtual screening methods to non-specialists. Community-accessible chemical libraries are also indispensable and rapidly evolving. The National Cancer Institute (NCI) provides calculated structures for approximately 140,000 of its compounds, and even provides some of these for free to academics for experimental testing. ZINC is a free, web-accessible, database that currently stores more than 17 million purchasable compounds. The chemical universe database GDB-13 is also accessible for download and stores more than 977 million organic molecules (containing up to 13 atoms of C, N, O, S and Cl) that are predicted to be synthetically accessible.

As a result of the growing amount of publicly accessible resources, virtual screening is currently evolving from classic similarity searches based on isolated reference compounds into a more complex application of machine learning and data mining techniques, which can make use of high-information content training sets. In fact, techniques like (*i*) support vector machines (SVMs), (*ii*) decision trees, (*iii*) Bayesian methods, (*iv*) inductive logic programming (ILP), and (v) belief theory are already being applied in virtual screening. This expansion of the repertoire of methods is opening up exciting new applications areas for virtual screening. The concurrent emergence of chemogenomics is calling for the discovery of small molecules that can act as tool compounds to probe biological functions by being selective for subfamilies within certain target families. Equally, the use of such computational methods to predict protein targets of small molecules is gaining momentum. The applications of *reverse* virtual screening methods now range from the so-called "target fishing", through mode-of-action studies, all the way to "rational" polypharmacology. The identification of ligands for undrugged proteins and new members of protein families is yet another application of high interest in chemoinformatics

and chemical biology, often referred to as *orphan* screening. The traditional screening methods usually require the knowledge of the target structure and/or previous ligands to use as reference. Currently, some researchers are using SVM-based approaches to exploit the concept of "homology-based similarity searches", which is based on the idea that homologous proteins present similar binding sites and, thus, interact with similar ligands.

Further expectations in the field of virtual screening regard the development of community-wide efforts to standardise the statistical evaluation of the computational methods, as well as the preparation and the sharing of data sets. This is a critical aspect, since method comparisons are often difficult to reproduce and it is hard to evaluate the relative performance of the methods based on their original reports. With this in mind, it would be very important that researchers working on molecular modelling provided detailed information on the design and setup of their system, showed more willingness to share their benchmarking sets and other relevant data, and stated more clearly the objective of their studies, such as comparing one specific method to other methods in the literature. Community-based efforts have already been taken to aid the implementation of standards in decoy set preparation, such as the Directory of Useful Decoys for docking, but it would be interesting to see these efforts extended to ligand-based virtual screening approaches. More importantly, perhaps, a better awareness of the risks of artificial enrichment should be raised in order to eschew virtual screening artifacts linked to certain activity classes and to the problem of analogue bias discussed in this thesis. Other sources of bias are also important, especially when machine learning methods are being explored based on the concept of similarity, which holds itself a considerable level of variability and subjectiveness, and may itself introduce bias. The establishment of community standards regarding the metrics used to evaluate the performance of virtual screening methods is another basic scientific requirement that should be prioritised by researchers working in the field. Undoubtedly, the myriad of methods that are currently being explored for virtual screening represents a thrilling playground for further improvements to be made in the field.

Appendix

Section A

(Chapter 2)

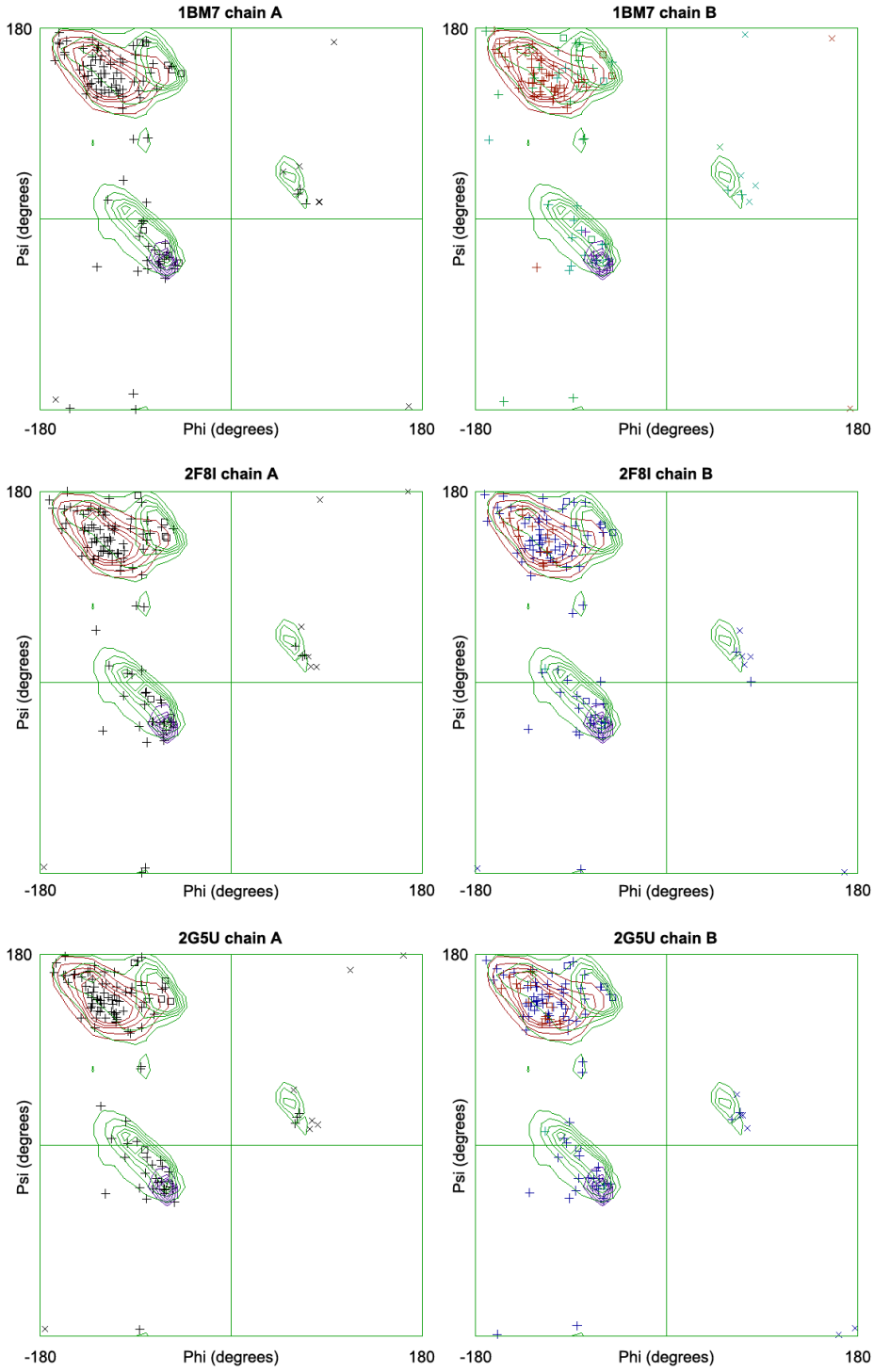


Figure A.1. Ramachandran plots for the three best-performing TTR structures identified through the crossdocking studies. X-signs represent glycines, squares represent prolines, and plus-signs represent the other residues. Residues that are part of a helix are shown in blue, strand residues in red. "Allowed" regions for α-helices are drawn in blue, for β-strands in red, and for all other structures in green.

	2g9k	1bm7	1dvs	1dvt	1dvu	1dvx	1dvy	1dvz	1e4h	1tt6	2b15	2b77	2f7i	2f8i	2g5u	1y1d	2gab	2qgd	2qge	2rox	3b56		Score
2g9k	0	435	442	444	446	442	445	447	448	434	442	431	440	431	449	445	445	446	453	442	441	8848	**8848**
1bm7	0	0	439	440	433	441	440	439	432	411	431	416	428	424	435	438	429	416	425	413	418	8148	8583
1dvs	0	0	0	458	458	458	458	458	440	413	426	434	438	433	444	453	444	429	425	442	427	7938	8819
1dvt	0	0	0	0	458	458	458	458	434	403	424	435	437	432	446	456	444	429	435	443	421	7471	8813
1dvu	0	0	0	0	0	458	458	458	442	395	430	437	436	429	447	454	446	424	441	440	423	7018	8813
1dvx	0	0	0	0	0	0	458	454	437	408	431	431	434	425	443	452	439	418	440	433	427	6530	8787
1dvy	0	0	0	0	0	0	0	458	442	407	427	436	435	428	447	454	446	424	440	447	419	6110	8827
1dvz	0	0	0	0	0	0	0	0	432	415	439	438	437	429	443	458	442	430	440	440	424	5667	8839
1e4h	0	0	0	0	0	0	0	0	0	408	442	419	421	411	449	438	437	429	444	432	419	5149	8656
1tt6	0	0	0	0	0	0	0	0	0	0	456	419	427	420	415	438	376	432	416	372	455	4626	8320
2b15	0	0	0	0	0	0	0	0	0	0	0	432	441	427	438	439	405	438	436	411	460	4327	8675
2b77	0	0	0	0	0	0	0	0	0	0	0	0	446	447	431	440	426	434	430	428	429	3911	8639
2f7i	0	0	0	0	0	0	0	0	0	0	0	0	0	449	438	439	435	441	437	433	427	3499	8719
2f8i	0	0	0	0	0	0	0	0	0	0	0	0	0	0	432	434	427	438	430	419	424	3004	8589
2g5u	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	446	445	433	444	434	431	2633	8790
1y1d	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	437	419	439	422	427	2144	8828
2gab	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	423	435	435	405	1698	8621
2qgd	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	449	422	441	1312	8615
2qge	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	437	439	876	8735
2rox	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	401	401	8546
3b56	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8558
	0	435	881	1342	1795	2257	2717	3172	3507	3694	4348	4728	5220	5585	6157	6684	6923	7303	7859	8145	8558		

Figure A.2. The all-versus-all Cα-atom alignment matrix produced by the program GH8 for 21 high-quality TTR complexes. The total scores were calculated by summing up the row and column scores for each complex. 2g9k was found to be the highest scoring structure. GH8 places a 0 in red for every structure aligned against itself (matrix diagonal) and for redundant scores a 0 in green.

	2g9k	1y1d	1bm7	1dvs	1dvt	1dvu	1dvx	1dvy	1dvz	1e4h	1tt6	2b15	2b77	2f7i	2f8i	2g5u	2gab	2qgd	2qge	2rox	3b56		
2g9k_poc	0	228	353	328	338	321	303	322	345	324	347	329	355	356	344	387	377	339	335	320	332	6683	6683
1y1d_poc	0	0	258	244	250	222	205	230	267	303	311	364	327	353	319	294	254	340	337	207	367	5452	5680
1bm7_poc	0	0	0	366	370	338	308	355	391	339	368	372	369	370	350	365	349	368	335	321	367	6401	7012
1dvs_poc	0	0	0	0	384	349	304	371	380	349	360	361	363	367	371	364	347	367	341	342	359	6079	7017
1dvt_poc	0	0	0	0	0	351	314	381	390	360	374	362	377	380	379	371	351	378	341	329	347	5785	7127
1dvu_poc	0	0	0	0	0	0	296	359	363	371	301	322	333	352	328	345	342	314	320	298	329	4973	6554
1dvx_poc	0	0	0	0	0	0	0	288	321	285	293	298	305	318	304	322	307	311	305	266	294	4217	5947
1dvy_poc	0	0	0	0	0	0	0	0	375	347	342	328	347	348	348	352	344	350	323	346	333	4483	6789
1dvz_poc	0	0	0	0	0	0	0	0	0	353	376	389	384	393	377	380	364	383	349	332	378	4458	7290
1e4h_poc	0	0	0	0	0	0	0	0	0	0	395	419	385	386	395	376	356	347	375	298	375	4107	7138
1tt6_poc	0	0	0	0	0	0	0	0	0	0	0	477	440	427	425	397	342	438	362	334	371	4013	7480
2b15_poc	0	0	0	0	0	0	0	0	0	0	0	0	457	462	434	429	355	447	436	304	551	3875	7896
2b77_poc	0	0	0	0	0	0	0	0	0	0	0	0	0	492	457	411	391	447	417	349	436	3400	7842
2f7i_poc	0	0	0	0	0	0	0	0	0	0	0	0	0	0	485	427	382	470	427	347	476	3014	8018
2f8i_poc	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	407	368	443	414	341	426	2399	7715
2g5u_poc	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	407	415	404	352	414	1992	7619
2gab_poc	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	355	367	334	340	1396	7032
2qgd_poc	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	436	339	451	1226	7738
2qge_poc	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	312	444	756	7380
2rox_poc	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	334	334	6405
3b56_poc	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7724
	0	228	611	938	1342	1581	1730	2306	2832	3031	3467	4021	4442	5004	5316	5627	5636	6512	6624	6071	7724		

Figure A.3. The all-versus-all all-atom alignment matrix produced by the program GH8 for the binding site residues extracted from 21 high-quality TTR complexes. The total scores were calculated by summing up the row and column scores for each complex. 2f7i was found to be the highest scoring structure.

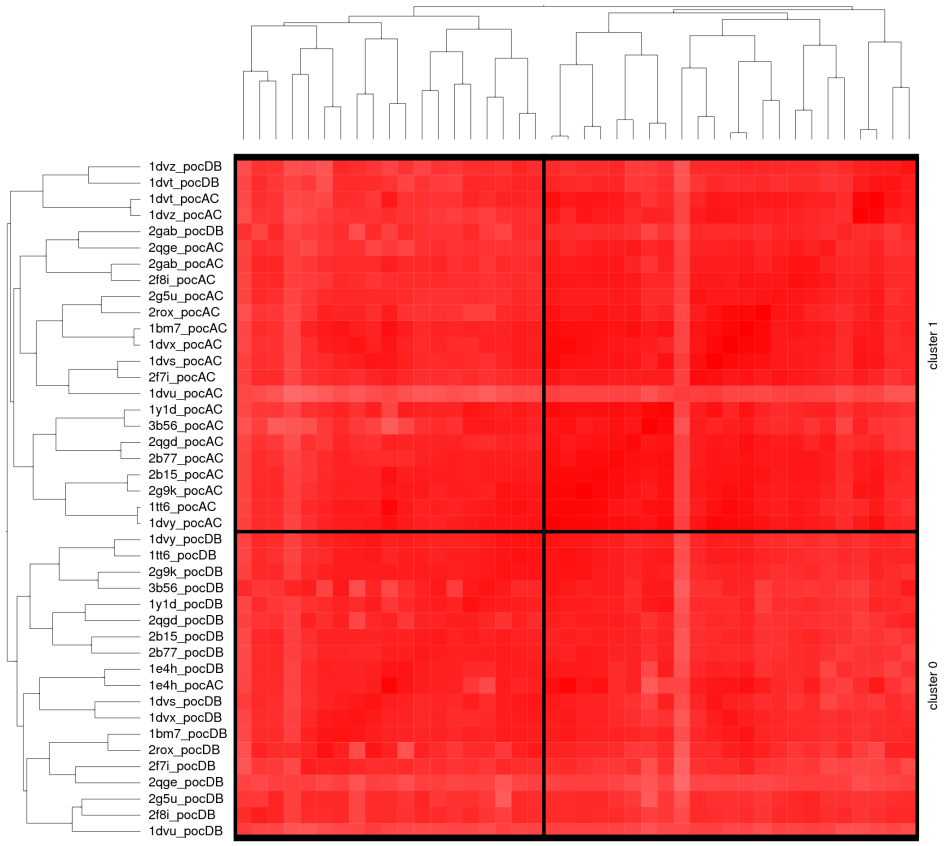


Figure A.4. Clustering analysis performed with the program CLUTO on the all-versus-all Cα-atom alignment matrix produced by the program GH8 for the binding site residues of 21 high-quality TTR complexes. Absence of visually distinguishable clusters indicates a close similarity in scores for the two pockets.

```
npts 50 44 70    # number of grid points in xyz

gridfld 1bm7_AC.maps.fld                # grid data file
spacing 0.375                           # spacing (Å)
receptor_types A C HD N OA SA           # receptor atom types
ligand_types A C F NA OA                # ligand atom types
receptor 1bm7.pdbqt                     # macromolecule
gridcentre 20.488 41.928 6.509          # xyz-coordinates or auto
smooth 0.5                              # store minimum energy w/in rad(Å)
map 1bm7_AC.A.map                       # atom-specific affinity map
map 1bm7_AC.C.map                       # atom-specific affinity map
map 1bm7_AC.F.map                       # atom-specific affinity map
map 1bm7_AC.NA.map                      # atom-specific affinity map
map 1bm7_AC.OA.map                      # atom-specific affinity map
elecmap 1bm7_AC.e.map                   # electrostatic potential map
dsolvmap 1bm7_AC.d.map                  # desolvation potential map
dielectric -0.1465                      # <0, AD4 distance-dependent diel.; >0, constant
```

Figure A.5. An example of a grid parameter file used as input for AutoGrid (part of AutoDock 4).

```
outlev 1                                # diagnostic output level
intelec                                 # calculate internal electrostatics
seed pid time                           # seeds for random generator
ligand_types A C F NA OA                # atoms types in ligand
fld 1bm7.maps.fld                       # grid data file
map 1bm7.A.map                          # atom-specific affinity map
map 1bm7.C.map                          # atom-specific affinity map
map 1bm7.F.map                          # atom-specific affinity map
map 1bm7.NA.map                         # atom-specific affinity map
map 1bm7.OA.map                         # atom-specific affinity map
elecmap 1bm7.e.map                      # electrostatics map
desolvmap 1bm7.d.map                    # desolvation map
move 1bm7_ligand_AC.pdbqt               # small molecule
about 20.488 41.928 6.509               # small molecule centre
tran0 random                            # initial coordinates/Å or random
quat0 random                            # initial quaternion
ndihe 4                                 # number of active torsions
dihe0 random                            # initial dihedrals (relative) or random
tstep 2.0                               # translation step/Å
qstep 50.0                              # quaternion step/deg
dstep 50.0                              # torsion step/deg
torsdof 4 0.274000                      # torsional degrees of freedom and coefficient
rmstol 2.0                              # cluster_tolerance/Å
extnrg 1000.0                           # external grid energy
e0max 0.0 10000                         # max initial energy; max number of retries
ga_pop_size 150                         # number of individuals in population
ga_num_evals 2500000                    # maximum number of energy evaluations
ga_num_generations 27000                # maximum number of generations
ga_elitism 1                            # top individuals to survive to next generation
ga_mutation_rate 0.02                   # rate of gene mutation
ga_crossover_rate 0.8                   # rate of crossover
ga_window_size 10                       # Number of preceding generations to consider
ga_cauchy_alpha 0.0                     # Alpha parameter of Cauchy distribution
ga_cauchy_beta 1.0                      # Beta parameter Cauchy distribution
set_ga                                  # set the above parameters for GA or LGA
sw_max_its 300                          # iterations of Solis & Wets local search
sw_max_succ 4                           # consecutive successes before changing rho
sw_max_fail 4                           # consecutive failures before changing rho
sw_rho 1.0                              # size of local search space to sample
sw_lb_rho 0.01                          # lower bound on rho
ls_search_freq 0.06                     # probability of performing local search on individ.
set_sw1                                 # set the above Solis & Wets parameters
compute_unbound_extended                # compute extended ligand energy
ga_run 10                               # do this many hybrid GA-LS runs
analysis                                # perform a ranked cluster analysis
```

Figure A.6. An example of a docking parameter file used as input for AutoDock 4.

```
GOLD CONFIGURATION FILE
generated by gold front end (GOLD v3.2)

  POPULATION
popsiz = 100
select_pressure = 1.1
n_islands = 5
maxops = 100000
niche_siz = 2

  GENETIC OPERATORS
pt_crosswt = 95
allele_mutatewt = 95
migratewt = 10

  FLOOD FILL
cavity_file = list_of_residues.txt
floodfill_centre = list_of_residues

  DATA FILES
protein_datafile = /home/fbscs/GOLD/docking/1bm7.mol2
ligand_data_file /home/fbscs/GOLD/ligands/1bm7_ligandAC.mol2 10
param_file = DEFAULT
set_ligand_atom_types = 1
set_protein_atom_types = 0
directory = .
tordist_file = DEFAULT
make_subdirs = 0
save_lone_pairs = 1
fit_points_file = fit_pts.mol2
read_fitpts = 0

  FLAGS
display = 0
internal_ligand_h_bonds = 0
n_ligand_bumps = 0
flip_free_corners = 0
flip_amide_bonds = 0
flip_planar_n = 1 flip_ring_NRR flip_ring_NHR
flip_pyramidal_n = 0
rotate_carboxylic_oh = flip
use_tordist = 1

 TERMINATION
early_termination = 1
n_top_solutions = 3
rms_tolerance = 1.5

CONSTRAINTS
force_constraints = 0

  COVALENT BONDING
covalent = 0

  SAVE OPTIONS
save_score_in_file = 1
save_protein_torsions = 1

  FITNESS FUNCTION SETTINGS
initial_virtual_pt_match_max = 2.5
relative_ligand_energy = 0
score_param_file = DEFAULT
start_vdw_linear_cutoff = 4

 WATER DATA
water 3525    toggle   spin
water 3528    toggle   spin
```

Figure A.7. An example of a docking parameter file used as input for GOLD.

Section B
(Chapter 3)

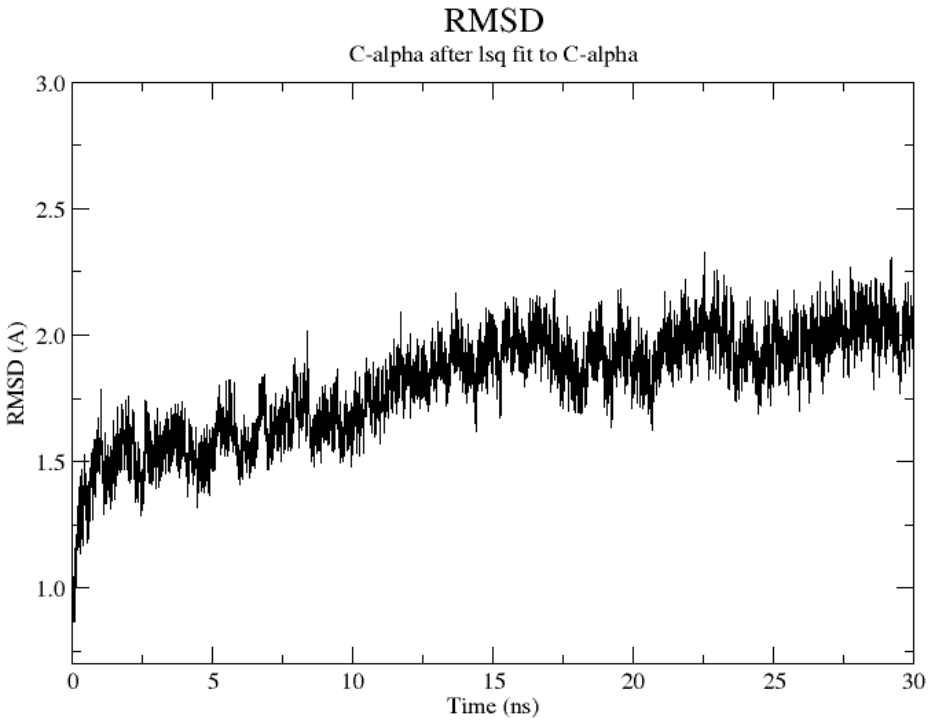


Figure B.1. Variation of the C_α atom root mean square deviation (RMSD) throught one 30-ns long equilibrium MD simulation. RMSDs were calculated over all residues except the N- and C-terminal tails, and with respect to the input tetrameric structure of TTR down-loaded from the PQS (PDB/PQS entry 2g9k). The trajectory was performed at 310 K and under NPT conditions.

Table B.1. Relative interaction energy values (in kcal.mol^{-1}) of the populated affinity minima for the various probe types. The points (minima) are distributed along 3 sub-binding pockets (BP) for each dimer comprising the TTR binding site (BP 1, 2, 3 for one dimer and 1',2',3' for the adjacent dimer), which are identified within parenthesis by their numerals.

Probe type	TTR apo (1f41)	TTR:T4 (2rox)	TTR:DES (1tt6)	TTR:FF (1bm7)	TTR:PCB (2g5u)
DRY (aromatic)	-1.89/-1.88 (2/2') -1.52/-1.52 (2/2') -2.13/-2.13 (3/3') -2.02/-2.02 (3/3') -1.54 (3*)	-3.03 (1*) -1.70/-1.76 (1/1') -1.76/-1.93 (2/2') -2.68/-1.78 (2/2') -2.43/-2.04 (3/3') -2.84/-1.85 (3/3')	-1.59 (1*) -1.69/-1.70 (2/2') -2.10/-2.08 (3/3') -2.09/-2.00 (3/3')	-1.63 (1*) -1.85/-1.85 (2/2') -1.70/-1.71 (2/2') -2.02/-2.02 (3/3') -2.16/-2.16 (3/3')	-1.91/-1.92 (2/2') -1.72/-1.72 (2/2') -2.12/-2.12 (3/3') -2.14/-2.08 (3/3') -1.56 (3*)
C3 (methyl)	-5.26/-5.34 (1/1') -3.18/-3.20 (1/1') -3.11/-3.10 (2/2')	-9.80/-5.35 (1/1') -4.75 (1*) -8.87/-4.52 (2/2') -4-.43/-4.44 (3/3')	-3.15 (1*) -4.86/-4.84 (1/1') -4.74/-4.75 (2/2') -3.69/-3.70 (3/3')	-3.43 (1*) -5.03/-5.03 (1/1') -3.13/-3.11 (2/2') -3.03/-3.03 (2/2') -3.29/-3.30 (3/3') -3.27/-3.27 (3/3')	-8.87/-8.39 (1/1') -5.35/-5.34 (2/2')
OH (phenolic or carboxy oxygen)	-9.94 (1*) -11.36/-11.60 (1/1') -12.01/-12.06 (3/3') -9.19/-9.20 (3/3')	-10.13 (1*) -10.85 (1) -11.52 (1) -9.23/-12.01 (3/3') -9.36 (3*)	-9.83 (1*) -9.68 (3*) -11.63/-11.57 (3/3')	-9.65/-9.65 (3/3')	-10.74 (1*) -9.03/-9.07 (3/3')
F (fluorine)	-3.12/-3.21 (1/1') -3.08/-3.08 (2/2') -3.20/-3.20 (3/3')	-5.61 (1) -3.15 (1*) -5.91 (2) -3.82 (3*)	-3.02/-3.02 (2/2') -3.20/-3.20 (2/2')	-3.19/-3.19 (2/2') -3.29/-3.30 (2/2') -3.15/-3.15 (3/3')	-3.10/-3.10 (1/1') -3.09/-3.09 (2/2') -3.28/-3.28 (2/2')
Cl (chlorine)	-7.93/-8.07 (1/1') -6.68/-6.68 (2/2')	-14.72/-7.86 (1/1') -7.24 (1*) -14.21/-7.17 (2/2') -6.61/-6.80 (3/3')	-7.13/-7.16 (1/1') -7.42/-7.40 (2/2')	-7.98/-7.98 (1/1')	-15.39/-15.26 (1/1') -13.03/-8.67 (2/2') -9.30/-8.86 (2/2')
I (iodine)	-12.73/-12.86 (1/1') -7.86/-7.86 (3/3')	-24.52/-13.75 (1/1') -19.40/-10.50 (2/2') -11.09/-10.41 (3/3') -16.90/-8.51 (3/3')	-12.35/-12.32 (1/1') -10.43/-10.43 (2/2') -9.18/-9.16 (3/3') -7.07/-7.08 (3/3')	-12.71/-12.72 (1/1') -7.07/-7.81 (3/3') -8.16/-8.17 (3/3')	-19.40/-17.37 (1/1') -13.75/-12.10 (3/3')
N2 (donor nitrogen)	-10.66 (1*) -14.11/-14.11 (1/1') -9.07/-9.10 (2/2') -11.07/-11.00 (3/3')	-14.60 (1*) -15.03/-10.81 (1/1')	-9.14/-9.15 (2/2')	-9.42/-9.41 (2/2') -10.64/-10.53 (3/3') -10.31/-10.36 (3/3')	-13.08 (1*) -10.69/-11.37 (1/1') -9.17/-9.17 (2/2')

Probe type	TTR apo (1f41)	TTR:T4 (2rox)	TTR:DES (1tt6)	TTR:FF (1bm7)	TTR:PCB (2g5u)
O (acceptor oxygen)	-9.20 (1*) -9.94/-10.67 (1/1') -12.18/-12.27 (3/3') -11.15/-11.25 (3/3')	-9.77 (1) -11.95 (2) -9.32 (2*) -12.71/-18.61 (3/3') -11.83 (3*)	-18.23 (3*)	-18.26 (3*) -9.95/-9.96 (3/3')	-12.11/-12.10 (3/3')
* indicates a midpoint.					

Section C

(Chapter 4)

```
#This file defines the rules for filtering multi-structure files based on
#properties and substructure patterns.

MIN_MOLWT 200 "Minimum molecular weight"
MAX_MOLWT 600 "Maximum molecular weight"
MIN_NUM_HVY 15 "Minimum number of heavy atoms"
MAX_NUM_HVY 35 "Maximum number of heavy atoms"
MIN_RING_SYS 0 "Minumum number of ring systems"
MAX_RING_SYS 5 "Maximum number of ring systems"
MIN_RING_SIZE 0 "Minimum atoms in any ring system"
MAX_RING_SIZE 20 "Maximum atoms in any ring system"
MIN_CON_NON_RING 0 "Minimum number of connected non-ring atoms"
MAX_CON_NON_RING 15 "Maximum number of connected non-ring atoms"
MIN_FCNGRP 0 "Minimum number of functional groups"
MAX_FCNGRP 18 "Maximum number of functional groups"
MIN_UNBRANCHED 0 "Minimum number of connected unbranched non-ring atoms"
MAX_UNBRANCHED 6 "Maximum number of connected unbranched non-ring atoms"
MIN_CARBONS 7 "Minimum number of carbons"
MAX_CARBONS 35 "Maximum number of carbons"
MIN_HETEROATOMS 2 "Minimum number of heteroatoms"
MAX_HETEROATOMS 20 "Maximum number of heteroatoms"
MIN_Het_C_Ratio 0.10 "Minimum heteroatom to carbon ratio"
MAX_Het_C_Ratio 1.0 "Maximum heteroatom to carbon ratio"
MIN_HALIDE_FRACTION 0.0 "Minimum Halide Fraction"
MAX_HALIDE_FRACTION 0.6 "Maximum Halide Fraction"    # This threshold was increased by 0.1,
following TTR actives Halide Fraction distribution

#count ring degrees of freedom = (#BondsInRing) - 4 - (RigidBondsInRing) - (BondsSharedWithOther
#must be >= 0, from JCAMD 14:251-265,2000.
ADJUST_ROT_FOR_RING true "BOOLEAN for whether to estimate degrees of freedom in rings"
MIN_ROT_BONDS 0 "Minimum number of rotatable bonds"
MAX_ROT_BONDS 20 "Maximum number of rotatable bonds"
MIN_RIGID_BONDS 0 "Minimum number of rigid bonds"
MAX_RIGID_BONDS 35 "Maximum number of rigid bonds"
MIN_HBOND_DONORS 0 "Minimum number of hydrogen-bond donors"
MAX_HBOND_DONORS 6 "Maximum number of hydrogen-bond donors"
MIN_HBOND_ACCEPTORS 0 "Minimum number of hydrogen-bond acceptors"
MAX_HBOND_ACCEPTORS 8 "Maximum number of hydrogen-bond acceptors"
MIN_LIPINSKI_DONORS 0 "Minimum number of hydrogens on O & N atoms"
MAX_LIPINSKI_DONORS 5 "Maximum number of hydrogens on O & N atoms"
MIN_LIPINSKI_ACCEPTORS 0 "Minimum number of oxygen & nitrogen atoms"
MAX_LIPINSKI_ACCEPTORS 10 "Maximum number of oxygen & nitrogen atoms"
MIN_COUNT_FORMAL_CRG 0 "Minimum number formal charges"
MAX_COUNT_FORMAL_CRG 3 "Maximum number of formal charges"
MIN_SUM_FORMAL_CRG -2 "Minimum sum of formal charges"
MAX_SUM_FORMAL_CRG 0 "Maximum sum of formal charges"    # This parameter was changed
from 2 to 0 (none of TTR actives as sum of formal charges > 0)
MIN_CHIRAL_CENTRES 0 "Minimum chiral centres"
MAX_CHIRAL_CENTRES 4 "Maximum chiral centres"
MIN_XLOGP -5.0 "Minimum XLogP"
MAX_XLOGP 6.5 "Maximum XLogP"    # This value was increased by 0.5, following TTR actives
XLogP distribution

#choices are insoluble<poorly<moderately<soluble<very<highly
MIN_SOLUBILITY moderately "Minimum solubility"  # This compensates for the higher XlogP
PSA_USE_SandP false "Count S and P as polar atoms"
MIN_2D_PSA 0.0 "Minimum 2-Dimensional (SMILES) Polar Surface Area"
MAX_2D_PSA 150.0 "Maximum 2-Dimensional (SMILES) Polar Surface Area"
AGGREGATORS true "Eliminate known aggregators"
PRED_AGG true "Eliminate predicted aggregators"
#secondary filters (based on multiple primary filters)
GSK_VEBER true "PSA>140 or >10 rot bonds"
MAX_LIPINSKI 1 "Maximum number of Lipinski violations"
```

MIN_ABS 0.5 "Minimum probability F>10% in rats"
PHARMACOPIA true "LogP > 5.88 or PSA > 131.6"
ALLOWED_ELEMENTS H,C,N,O,F,S,Cl,Br,I # Included Iodine, given TTR's high propensity to bind compounds bearing this element
ELIMINATE_METALS Sc,Ti,V,Cr,Mn,Fe,Co,Ni,Cu,Zn,Y,Zr,Nb,Mo,Tc,Ru,Rh,Pd,Ag,Cd

#acceptable molecules must have <= instances of each of the patterns below
#specific, undesirable functional groups
RULE 0 quinone
RULE 0 pentafluorophenyl_esters
RULE 0 paranitrophenyl_esters
RULE 0 HOBT_esters
RULE 0 triflates
RULE 0 lawesson_s_reagent
RULE 0 phosphoramides
RULE 0 beta_carbonyl_quat_nitrogen
RULE 0 acylhydrazide
RULE 0 cation_C_Cl_I_P_or_S
RULE 0 phosphoryl
RULE 0 alkyl_phosphate
RULE 0 phosphinic_acid
RULE 0 phosphanes
RULE 0 phosphoranes
RULE 0 imidoyl_chlorides
RULE 0 nitroso
RULE 0 N_P_S_Halides
RULE 0 carbodiimide
RULE 0 isonitrile
RULE 0 triacyloxime
RULE 0 cyanohydrins
RULE 0 acyl_cyanides
RULE 0 sulfonylnitrile
RULE 0 phosphonylnitrile
RULE 0 azocyanamides
RULE 0 beta_azo_carbonyl
RULE 0 polyenes
RULE 0 saponin_derivatives
RULE 1 cytochalasin_derivatives
RULE 4 cycloheximide_derivatives
RULE 1 monensin_derivatives
RULE 1 squalestatin_derivatives

#functional groups which often eliminate compounds from consideration
RULE 0 acid_halide
RULE 0 aldehyde
RULE 0 alkyl_halide
RULE 0 anhydride
RULE 0 azide
RULE 0 azo
RULE 0 di_peptide
RULE 0 michael_acceptor
RULE 0 beta_halo_carbonyl
RULE 0 nitro
RULE 0 oxygen_cation
RULE 0 peroxide
RULE 0 phosphonic_acid
RULE 0 phosphonic_ester
RULE 0 phosphoric_acid
RULE 0 phosphoric_ester
RULE 0 sulfonic_acid
RULE 0 sulfonic_ester
RULE 0 tricarbo_phosphene
RULE 0 epoxide
RULE 0 sulfonyl_halide
RULE 0 halopyrimidine
RULE 0 perhalo_ketone
RULE 0 aziridine
RULE 1 oxalyl
RULE 0 alphahalo_amine
RULE 0 halo_amine
RULE 0 halo_alkene

```
RULE 0 acyclic_NCN
RULE 0 acyclic_NS
RULE 0 SCN2
RULE 0 terminal_vinyl
RULE 0 hetero_hetero
RULE 0 hydrazine
RULE 0 N_methoyl
RULE 0 NS_beta_halothyl
RULE 0 propiolactones
RULE 0 nitroso
RULE 0 iodoso
RULE 0 iodoxy
RULE 0 noxide

#groups of molecules
RULE 0 dye

#functional groups which are allowed, but may not be wanted in high quantities
#common functional groups
RULE 6 alcohol
RULE 4 alkene
RULE 4 amide
RULE 4 amino_acid
RULE 2 amine
RULE 4 primary_amine
RULE 4 secondary_amine
RULE 4 tertiary_amine
RULE 2 carboxylic_acid
RULE 6 halide
RULE 0 iodine
RULE 2 ketone
RULE 4 phenol
RULE 1 imine
RULE 1 methyl_ketone
RULE 1 alkylaniline
RULE 4 sulfonamide
RULE 1 sulfonylurea
RULE 0 phosphonamide
RULE 0 alphahalo_ketone
RULE 0 oxaziridine
RULE 1 cyclopropyl
RULE 2 guanidine
RULE 0 sulfonimine
RULE 0 sulfinimine
RULE 1 hydroxamic_acid
RULE 0 phosphoryl
RULE 0 sulfinylthio
RULE 0 disulfide
RULE 0 enol_ether
RULE 0 enamine
RULE 0 organometallic
RULE 0 dithioacetal
RULE 1 oxime
RULE 0 isothiocyanate
RULE 0 isocyanate
RULE 3 lactone
RULE 3 lactam
RULE 1 thioester
RULE 1 carbonate
RULE 0 carbamic_acid
RULE 1 thiocarbamate
RULE 0 triazine
RULE 1 malonic

#other functional groups
RULE 2 alkyne
RULE 4 aniline
RULE 4 aryl_halide
RULE 2 carbamate
RULE 3 ester
RULE 5 ether
```

```
RULE 1 hydrazone
RULE 0 nonacylhydrazone
RULE 1 hydroxylamine
RULE 2 nitrile
RULE 2 sulfide
RULE 2 sulfone
RULE 2 sulfoxide
RULE 0 thiourea
RULE 1 thioamide
RULE 1 thiol
RULE 2 urea
RULE 0 hemiketal
RULE 0 hemiacetal
RULE 0 ketal
RULE 1 acetal
RULE 0 aminal
RULE 0 hemiaminal

#protecting groups
RULE 0 benzyloxycarbonyl_CBZ
RULE 0 t_butoxycarbonyl_tBOC
RULE 0 fluorenylmethoxycarbonyl_Fmoc
RULE 1 dioxolane_5MR
RULE 1 dioxane_6MR
RULE 1 tetrahydropyran_THP
RULE 1 methoxyethoxymethyl_MEM
RULE 2 benzyl_ether
RULE 2 t_butyl_ether
RULE 0 trimethylsilyl_TMS
RULE 0 t_butyldimethylsilyl_TBDMS
RULE 0 triisopropylsilyl_TIPS
RULE 0 t_butyldiphenylsilyl_TBDPS
RULE 1 phthalimides_PHT
RULE 2 arenesulfonyl
```

Figure C.1. FILTER definitions used to trim down the initial library downloaded from the ZINC database.

Section D

(Chapter 5)

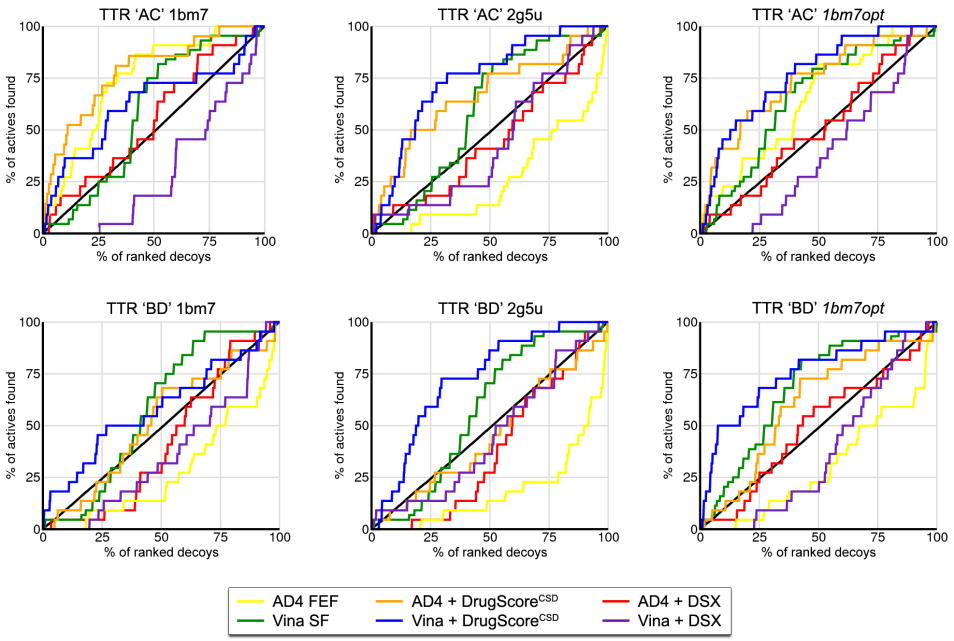


Figure D.1. ROC curves comparing the VS performance of six docking and scoring protocols against TTR. The results obtained with the 1bm7 X-ray structure are shown in the left panels, the middle panels show the results structure 2g5u, and the right panels correspond to the results of the *1bm7opt* model. In all cases, the results obtained for the "AC" binding site of the protein are in the top panels and the bottom panels show the results for the "BD" binding site. The protocols are coloured as follows: AD4 Free Energy Function (yellow), AD4 + DrugScore^{CSD} (orange), AD4 + DSX (red), Vina Scoring Function (green), Vina + DrugScore^{CSD} (blue), Vina + DSX (purple), and random performance (black).

Section E
(Chapter 6)

Table E.1. Summary description of nuclear hormone receptors in the DUD.

Target	PDB code	Resolution (Å)	Description	No. of actives[a]	No. of decoys[a]	No. of *lead-like* actives[a]	No. of clusters (actives)
AR	1z95	1.80	Androgen receptor in complex R-bicalutamide	74 (74/74)	2630 (2628/2628)	63 (63/63)	10
ER agonist	1l2i	1.95	Estrogen receptor alpha ligand-binding domain in complex with tetrahydrochrysene derivative	67 (67/67)	2361 (2355/2355)	63 (63/63)	10
ER antagonist	3ert	1.90	Estrogen receptor alpha ligand-binding domain in complex with 4-hydroxytamoxifen	39 (39/39)	1399 (1395/1395)	18 (18/18)	8
GR	1m2z	2.50	Human glucocorticoid receptor ligand-binding domain bound to dexamethasone and a TIF2 coactivator motif	78 (78/78)	2804 (2804/2804)	32 (32/32)	2
MR	2aa2	1.95	Mineralocorticoid receptor in complex with aldosterone	15 (15/15)	535 (535/535)	13 (13/13)	2
PPAR gamma	1fm9	2.10	Human PPAR gamma ligand binding domain bound with gi262570 and a co-activator peptide	81 (81/81)	2910 (2906/2905)	7 (6/6)	6
PR	1sr7	1.46	Progesterone receptor hormone binding domain with bound mometasone furoate	27 (27/27)	967 (967/966)	22 (22/22)	4
RXR alpha	1mvc	1.90	Retinoid X receptor alpha ligand-binding domain in complex with the BMS 649 agonist	20 (20/20)	708 (708/708)	18 (18/18)	3

[a] The numbers within brackets correspond to the number of compounds that have been successfully docked with AD4/Vina, respectively.

Table E.2. Summary description of protein kinases in the DUD.

Target	PDB code	Resolution (Å)	Description	No. of actives[a]	No. of decoys[a]	No. of *lead-like* actives[a]	No. of clusters (actives)
CDK2	1ckp	2.05	Human cyclin dependent kinase 2 complexed with the inhibitor purvalanol B	50 (50/48)	1780 (1779/1779)	55 (47/47)	32
EGFr	1m17	2.60	Epidermal growth factor receptor tyrosine kinase domain with the inhibitor erlotinib	458 (444/444)	14894 (14894/14894)	379 (365/365)	40
FGFr1	1agw	2.40	Tyrosine kinase domain of fibroblast growth factor receptor 1 in complex with the su4984 inhibitor	118 (118/118)	4216 (4205/4205)	73 (71/71)	12
HSP90	1uy6	1.90	Human heat shock protein 90 kinase in complex with a purine-6-ylamine analogue	24 (24/24)	860 (860/860)	24 (23/23)	4
P38 MAP	1kv2	2.80	Human p38 MAP kinase in complex with BIRB 796	353 (256/256)	8399 (8387/8387)	219 (137/137)	20
PDGFrb	model	n/a	Platlet derived growth factor receptor kinase	169 (157/157)	5625 (5614/5614)	136 (124/124)	22
SRC	2src	1.50	C-src tyrosine kinase in complex with AMP-PNP	159 (155/155)	5801 (5793/5793))	102 (98/98)	21
TK	1kim	2.14	Thymidine kinase from type-1 HSV in complex with deoxythymidine	22 (22/22)	785 (783/783)	22 (22/22)	7
VEGFr2	1vr2	2.40	Human vascular endothelial growth factor receptor 2 kinase	78 (74/73)	2647 (2641/2641)	49 (48/47)	31

[a] The numbers within brackets correspond to the number of compounds that have been successfully docked with AD4/Vina, respectively.

Table E.3. Summary description of serine proteases in the DUD.

Target	PDB code	Resolution (Å)	Description	No. of actives[a]	No. of decoys[a]	No. of *lead-like* actives[a]	No. of clusters (actives)
FXa	1f0r	2.10	Human coagulation factor Xa in complex with RPR208815	146 (142/142)	5102 (5095/5095)	n/a	19
Thrombin	1ba8	1.80	Thrombin bound to an inhibitor with a rigid tripeptidyl aldehydes	68 (65/65)	2294 (2294/2294)	26 (24/24)	14
Trypsin	1bju	1.80	Beta-trypsin in complex with ACPU	46 (44/44)	1545 (1545/1545)	10 (9/9)	7

[a] The numbers within brackets correspond to the number of compounds that have been successfully docked with AD4/Vina, respectively.

Table E.4. Summary description of metalloenzymes in the DUD.

Target	PDB code	Resolution (Å)	Description	No. of actives[a]	No. of decoys[a]	No. of *lead-like* actives[a]	No. of clusters (actives)
ACE	1o86	2.00	Human angiotensin converting enzyme in complex with lisinopril	49 (49/49)	1728 (1727/1727)	46 (46/46)	18
ADA	1ndw	2.00	Adenosine deaminase in complex FR221647	23 (23/23)	822 (821/821)	23 (23/23)	8
COMT	1h1d	2.00	Catechol-o-methyltransferase complex with co-substrate and a novel inhibitor	12 (11/11)	430 (430/430)	11 (11/11)	2
PDE5	1xp0	1.89	Catalytic domain of human phosphodiesterase 5A in complex with vardenafil	51 (51/51)	1810 (1809/1809)	34 (26/26)	22

[a] The numbers within brackets correspond to the number of compounds that have been successfully docked with AD4/Vina, respectively.

Table E.5. Summary description of folate enzymes in the DUD.

Target	PDB code	Resolution (Å)	Description	No. of actives[a]	No. of decoys[a]	No. of *lead-like* actives[a]	No. of clusters (actives)
DHFR	3dfr	1.70	Dihydrofolate reductase in complex with methotrexate	201 (201/201)	7150 (7145/7145)	191 (191/191)	14
GART	1c2t	2.10	GAR transformylase in complex with beta-GAR-10-formyl-TDAF	21 (21/21)	753 (753/753)	13 (8/8)	5

[a] The numbers within brackets correspond to the number of compounds that have been successfully docked with AD4/Vina, respectively.

Table E.6. Summary description of other enzymes in the DUD.

Target	PDB code	Resolution (Å)	Description	No. of actives[a]	No. of decoys[a]	No. of *lead-like* actives[a]	No. of clusters (actives)
AChE	1eve	2.50	Acetylcholinesterase in complex with the anti-Alzheimer drug E2020 (Aricept)	106? 105 (105/105)	3732 (3714/3714)	100 (100/100)	18
ALR2	1ah3	2.30	Aldose reductase in complex with tolrestat	26 (26/26)	920 (918/918)	26 (26/26)	14
AmpC	1xgj	1.97	AmpC beta-lactamase in complex with 3-(4-carboxy-2-hydroxy-phenylsulfamoyl)-thiophene-2-carboxylic acid	21 (21/21)	734 (732/732)	21 (21/21)	6
COX1	1q4g	2.00	Ovine prostaglandin H2 synthase-1 (cox1) in complex with alpha-methyl-4-biphenylacetic acid	25 (25/25)	850 (849/849)	23 (23/23)	11
COX2	1cx2	3.00	Prostaglandin synthase-2 (cox2) in complex with a SC-558	412? 349 (348/348)	12491 (12464/12464)	250 (212/212)	44
GPB	1a8i	1.78	Glycogen phosphorylase in complex with spirohydantoin	52 (52/52)	1851 (1848/1848)	52 (52/52)	10
HIVPR	1hpx	2.00	HIV-1 protease in complex with KNI-272 (inhibitor)	62? 53 (53/53)	1888 (1885/1885)	6 (4/4)	3
HIVRT	1rt1	2.55	HIV-1 reverse transcriptase in complex with MKC-442	41? 40 (40/40)	1439 (1437/1437)	35 (34/34)	17
HMGR	1hw8	2.10	HMG-CoA reductase in complex with compactin (aka mevastatin)	35 (35/35)	1242 (1241/1241)	25 (25/25)	4
InhA	1p44	2.70	Enoyl reductase in complex with a piperazine-indol	86? 85 (85/85)	3043 (3035/3035)	58 (57/57)	23
NA	1a4g	2.20	Neuraminidase (influenza virus B/Beijing/1/87) in complex with zanamivir	49 (49/49)	1745 (1745/1745)	49 (49/49)	7
PARP	1efy	2.20	Poly(ADP-ribose) polymerase (PARP) in complex with benzimidazole inhibitor	35? 33 (33/33)	1178 (1176/1176)	33 (31/31)	7
PNP	1b8o	1.50	Purine nucleoside phosphorylase complexed with a transition-state analog	30? 25 (25/25)	884 (882/882)	30 (25/25)	4
SAHH	1a7a	2.80	S-adenosylhomocysteine hydrolase in complex with nicotinamide-adenine-dinucleotide	33 (33/33)	1159 (1158/1158)	33 (33/33)	2

[a] The numbers inside the brackets correspond to the number of compounds that have been successfully docked with AD4/Vina, respectively.

Table E.7. Statistical results for the docking and scoring protocols using DUD, broken down by protein family.

	AutoDock 4 FEF [a]		AD4 + Drug-Score[CSD][b]	AD4 + DSX [c]			Vina SF	Vina + Drug-Score[CSD][b]	Vina + DSX [c]		
	LE	LC		total	PAS	PCS			total	PAS	PCS
Mean ROC AUC[d]											
Entire DUD	**0.58**	0.57	0.55	0.53	0.54	**0.59**	0.57	0.54	0.54	0.56	0.54
NHRs	**0.74**	0.73	0.65	0.62	0.59	0.64	0.68	0.67	0.63	0.62	0.62
Kinases	0.47	0.48	0.54	0.45	0.45	0.53	**0.58**	0.54	0.51	0.49	0.43
Serine Proteases	**0.65**	0.63	0.51	0.57	0.52	0.60	0.62	0.50	0.56	0.52	0.58
Metalloenzymes	0.52	0.51	**0.56**	0.54	0.57	**0.60**	0.45	0.53	0.52	0.57	0.50
Folate enzymes	**0.77**	0.64	–	0.69	0.72	0.64	0.64	–	0.71	**0.77**	0.57
Other enzymes	**0.55**	0.52	0.47	0.50	0.54	**0.59**	0.51	0.45	0.50	0.55	0.57
Mean/median ROCE at 0.5% FPR											
Entire DUD	4.9/ 13.2	2.3/ 14.9	4.9 / 9.7	4.9/ 13.8	4.8/ 15.8	3.5/ 15.7	6.5/ 22.89	**17.0 / 38.7**	16.2/ 26.5	2.1/ 16.8	0.0/ 10.2
NHRs	18.6/ 20.1	14.0/ 32.3	20.8 / 24.6	11.0/ 14.6	23.6/ 49.7	47.3/ 49.8	33.9/ 65.1	**42.0 / 55.7**	21.3/ 35.4	14.4/ 39.4	17.3/ 15.9
Kinases	2.0/ 7.6	4.0/ 6.5	4.5 / 6.9	3.3/ 5.7	2.0/ 1.8	0.6/ 2.5	0.8/ 9.4	**17.0 / 37.5**	**14.0/ 18.3**	0.0/ 0.8	0.0/ 0.5
Serine Proteases	7.5/ 10.6	15.0/ 10.7	2.6 / 2.6	5.3/ 4.3	5.5/ 4.9	8.8/ 9.2	5.5/ 8.4	4.4 / 4.4	3.8/ 7.5	5.5/ 6.1	0.0/ 0.6
Metalloenzymes	9.0/ 15.8	3.7/ 13.1	2.5 / 2.5	2.5/ 4.5	16.8/ 20.9	0.0/ 1.2	7.4/ 14.9	17.3 / 17.3	**19.8/ 21.2**	9.7/ 15.6	0.0/ 10.8
Folate enzymes	17.7/ 17.7	9.0/ 9.0	–	9.8/ 9.8	19.8/ 19.8	0.4/ 0.4	11.9/ 11.9	–	12.3/ 12.3	**28.8/ 28.8**	0.0/ 0.0
Other enzymes	0.6/ 12.6	0.0/ 11.8	0.0 / 1.4	2.2/ 23.8	4.6/ 5.8	5.5/ 12.5	2.4/ 14.4	**6.0 / 37.0**	12.8/ 34.3	0.0/ 15.2	0.0/ 16.6
Mean/median ROCE at 1% FPR											
Entire DUD	4.4/ 15.6	3.5/ 16.7	3.1 / 16.2	3.1/ 8.0	3.4/ 11.4	3.2/ 9.7	4.7/ 21.5	8.0 / 14.9	**8.8/ 16.1**	3.4/ 9.3	1.0/ 5.8
NHRs	14.4/ 21.7	17.4/ 28.1	16.8 / 51.6	6.1/ 10.1	15.4/ 35.1	18.0/ 26.2	**30.8/ 78.1**	24.0 / 24.6	11.3/ 31.0	9.3/ 23.0	7.5/ 7.9
Kinases	1.8/ 3.4	1.7/ 3.4	2.7 / 3.7	1.8/ 4.0	1.2/ 1.7	0.9/ 1.5	0.9/ 5.5	2.8 / 11.6	**8.0/ 9.3**	0.4/ 1.2	0.2/ 0.6
Serine Proteases	**7.6/ 7.1**	7.6/ 5.6	1.6 / 1.6	3.1/ 4.4	2.4/ 3.3	5.0/ 4.9	2.4/ 5.6	4.3 / 4.3	3.4/ 3.8	3.1/ 3.0	0.0/ 0.5
Metalloenzymes	4.4/ 5.9	4.9/ 6.2	2.2 / 2.2	2.2/ 3.5	11.0/ 17.4	1.1/ 1.6	6.2/ 14.4	10.0 / 10.0	**13.4/ 18.0**	12.4/ 11.6	2.2/ 6.5
Folate enzymes	11.3/ 11.3	9.5/ 9.5	—	6.7/ 6.7	9.5/ 9.5	0.2/ 0.2	6.9/ 6.9	—	10.0/ 10.0	**18.8/ 18.8**	0.0/ 0.0
Other enzymes	2.9/ 25.1	0.8/ 25.2	1.0/ 2.1	2.6/ 11.6	2.9/ 4.3	4.7/ 10.1	2.3/ 7.0	3.2 / 12.7	**9.8/ 14.8**	1.9/ 6.0	1.4/ 9.6
Mean/median ROCE at 2% FPR											
Entire DUD	4.3/ 8.5	3.2/ 7.6	2.6 / 5.6	2.5/ 5.1	2.4/ 4.5	2.3/ 5.9	3.4/ 8.0	**5.6/ 9.7**	6.7/ 9.1	2.6/ 6.1	1.0/ 3.8

	AutoDock 4 FEF [a]		AD4 + DrugScore^{CSD} [b]	AD4 + DSX [c]			Vina SF	Vina + DrugScore^{CSD} [b]	Vina + DSX [c]		
	LE	LC		total	PAS	PCS			total	PAS	PCS
NHRs	11.8/ 19.2	11.3/ 19.8	11.4 / 14.2	6.3/ 9.2	9.2/ 12.3	12.5/ 15.5	14.9/ 22.6	**14.6/ 19.0**	9.3/ 17.7	12.5/ 14.8	5.2/ 7.7
Kinases	1.0/ 2.5	2.4/ 2.8	2.2 / 2.6	2.3/ 3.0	1.0/ 1.5	1.6/ 1.6	1.4/ 3.1	**5.2/ 5.9**	4.6/ 5.3	1.3/ 2.6	0.8/ 1.0
Serine Proteases	5.6/ 4.9	**8.6/ 6.3**	1.3 / 1.3	2.4/ 3.4	2.2/ 2.5	2.6/ 3.1	1.1/ 4.6	3.3/ 3.3	5.1/ 3.8	1.5/ 1.7	0.8/ 0.6
Metalloenzymes	3.8/ 3.1	2.6/ 4.1	4.0 / 4.0	4.8/ 3.8	7.3/ 6.7	2.2/ 1.9	4.5 / 7.5	6.3/ 6.3	**8.7/ 8.2**	6.3 / 5.3	1.6/ 2.7
Folate enzymes	**10.7/ 10.7**	9.1/ 9.1	—	3.6/ 3.6	**10.2/ 10.2**	1.8/ 1.8	5.1/ 5.1	—	8.4/ 8.4	12.3/ 12.3	0.0/ 0.0
Other enzymes	2.3/ 8.4	1.9/ 6.9	1.0 / 1.6	1.1/ 2.1	1.9/ 2.4	3.8/ 5.5	2.0/ 4.1	2.1 / 7.1	**6.6/ 8.3**	2.5/ 4.6	1.6/ 5.0
Mean/median ROCE at 5% FPR											
Entire DUD	3.5/ 4.4	2.0/ 3.8	3.0/ 7.8	2.3/ 3.1	2.1/ 3.1	2.0/ 3.0	2.9 /3.9	4.5/ 8.8	3.8/ 4.3	2.0/ 3.2	1.4/ 2.3
NHRs	10.3/ 9.6	6.9/ 8.3	7.3/ 6.7	4.4/ 5.5	4.7/ 5.6	5.3/ 6.2	6.8 / 7.6	8.3/ 8.2	7.4/ 7.5	5.2/ 5.9	3.7/ 4.1
Kinases	0.9/ 1.8	1.9/ 2.0	1.8/ 1.7	1.3/ 1.5	1.3/ 1.3	1.3/ 1.4	1.6 / 2.4	2.8/ 3.0	3.3/ 3.0	1.4/ 1.2	0.8/ 0.9
Serine Proteases	5.7/ 4.4	4.5/ 3.4	0.8/ 0.8	1.4/ 2.1	1.6/ 2.0	2.0/ 2.3	1.9/ 2.8	1.8 / 1.8	1.9/ 1.8	0.8/ 1.1	0.6/ 1.2
Metalloenzymes	2.3/ 2.1	1.8/ 1.9	11.4/ 11.7	2.2/ 2.4	3.6/ 3.5	1.7/ 1.7	3.5/ 3.3	**12.3 / 12.3**	4.6 / 4.1	4.4/ 3.6	1.5/ 1.7
Folate enzymes	7.1/ 7.1	4.2/ 4.2	—	4.3/ 4.3	8.8/ 8.8	1.6/ 1.6	4.5 / 4.5	—	6.1 / 6.1	7.6/ 7.6	1.5/ 1.5
Other enzymes	1.3/ 3.3	1.5/ 2.9	3.1/ 10.2	2.4/ 3.1	2.2/ 2.2	2.9/ 2.9	1.8/ 3.0	7.9 / 11.0	3.1/ 3.5	2.0/ 2.7	1.8/ 2.8

[a] For AutoDock 4 free energy function, both ranking by lowest energy conformation scores and by largest cluster conformation (grey shade) scores are reported. [b] Results for DrugScore^{CSD} were based only on 30 DUD targets, due to failures while handling protein metals/co-factors. [c] For DSX, besides the total score, ranking by per-atom scores and per-contact scores are compared (grey shades). [d] A mean ROC AUC of 0.5 represents random performance.

Section F
(Chapter 7)

Table F.1. Median values for molecular properties predicted for TTR actives, composite templates, decoy molecules and top-100 VS hits.

VS protocol / data set	*MW (Da)*	*Cs (#)*	*HbA /D (#)*	*Rs (#)*	*Xlo gP*	*∑±*	*PSA (Å²)*	*Hal. (÷)*	*Solubility [a]*						*Aggregator [b]*	
									I	*P*	*M*	*S*	*V*	*H*	*P*	*F*
Actives	301	14	2/0	2.5	3.97	-1.0	60	0.22	6	7	5	6	2	0	18	8
Composite Templates	451	21	2/0	4.5	5.20	-3.0	132	0.16	0	0	1	2	1	0	0	4
Decoys	331	15	3/1	3.0	4.24	0.0	84	0.00	23	89	161	235	164	65	201	536
2D CHF Euclidean Phenox	252	14	4/1	3.8	2.25	-0.6	64	0.01	0	0	3	10	75	12	100	0
2D CHF Euclidean PCB18	221	13	2/2	2.0	2.89	-0.1	46	0.00	0	0	3	31	47	19	100	0
2D CHF Euclidean Thyroxine	223	12	4/1	5.0	-1.10	-0.2	82	0.00	0	0	0	3	71	26	100	0
2D CHF Tanimoto PCB18	254	14	3/1	2.7	2.89	-0.1	49	0.14	1	0	12	49	30	8	100	0
2D FP Tanimoto PCB18	265	14	4/0	2.0	1.9	-0.5	58	0.00	0	0	28	40	32	0	99	1
2D UNITY Tan. Phenox	307	16	5/1	6.0	2.3	-0.5	73	0.00	0	0	5	7	69	19	100	0
2D UNITY Tan. PCB18	248	13	2/1	2.6	3.19	-0.1	40	0.16	1	0	16	52	28	3	100	0
ROCS Combo Phenox	366	19	6/1	3.0	2.63	-0.1	76	0.02	0	3	20	38	36	3	97	3
ROCS Combo PCB18	268	13	6/1	3.0	1.64	0.0	74	0.02	0	0	7	15	62	16	97	3
ROCS ScaledColor Phenox	315	15	6/1	5.2	1.89	-0.9	88	0.01	0	1	19	39	27	14	100	0
ROCS ScaledColor Thyroxine	316	16	5/1	5.0	1.07	0.0	93	0.00	0	0	11	35	41	13	99	1
EON ET-combo Phenox	330	16	6/1	6.2	2.38	-1.9	109	0.02	0	0	5	24	53	18	100	0
LigMatch 2D + 3D 4 templates	315	16	5/1	5	1.22	-0.1	76	0.00	0	4	20	42	32	2	100	0
2D CHF Euclidean DB47OH	254	15	4/0	3.5	2.79	0.0	63	0.00	0	0	6	17	71	6	100	0
2D CHF Euclidean DB47COOH	254	15	4/0	3.0	2.78	0.0	63	0.00	0	0	5	18	70	6	99	1

VS protocol / data set	*MW (Da)*	*Cs (#)*	*HbA/D (#)*	*Rs (#)*	*XlogP*	*∑±*	*PSA (Å²)*	*Hal. (÷)*	*Solubility [a]*						*Aggregator [b]*	
									I	*P*	*M*	*S*	*V*	*H*	*P*	*F*
EON ET-combo PhenoxOH	350	17	7/1	7.0	2.50	-2.0	119	0.00	0	5	16	26	37	16	99	1
EON ET-combo Phenox-COOH	359	17	7/1	7.0	2.62	-2.0	119	0.00	0	5	8	29	42	16	98	2
FRED CG3	394	21	6/1	6	2.89	0.0	80	0.00	0	0	18	35	34	13	99	1
FRED ChemScore	302	16	5/3	4	2.05	0.0	74	0.00	0	0	12	32	49	7	99	1

Table F.2. Median and maximum scores assigned by the DrugScoreCSD (DS) and AutoDock 4 (AD4) scoring functions to the complexes of TTR with actives, composite templates, decoy molecules and the top-100 ranked compounds retrieved by each VS protocol. Median and maximum number of H-bonds (Hb) established within the TTR receptor sites are also given. Additionally, for each series of ligands, the number of hits detected by different 3D pharmacophore models (and combinations) is reported.

VS protocol / data set		*Docking and Scoring*						*Pharmacophore Searches*					
	Median / *Max.*	*DS AC*	*DS BD*	*AD4 AC*	*AD4 BD*	*Hb AC*	*Hb BD*	*QRY AC1*	*QRY AC2*	*QRY BD1*	*QRY BD2*	*QRY AC1/BD1*	*QRY AC1/BD2*
Actives		-117354	-91660	-7.91	-6.31	2	0	7	10	10	8	3	2
		-173603	-163629	-9.66	-9.05	4	1						
Composite Templates		-165206	-131603	-10.10	-7.53	2	0	3	0	0	0	3	3
		-203773	-145423	-11.06	-8.18	4	2						
Decoys		-68696	NA	-6.70	-7.23	2	2	6	10	2	9	0	0
		-171697	NA	-9.80	-11.22	4	4						
2D CHF Euclidean Phenox		-82363	-82763	-6.56	-6.57	2	0	6	4	7	4	3	0
		-147833	-180996	-8.69	-7.98	5	2						
2D CHF Euclidean PCB18		-93670	-97633	-5.93	-6.53	2	0	3	15	14	20	3	3
		-162406	-156750	-8.16	-9.24	5	1						
2D CHF Euclidean Thyroxine		-41053	-95593	-5.90	-6.20	2	0	0	1	4	0	0	0
		-97954	-140636	-7.80	-7.75	6	2						
2D CHF Tanimoto PCB18		-84482	-88868	-6.19	-6.39	2	0	1	14	10	16	0	1
		-162677	-154244	-8.08	-8.65	3	1						
2D PF Tanimoto PCB18		-80568	-61867	-7.11	-6.80	2	0	1	31	6	13	0	0
		-157548	-152155	-8.95	-8.94	2	2						
2D UNITY Tanimoto Phenox		-90557	-120846	-6.93	-6.89	2	0	3	1	1	0	0	0
		-173395	-171487	-8.98	-8.63	5	2						
2D UNITY Tanimoto PCB18		-84968	-88129	-5.93	-6.35	2	0	3	22	19	28	3	3
		-161473	-157423	-8.01	-8.57	4	0						
ROCS Combo Phenox		-74725	-74437	-7.36	-7.00	0	0	5	1	0	0	0	0
		-142131	-142471	-9.10	-9.76	5	2						
ROCS Combo PCB18		-95663	-92320	-6.35	-6.82	2	0	0	8	4	4	0	0
		-154217	-147791	-8.49	-9.42	3	2						
ROCS ScaledColor Phenox		-88949	-86073	-7.59	-7.05	2	0	6	3	0	0	0	0
		-145382	-182022	-9.55	-9.17	4	2						
ROCS ScaledColor		-92167	-86938	-7.09	-6.98	2	0	6	3	9	0	1	0

VS protocol / data set		*Docking and Scoring*						*Pharmacophore Searches*					
	Median / *Max.*	*DS AC*	*DS BD*	*AD4 AC*	*AD4 BD*	*Hb AC*	*Hb BD*	*QRY AC1*	*QRY AC2*	*QRY BD1*	*QRY BD2*	*QRY AC1/BD1*	*QRY AC1/BD2*
Thyroxine		-156931	-145906	-9.71	-9.36	4	1						
EON ETcombo Phenox		-78534	-85783	-7.72	-6.73	2	0	12	1	2	0	0	0
		-159861	-173793	-9.79	-8.88	4	2						
LigMatch 2D + 3D 4 templates		-80645	-98435	-7.06	-7.04	2	0	6	0	0	0	0	0
		-156793	-159928	-8.69	-9.53	4	2						
2D CHF Euclidean DBF47OH		-85827	-99648	-6.66	-6.26	2	0	1	0	1	0	0	0
		-163657	-146208	-8.29	-7.93	4	2						
2D CHF Euclidean DBF47COOH		-86120	-97346	-6.65	-6.20	2	0	1	0	1	0	0	0
		-155679	-145208	-9.07	-7.92	4	2						
EON ETcombo PhenoxOH		-74052	-85303	-7.34	-6.49	2	0	5	0	2	0	0	0
		-143950	-157417	-9.94	-9.04	5	2						
EON ETcombo PhenoxCOOH		-72830	-83665	-7.36	-6.86	2	0	11	0	0	0	0	0
		-162931	-163566	-9.83	-9.76	4	1						
FRED CG3		-98122	-98294	-7.88	-8.05	2	0	1	0	0	0	0	0
		-187325	-203683	-9.81	-9.76	4	1						
FRED ChemScore		-103904	-122752	-6.53	-7.69	2	0	0	4	10	1	0	0
		-167382	-182318	-8.92	-8.89	4	2						

References and Notes

1. Foye WO (1989) Principles of Medicinal Chemistry. 3rd ed. Philadelphia: Lea & Febiger. p.
2. Venkatesh S, Lipper RA (2000) Role of the development scientist in compound lead selection and optimization. J Pharm Sci 89: 145–154. doi:10.1002/(SICI)1520-6017(200002)89:2<145::AID-JPS2>3.0.CO;2-6.
3. DiMasi J a, Hansen RW, Grabowski HG (2003) The price of innovation: new estimates of drug development costs. J Health Econ 22: 151–185. doi:10.1016/S0167-6296(02)00126-1.
4. McGee P (2005) Modeling success with in silico tools. Drug Discov Today 8: 23–28.
5. Pang YP (2007) In Silico Drug Discovery: Solving the "'Target-rich and Lead-poor'" Imbalance Using the Genome-to-drug-lead Paradigm. Clinical Pharmacology & Therapeutics 81. doi:10.1038/sj.clpt.6100030.
6. Fischer E (1894) Einfluss der configuration auf die wirkung der enzyme. Ber Dt Chem Ges 27: 2985–2993.
7. Koshland DE (1958) Application of a theory of enzyme specificity to protein synthesis. Proc Natl Acad Sci USA 44: 98–104.
8. Foote J, Milstein C (1994) Conformational isomerism and the diversity of antibodies. Proc Natl Acad Sci USA 91: 10370–10374.
9. Martin YC (1978) Quantitative drug design: a critical introduction. Marcel Dekker I, editor New York. p.
10. Bourne PE, Weissig H, editors (2003) Structural Bioinformatics. Hoboken, NJ, USA: John Wiley & Sons, Inc. p. doi:10.1002/0471721204.
11. Böhm H-J, Klebe G (1996) What Can We Learn from Molecular Recognition in Protein–Ligand Complexes for the Design of New Drugs? Angew Chem Int Ed Engl 35: 2588–2614. doi:10.1002/anie.199625881.
12. Holdgate GA, Ward WHJ (2005) Measurements of binding thermodynamics in drug discovery. Drug Discov Today 10: 1543–1550.
13. Leach AR (2001) Molecular modelling: principles and applications. 2nd ed. Dorchester: Pearson Education. p.
14. Van Gunsteren WF, Bakowies D, Baron R, Chandrasekhar I, Christen M, et al. (2006) Biomolecular Modeling: Goals, Problems, Perspectives. Angew Chem Int Ed 45: 4064 - 4092. doi:10.1002/anie.200502655.
15. Klabunde T, Hessler G (2002) Drug design strategies for targeting G-protein-coupled receptors. Chembiochem 3: 928–944. doi:10.1002/1439-7633(20021004)3:10<928::AID-CBIC928>3.0.CO;2-5.
16. Moult J (2005) A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. Curr Opin Struct Biol 15: 285–289. doi:10.1016/j.sbi.2005.05.011.
17. Zhang Y, Skolnick J (2005) The protein structure prediction problem could be solved using the current PDB library. Proc Natl Acad Sci USA 102: 1029–1034. doi:10.1073/pnas.0407152101.
18. Chothia C, Lesk AM (1986) The relation between the divergence of sequence and structure in proteins. EMBO J 5: 823–826.
19. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403–410. doi:10.1006/jmbi.1990.9999.
20. Xu D (2009) Computational methods for protein sequence comparison and search. Curr Protoc Protein Sci Chapter 2: Unit2.1. doi:10.1002/0471140864.ps0201s56.
21. Sander C, Schneider R (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins 9: 56–68. doi:10.1002/prot.340090107.
22. Krieger E, Nabuurs SB, Vriend G (2003) Homology modeling. Methods Biochem Anal 44: 509–523.

23. Martí-Renom MA, Stuart AC, Fiser A, Sánchez R, Melo F, et al. (2000) Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct 29: 291–325. doi:10.1146/annurev.biophys.29.1.291.
24. Leach AR (2001) Empirical force field methods: molecular mechanics. Molecular modelling: principles and applications 2nd Edition. Dorchester: Pearson Education. pp. 165–252.
25. Case D a, Cheatham TE, Darden T, Gohlke H, Luo R, et al. (2005) The Amber biomolecular simulation programs. J Comput Chem 26: 1668–1688. doi:10.1002/jcc.20290.
26. Jr AM, Bashford D, Bellott M (1998) All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. J Phys Chem B 102: 3586–3616. doi:10.1021/jp973084f.
27. Spoel DVD, Lindahl E, Hess B, Groenhof G, Mark AE, et al. (2005) GROMACS: Fast, Flexible, and Free. J Comp Chem 26: 1701–1718. doi:10.1002/jcc.20291.
28. Kaminski GA, Friesner RA, Tirado-rives J, Jorgensen WL (2001) Comparison with Accurate Quantum Chemical Calculations on Peptides. J Phys Chem B 2: 6474–6487.
29. Wang J, Wolf RM, Caldwell JW, Kollman P a, Case D a (2004) Development and testing of a general amber force field. J Comput Chem 25: 1157–1174. doi:10.1002/jcc.20035.
30. Halgren TA (1996) Merck Molecular Force Field: I. Basis, Form, Scope, Parameterization and Performance of MMFF94. J Comp Chem 17: 490–519.
31. Leach AR (2001) Energy minimisation and related methods for exploring the energy surface. Molecular modelling: principles and applications. 2nd ed. Dorchester: Pearson Education. pp. 253–302.
32. Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by simulated annealing. Science 220: 671–680. doi:10.1126/science.220.4598.671.
33. Jones G, Willett P, Glen RC, Leach AR, Taylor R (1997) Development and Validation of a Genetic Algorithm for Flexible Docking. J Mol Biol 267: 727–748. doi:10.1006/jmbi.1996.0897.
34. Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, et al. (1998) Automated Docking Using a Lamarckian Genetic Algorithm and An Empirical Binding Free Energy Function. J Comput Chem 19: 1639–1662. doi:10.1002/(SICI)1096-987X(19981115)19:14<1639::AID-JCC10>3.0.CO;2-B.
35. Venkatachalam CM, Jiang X, Oldfield T, Waldman M (2003) LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites. J Mol Graph Model 21: 289–307.
36. Liu M, Wang S (1999) MCDOCK: a Monte Carlo simulation approach to the molecular docking problem. J Comput Aided Mol Des 13: 435–451.
37. Still W, Tempczyk A, Hawley R (1990) Semianalytical treatment of solvation for molecular mechanics and dynamics. J Am Chem Soc 112: 6127–6129.
38. Leach A (2001) An introduction to computational quantum mechanics. Molecular modelling: principles and applications. 2nd ed. Dorchester. pp. 26–107.
39. Goodford PJ (1985) A Computational Procedure for Determining Energetically Favorable Binding Sites on Biologically Important Macromolecules. J Med Chem 28: 849–857.
40. Pastor M, Cruciani G, McLay I, Pickett S, Clementi S (2000) GRid-INdependent Descriptors (GRIND): A Novel Class of Alignment-Independent Three-Dimensional Molecular Descriptors. J Med Chem 43: 3233–3243. doi:10.1021/jm000941m.
41. Li Q, Jørgensen FS, Oprea T, Brunak S, Taboureau O (2008) hERG classification model based on a combination of support vector machine method and GRIND descriptors. Mol Pharmaceutics 5: 117–127. doi:10.1021/mp700124e.
42. Durán A, Zamora I, Pastor M (2009) Suitability of GRIND-based principal properties for the description of molecular similarity and ligand-based virtual screening. J Chem Inf Model 49: 2129–2138. doi:10.1021/ci900228x.
43. Durán A, Martínez GC, Pastor M (2008) Development and validation of AMANDA, a new algorithm for selecting highly relevant regions in Molecular Interaction Fields. J Chem Inf Model 48: 1813–1823. doi:10.1021/ci800037t.

44. Ortuso F, Langer T, Alcaro S (2006) GBPM: GRID-based pharmacophore model: concept and application studies to protein-protein recognition. Bioinformatics 22: 1449–1455. doi:10.1093/bioinformatics/btl115.
45. Cross S, Baroni M, Carosati E, Benedetti P, Clementi S (2010) FLAP: GRID molecular interaction fields in virtual screening. Validation using the DUD data set. J Chem Inf Model 50: 1442–1450. doi:10.1021/ci100221g.
46. Kubinyi H (1997) QSAR and 3D QSAR in drug design.Part 1: methodology. Drug Discov Today 2: 457–467.
47. Cramer RD, Patterson DE, Bunce JD (1988) Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. J Am Chem Soc 110: 5959–5967. doi:10.1021/ja00226a005.
48. Klebe G, Abraham U, Mietzner T (1994) Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. J Med Chem 37: 4130–4146.
49. Klebe G (1993) Structural alignment of molecules. In: Kubinyi H, editor. 3D-QSAR in Drug Design: Theory, Methods and Applications. Leiden: ESCOM. pp. 173–199.
50. Natesan S, Wang T, Lukacova V, Bartus V, Khandelwal A, et al. (2011) Rigorous Treatment of Multispecies Multimode Ligand-Receptor Interactions in 3D-QSAR : CoMFA Analysis of Thyroxine Analogs Binding to Transthyretin. J Chem Inf Model.
51. Doweyko AM (2008) QSAR: dead or alive? J Comput Aided Mol Des 22: 81–89. doi:10.1007/s10822-007-9162-7.
52. Marshall GR, Barry CD, Bosshard HE, Dammkoehler RA, Dunn DA (1979) The conformational parameter in drug design: The active analog approach. In: Olsson EC, Christoffersen RE, editors. Computer-Assisted Drug Design. Washington, DC: ACS Publications. pp. 205–226. doi:10.1021/bk-1979-0112.ch009.
53. Crippen GM (1979) Distance geometry approach to rationalizing binding data. J Med Chem 22: 988–997.
54. Crippen GM (1980) Quantitative structure-activity relationships by distance geometry: systematic analysis of dihydrofolate reductase inhibitors. J Med Chem 23: 599–606.
55. Crippen GM (1981) Quantitative structure-activity relationships by distance geometry: thyroxine binding site. J Med Chem 24: 198–203.
56. Aqvist J, Medina C, Samuelsson JE (1994) A new method for predicting binding affinity in computer-aided drug design. Protein Eng 7: 385–391.
57. Ortiz AR, Pisabarro MT, Gago F, Wade RC (1995) Prediction of drug binding affinities by comparative binding energy analysis. J Med Chem 38: 2681–2691.
58. Head RD, Smythe ML, Oprea TI, Waller CL, Green SM, et al. (1996) VALIDATE: A New Method for the Receptor-Based Prediction of Binding Affinities of Novel Ligands. J Am Chem Soc 118: 3959–3969. doi:10.1021/ja9539002.
59. Vedani A, Briem H, Dobler M, Dollinger H, McMasters DR (2000) Multiple-conformation and protonation-state representation in 4D-QSAR: the neurokinin-1 receptor system. J Med Chem 43: 4416–4427.
60. Karplus M, Mccammon JA (2002) Molecular dynamics simulations of biomolecules. Nat Struct Biol 9.
61. Hansson T, Oostenbrink C, van Gunsteren WF (2002) Molecular dynamics simulations. Curr Opin Struct Biol 12: 190–196.
62. Shaw DE, Bowers KJ, Chow E, Eastwood MP, Ierardi DJ, et al. (2009) Millisecond-scale molecular dynamics simulations on Anton. Proc. Conf. High Perform. Comput. Network. Storage Anal. New York, New York, USA: ACM Press. p. 1. doi:10.1145/1654059.1654099.
63. Shaw DE, Maragakis P, Lindorff-Larsen K, Piana S, Dror RO, et al. (2010) Atomic-level characterization of the structural dynamics of proteins. Science 330: 341–346. doi:10.1126/science.1187409.
64. Dror RO, Jensen MØ, Borhani DW, Shaw DE (2010) Exploring atomic resolution physiology on a femtosecond to millisecond timescale using molecular dynamics simulations. J Gen Physiol 135: 555–562. doi:10.1085/jgp.200910373.

65. Woo H-J, Roux B (2005) Calculation of absolute protein-ligand binding free energy from computer simulations. Proc Natl Acad Sci USA 102: 6825–6830. doi:10.1073/pnas.0409005102.
66. Wang J, Deng Y, Roux B (2006) Absolute binding free energy calculations using molecular dynamics simulations with restraining potentials. Biophys J 91: 2798–2814. doi:10.1529/biophysj.106.084301.
67. Malmstrom RD, Watowich SJ (2011) Using free energy of binding calculations to improve the accuracy of virtual screening predictions. J Chem Inf Model 51: 1648–1655. doi:10.1021/ci200126v.
68. Carlson H a, McCammon J a (2000) Accommodating protein flexibility in computational drug design. Mol Pharmacol 57: 213–218.
69. Totrov M, Abagyan R (2008) Flexible ligand docking to multiple receptor conformations: a pratical alternative. Curr Opin Struct Biol 18: 178–184.
70. Carlson HA, Masukawa KM, McCammon JA (1999) Method for Including the Dynamic Fluctuations of a Protein in Computer-aided Drug Design. J Phys Chem A 103: 10213–10219.
71. Masukawa KM, Carlson HA, McCammon JA (2000) Technique for developing a pharmacophore model that accommodates inherent protein flexibility: an application to HIV-1 integrase. In: Guner OF, editor. Pharmacophore Perception, Development, and Use in Drug Design. La Jolla, CA: International University Line.
72. Carlson HA, Masukawa KM, Rubins K, Bushman FD, Jorgensen WL, et al. (2000) Developing a dynamic pharmacophore model for HIV-1 integrase. J Med Chem 43: 2100–2114.
73. Kuhn B, Jacobsen W, Christians U, Benet LZ, Kollman P a (2001) Metabolism of sirolimus and its derivative everolimus by cytochrome P450 3A4: insights from docking, molecular dynamics, and quantum chemical calculations. J Med Chem 44: 2027–2034.
74. Correia BE, Loureiro-Ferreira N, Rodrigues JR, Brito RMM (2006) A structural model of an amyloid protofilament of transthyretin. Protein Sci 15: 28–32.
75. Tembe BL, McCammon JA (1984) Ligand-receptor interactions. Comput Chem 8: 281.
76. Oostenbrink C, van Gunsteren WF (2005) Free energies of ligand binding for structurally diverse compounds. Proc Natl Acad Sci USA 102: 6750–6754. doi:10.1073/pnas.0407404102.
77. Jorgensen WL, Thomas LL (2009) Perspective on Free-Energy Perturbation Calculations for Chemical Equilibria. J Chem Theory Comput 4: 869–876. doi:10.1021/ct800011m.Perspective.
78. Knight JL, Brooks CL (2009) Lambda-dynamics free energy simulation methods. J Comput Chem 30: 1692–1700. doi:10.1002/jcc.21295.
79. Michel J, Essex JW (2010) Prediction of protein-ligand binding affinity by free energy simulations: assumptions, pitfalls and expectations. J Comput Aided Mol Des 24: 639–658. doi:10.1007/s10822-010-9363-3.
80. Gilson MK, Zhou H-xiang (2007) Calculation of Protein-Ligand Binding Affnities. Annu Rev Biophys Biomol Struct 36: 21–42. doi:10.1146/annurev.biophys.36.040306.132550.
81. Shirts MR, Mobley DL, Chodera JD (2007) Alchemical free energy calculations: ready for prime time? Annu Rep Comput Chem 3: 41–59.
82. Deng Y, Roux B (2009) Computations of standard binding free energies with molecular dynamics simulations. J Phys Chem B 113: 2234–2246. doi:10.1021/jp807701h.
83. Mobley DL, Dill K a (2009) Binding of small-molecule ligands to proteins: "what you see" is not always "what you get." Structure 17: 489–498. doi:10.1016/j.str.2009.02.010.
84. Chodera JD, Mobley DL, Shirts MR, Dixon RW, Branson K, et al. (2011) Alchemical free energy methods for drug discovery: progress and challenges. Curr Opin Struct Biol 21: 150–160. doi:10.1016/j.sbi.2011.01.011.
85. Blaney JM, Dixon JS (1993) A good ligand is hard to find: automated docking methods. Perspect Drug Discov Des 1: 301–319.

86. Kuntz ID (1982) Structure-based strategies for drug design and discovery. Science 257: 1078–1082.
87. Leach AR (2001) Monte Carlo simulation methods. Molecular modelling: principles and applications. 2nd ed. Dorchester: Pearson Education. pp. 410–456.
88. Goldberg DE (1989) Genetic Algorithms in Search, Optimization, and Machine Learning. London: Addison-Wesley Professional. p.
89. Bissantz C, Folkers G, Rognan D (2000) Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. J Med Chem 43: 4759–4767.
90. Meng EC, Shoichet BK, Kuntz ID (1992) Automated Docking with Grid-Based Energy Evaluation. J Comput Chem 13: 505–524.
91. Eldridge MD, Murray CW, Auton TR, Paolini GV, Mee RP (1997) Empirical Scoring Functions: I. The Development of A Fast Empirical Scoring Function to Estimate the Binding Affinity of Ligands in Receptor Complexes. J Comput Aided Mol Des 11: 425–445.
92. Murray CW, Auton TR, Eldridge MD (1998) Empirical scoring functions. II. The testing of an empirical scoring function for the prediction of ligand-receptor binding affinities and the use of Bayesian regression to improve the quality of the model. J Comput Aided Mol Des 12: 503–519.
93. Gohlke H, Hendlich M, Klebe G (2000) Knowledge-Based Scoring Function to Predict Protein-Ligand Interactions. J Mol Biol 295: 337–356. doi:10.1006/jmbi.1999.3371.
94. Muegge I, Martin YC (1999) A general and fast scoring function for protein–ligand interactions: a simplified potential approach. J Med Chem 42: 791–804.
95. Wermuth G, Ganellin CR, Lindberg P, Mitscher LA (1998) Glossay of terms used in medicinal chemistry (IUPAC Recommendations 1998). Pure Appl Chem 70: 1129–1143.
96. Khedkar SA, Malde AK, Coutinho EC, Srivastava S (2007) Pharmacophore modeling in drug discovery and development: an overview. Med Chem 3: 187–197.
97. Drie J (1997) Strategies for the determination of pharmacophoric 3D database queries. J Comput Aided Mol Des 11: 39–52.
98. Boström J (2001) Reproducing the conformations of protein-bound ligands: a critical evaluation of several popular conformational searching tools. J Comput Aided Mol Des 15: 1137–1152.
99. Güner OF (2002) History and evolution of the pharmacophore concept in computer-aided drug design. Curr Top Med Chem 2: 1321–1332.
100. Steindl T, Langer T (2004) Influenza virus neuraminidase inhibitors: generation and comparison of structure-based and common feature pharmacophore hypotheses and their application in virtual screening. J Chem Inf Comput Sci 44: 1849–1856. doi:10.1021/ci049844i.
101. Pirard B, Brendel J, Peukert S (2005) The discovery of Kv1.5 blockers as a case study for the application of virtual screening approaches. J Chem Inf Model 45: 477–485. doi:10.1021/ci0400011.
102. Rella M, Rushworth C a, Guy JL, Turner AJ, Langer T, et al. (2006) Structure-based pharmacophore design and virtual screening for novel angiotensin converting enzyme 2 inhibitors. J Chem Inf Model 46: 708–716. doi:10.1021/ci0503614.
103. Wolber G, Langer T (2005) LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. J Chem Inf Model 45: 160–169. doi:10.1021/ci049885e.
104. Johnson M, Maggiora GM (1990) Concepts and Applications of Molecular Similarity. Johnson M, Maggiora GM, editors New York: John Wiley & Sons. p.
105. Maldonado AG, Doucet JP, Petitjean M, Fan B-T (2006) Molecular similarity and diversity in chemoinformatics: from theory to applications. Mol Diver 10: 39–79. doi:10.1007/s11030-006-8697-1.
106. Wang N, DeLisle RK, Diller DJ (2005) Fast small molecule similarity searching with multiple alignment profiles of molecules represented in one-dimension. J Med Chem 48: 6980–6990. doi:10.1021/jm050563r.

107. Bender A, Glen RC (n.d.) A discussion of measures of enrichment in virtual screening: comparing the information content of descriptors with increasing levels of sophistication. J Chem Inf Model 45: 1369–1375. doi:10.1021/ci0500177.
108. James CA, Weininger D (2006) Daylight theory manual. Daylight Chemical Information Systems.
109. Barnard JM, Downs GM (1997) Chemical Fragment Generation and Clustering Software. J Chem Inf Comput Sci 37: 141–142. doi:10.1021/ci960090k.
110. Bender A, Mussa HY, Glen RC, Reiling S (2004) Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): evaluation of performance. J Chem Inf Comput Sci 44: 1708–1718. doi:10.1021/ci0498719.
111. Willet P (2005) Searching techniques for databases of two- and three-dimensional chemical structures. J Med Chem 48: 4183–4199.
112. Hert J, Willett P, Wilton DJ, Acklin P, Azzaoui K, et al. (2005) Enhancing the effectiveness of similarity-based virtual screening using nearest-neighbor information. J Med Chem 48: 7049–7054. doi:10.1021/jm050316n.
113. Xue L, Godden JW, Stahura FL, Bajorath J (2003) Profile scaling increases the similarity search performance of molecular fingerprints containing numerical descriptors and structural keys. J Chem Inf Comput Sci 43: 1218–1225. doi:10.1021/ci030287u.
114. Schuffenhauer A, Floersheim P, Acklin P, Jacoby E (2003) Similarity metrics for ligands reflecting the similarity of the target proteins. J Chem Inf Comput Sci 43: 391–405. doi:10.1021/ci025569t.
115. Hert J, Willett P, Wilton DJ, Acklin P, Azzaoui K, et al. (2004) Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. J Chem Inf Comput Sci 44: 1177–1185. doi:10.1021/ci034231b.
116. Tovar A, Eckert H, Bajorath J (2007) Comparison of 2D fingerprint methods for multiple-template similarity searching on compound activity classes of increasing structural diversity. ChemMedChem 2: 208–217. doi:10.1002/cmdc.200600225.
117. Willett P (2006) Similarity-Based Virtual Screening Using 2D Fingerprints. Drug Discov Today 11: 1046–1053. doi:10.1016/j.drudis.2006.10.005.
118. Rush TS, Grant JA, Mosyak L, Nicholls A (2005) A Shape-Based 3-D Scaffold Hopping Method and Its Application to a Bacterial Protein-Protein Interaction. J Med Chem 48: 1489–1495. doi:10.1021/jm040163o.
119. Haigh J a, Pickup BT, Grant JA, Nicholls A (2005) Small molecule shape-fingerprints. J Chem Inf Model 45: 673–684. doi:10.1021/ci049651v.
120. Jenkins JL, Glick M, Davies JW (2004) A 3D similarity method for scaffold hopping from known drugs or natural ligands to new chemotypes. J Med Chem 47: 6144–6159. doi:10.1021/jm049654z.
121. Cheeseright T, Mackey M, Rose S, Vinter A (2006) Molecular field extrema as descriptors of biological activity: definition and validation. J Chem Inf Model 46: 665–676. doi:10.1021/ci050357s.
122. Saeh JC, Lyne PD, Takasaki BK, Cosgrove D a (2005) Lead hopping using SVM and 3D pharmacophore fingerprints. J Chem Inf Model 45: 1122–1133. doi:10.1021/ci049732r.
123. Dunn D, Orlowski M, McCoy P, Gastgeb F, Appell K, et al. (2000) Ultra-high throughput screen of two-million-member combinatorial compound collection in a miniaturized, 1536-well assay format. J Biomol Screen 5: 177–188.
124. Moore K, Turconi S, Miles-Williams A, Djaballah H, Hurskainen P, et al. (1999) A Homogenous 384-Well High Throughput Screen for Novel Tumor Necrosis Factor Receptor: Ligand Interactions Using Time Resolved Energy Transfer. J Biomol Screen 4: 205–214.
125. Hamasaki K, Rando RR (1998) A high-throughput fluorescence screen to monitor the specific binding of antagonists to RNA targets. Anal Biochem 261: 183–190. doi:10.1006/abio.1998.2740.
126. Young K, Lin S, Sun L, Lee E, Modi M, et al. (1998) Identification of a calcium channel modulator using a high throughput yeast two-hybrid screen. Nat Biotechnol 16: 946–950. doi:10.1038/nbt1098-946.

127. Chen J, Zhang Z, Stebbins JL, Zhang X, Hoffman R, et al. (2007) A fragment-based approach for the discovery of isoform-specific p38alpha inhibitors. ACS Chem Biol 2: 329–336. doi:10.1021/cb700025j.
128. Carr RAE, Congreve M, Murray CW, Rees DC (2005) Fragment-based lead discovery: leads by design. Drug Discov Today 10: 10–15.
129. Hann MM, Leach AR, Harper G (n.d.) Molecular complexity and its impact on the probability of finding leads for drug discovery. J Chem Inf Comput Sci 41: 856–864.
130. Geldenhuys WJ, Gaasch KE, Watson M, Allen DD (2006) Optimizing the use of open-source software applications in drug discovery. Drug Discov Today 11: 127–132.
131. Gaulton A, Overington JP (2010) Role of Open Chemical Data in Aiding Drug Discovery and Design. Future Med Chem 2: 903–907.
132. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, et al. (2011) ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res 44: 1–8. doi:10.1093/nar/gkr777.
133. Liu T, Lin Y, Wen X, Jorissen RN, Gilson MK (2007) BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. Nucleic Acids Res 35: D198–201. doi:10.1093/nar/gkl999.
134. Hu L, Benson ML, Smith RD, Lerner MG, Carlson HA (2005) Binding MOAD (Mother Of All Databases). Proteins 60: 333–340. doi:10.1002/prot.20512.
135. PDSP Database (n.d.). Available:http://pdsp.med.unc.edu/pdsp.php.
136. Wang Y, Bolton E, Dracheva S, Karapetyan K, Shoemaker B a, et al. (2010) An overview of the PubChem BioAssay resource. Nucleic Acids Res 38: D255–66. doi:10.1093/nar/gkp965.
137. Irwin JJ, Shoichet BK (2005) ZINC - A Free Database of Commercially Available Compounds for Virtual Screening. J Chem Inf Model 45: 177–182. doi:10.1021/ci049714+.
138. Blum LC, Reymond J-louis (2009) 970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13. J Am Chem Soc 131: 8732–8733.
139. Villoutreix BO, Renault N, Lagorce D, Sperandio O, Montes M, et al. (2007) Free resources to assist structure-based virtual ligand screening experiments. Curr Protein Pept Sci 8: 381–411.
140. Knox C, Law V, Jewison T, Liu P, Ly S, et al. (2010) DrugBank 3.0: a comprehensive resource for “Omics” research on drugs. Nucleic Acids Res 39: D1035–D1041. doi:10.1093/nar/gkq1126.
141. DailyMed (n.d.). Available:http://dailymed.nlm.nih.gov/dailymed/about.cfm. Accessed 7 February 2011.
142. Richard AM, Williams CR (2002) Distributed structure-searchable toxicity (DSSTox) public database network: a proposal. Mutation Res 499: 27–52.
143. Hochstein C, Arnesen S, Goshorn J (2007) Environmental Health and Toxicology Resources of the United States National Library of Medicine. Med Ref Serv Q 26: 21–45. doi:10.1300/J115v26n03.
144. Collins FS, Gray GM, Bucher JR (2008) Toxicology. Transforming environmental health protection. Science 319: 906–907. doi:10.1126/science.1154619.
145. Judson RS, Houck K a, Kavlock RJ, Knudsen TB, Martin MT, et al. (2010) In vitro screening of environmental chemicals for targeted testing prioritization: the ToxCast project. Environ Health Perspect 118: 485–492. doi:10.1289/ehp.0901392.
146. Haynes R (2010) Toxcast on Target: In Vitro Assays and Computer Modeling Show Promise for Screening Chemicals. Environ Health Perspect.
147. Tox21: Putting a lens on the vision of toxicity testing in the 21st century (n.d.). Available:www.alttox.org/ttrc/overarching-challenges/way-forward/austin-kavlock-tice. Accessed 6 February 2011.
148. Bailey DS, Zanders ED (2008) Drug Discovery in the Era of Facebook - New Tools for Scientific Networking. Drug Discov Today 13: 863–868. doi:10.1016/j.drudis.2008.07.003.

149. Hohman M, Gregory K, Chibale K, Smith PJ, Ekins S, et al. (2009) Novel web-based tools combining chemistry informatics , biology and social networks for drug discovery. Drug Discov Today 14: 261–270. doi:10.1016/j.drudis.2008.11.015.
150. Muchmore SW, Edmunds JJ, Stewart KD, Hajduk PJ (2010) Cheminformatic tools for medicinal chemists. J Med Chem 53: 4830–4841. doi:10.1021/jm100164z.
151. Doman TN, McGovern SL, Witherbee BJ, Kasten TP, Kurumbail R, et al. (2002) Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. J Med Chem 45: 2213–2221.
152. Virchow R (1854) Zur Cellulosefrage. Virchows Arch Pathol Anat Physiol 6: 416–426.
153. Westermark P, Araki S, Benson MD, Cohen AS, Frangione B, et al. (1999) Nomenclature of amyloid fibril proteins. Report from the meeting of the International Nomenclature Committee on Amyloidosis, August 8-9, 1998. Part 1. Amyloid 6: 63–66.
154. Katzman R (1986) Alzheimer's disease. N Engl J Med 314: 964–973. doi:10.1056/NEJM198604103141506.
155. Ghiso J, Wisniewski T, Frangione B (1994) Unifying features of systemic and cerebral amyloidosis. Mol Neurobiol 8: 49–64. doi:10.1007/BF02778007.
156. Glenner GG, Wong CW (1984) Alzheimer's disease: initial report of the purification and characterization of a novel cerebrovascular amyloid protein. Biochem Biophys Res Commun 120: 885–890.
157. Masters CL, Simms G, Weinman NA, Multhaup G, McDonald BL, et al. (1985) Amyloid plaque core protein in Alzheimer disease and Down syndrome. Proc Natl Acad Sci USA 82: 4245–4249.
158. Esch FS, Keim PS, Beattie EC, Blacher RW, Culwell AR, et al. (1990) Cleavage of amyloid beta peptide during constitutive processing of its precursor. Science 248: 1122–1124.
159. Haass C, Schlossmacher MG, Hung AY, Vigo-Pelfrey C, Mellon A, et al. (1992) Amyloid beta-peptide is produced by cultured cells during normal metabolism. Nature 359: 322–325. doi:10.1038/359322a0.
160. Hill CS, Ibanez ML, Samaan NA, Ahearn MJ, Clark RL (1973) Medullary (solid) carcinoma of the thyroid gland: an analysis of the M. D. Anderson hospital experience with patients with the tumor, its special features, and its histogenesis. Medicine 52: 141–171.
161. Sletten K, Westermark P, Natvig JB (1976) Characterization of amyloid fibril proteins from medullary carcinoma of the thyroid. J Exp Med 143: 993–998.
162. Westermark P, Wernstedt C, O'Brien TD, Hayden DW, Johnson KH (1987) Islet amyloid in type 2 human diabetes mellitus and adult diabetic cats contains a novel putative polypeptide hormone. Am J Pathol 127: 414–417.
163. Johnson KH, O'Brien TD, Betsholtz C, Westermark P (1989) Islet amyloid, islet-amyloid polypeptide, and diabetes mellitus. N Engl J Med 321: 513–518. doi:10.1056/NEJM198908243210806.
164. Andrade C (1952) A peculiar form of peripheral neuropathy. Familial atypical generalized amyloidosis with special involvement of peripheral nerves. Brain 75: 408–427.
165. Reilly MM, King RHM (1993) Familial amyloid polyneuropathy. Brain Pathol 3: 165–176.
166. Staunter H (1991) Familial amyloid polyneuropathies. In: Vinken PJ, Bruyn GW, Klawans HL, editors. Handbook of Clinical Pathology. Amsterdam: Elsevier. pp. 89–115.
167. Fonseca C, Ceia F, Nogueira JS, Alves M, Carvalho M, et al. (1991) Myocardiopathy caused by Portuguese-type familial amyloidotic polyneuropathy. Sequential morphologic and functional study of 60 patients. Rev Port Cardiol 10: 909–916.
168. Díaz Lobato S, Guerrero E, González P, Crespo M, Esteban R, et al. (1991) Pulmonary involvement in familial amyloid polyneuropathy type I. Rev Clin Esp 189: 335–337.

169. Allard SA, King RH, Thomas PK, Bourke BE (1991) Haemarthrosis due to fracture through amyloid deposits in bone in Portuguese familial amyloidosis. Ann Rheum Dis 50: 820–822.
170. Walsh DM, Hartley DM, Kusumoto Y, Fezoui Y, Condron MM, et al. (1999) Amyloid beta-protein fibrillogenesis. Structure and biological activity of protofibrillar intermediates. J Biol Chem 274: 25945–25952.
171. Teplow DB (1998) Structural and kinetic features of amyloid beta-protein fibrillogenesis. Amyloid 5: 121–142.
172. Selkoe DJ (1999) Translating cell biology into therapeutic advances in Alzheimer's disease. Nature 399: A23–31.
173. Saraiva MJ, Costa PP (1991) Molecular biology of the amyloidogenesis in the transthyretin related amyloidoses. In: Natvig JB, Forre O, Husby G, editors. Amyloid and Amyloidosis. Doordrecht, The Netherlands: Kluwer Academic Publishers. pp. 569–574.
174. Gustavsson A, Engström U, Westermark P (1991) Normal transthyretin and synthetic transthyretin fragments form amyloid-like fibrils in vitro. Biochem Biophys Res Commun 175: 1159–1164.
175. Jarvis JA, Kirkpatrick A, Craik DJ (1994) 1H NMR analysis of fibril-forming peptide fragments of transthyretin. Int J Pept Protein Res 44: 388–398.
176. Paci E, Gsponer J, Salvatella X, Vendruscolo M (2004) Molecular dynamics studies of the process of amyloid aggregation of peptide fragments of transthyretin. J Mol Biol 340: 555–569.
177. Tawara S, Nakazato M, Kangawa K, Matsuo H, Araki S (1983) Identification of amyloid prealbumin variant in familial amyloidotic polyneuropathy (Japanese type). Biochem Biophys Res Commun 116: 880–888.
178. Saraiva MJ, Birken S, Costa PP, Goodman DS (1984) Amyloid fibril protein in familial amyloidotic polyneuropathy, Portuguese type. Definition of molecular abnormality in transthyretin (prealbumin). J Clin Invest 74: 104–119.
179. Harper JD, Lansbury PT (1997) Models of amyloid seeding in Alzheimer's disease and scrapie: mechanistic truths and physiological consequences of the time-dependent solubility of amyloid proteins. Annu Rev Biochem 66: 385–407.
180. Kisilevsky R (2000) Review: amyloidogenesis-unquestioned answers and unanswered questions. J Struct Biol 130: 99–108. doi:10.1006/jsbi.2000.4222.
181. Hurshman AR, White JT, Powers ET, Kelly JW (2004) Transthyretin Aggregation Under Partially Denaturing Conditions Is A Downhill Polymerization. Biochemistry 43: 7365–7381. doi:10.1021/bi049621l.
182. Blake C, Serpell L (1996) Synchrotron X-ray studies suggest that the core of the transthyretin amyloid fibril is a continuous beta-sheet helix. Structure 4: 989–998.
183. Sunde M, Serpell LC, Bartlam M, Fraser PE, Pepys MB, et al. (1997) Common core structure of amyloid fibrils by synchrotron X-ray diffraction. J Mol Biol 273: 729–739.
184. Serpell LC, Fraser PE, Sunde M (1999) X-ray fiber diffraction of amyloid fibrils. Methods Enzymol 309: 526–536.
185. Tycko R (2000) Solid-state NMR as a probe of amyloid fibril structure. Curr Opin Chem Biol 4: 500–506.
186. Iwata K, Fujiwara T, Matsuki Y, Akutsu H, Takahashi S, et al. (2006) 3D structure of amyloid protofilaments of beta2-microglobulin fragment probed by solid-state NMR. Proc Natl Acad Sci USA 103: 18119–18124. doi:10.1073/pnas.0607180103.
187. Tycko R (2010) Solid-State NMR Studies of Amyloid Fibril Structure. Annu Rev Phys Chem Epub. doi:10.1146/annurev-physchem-032210-103539.
188. Serpell LC, Sunde M, Fraser PE, Luther PK, Morris EP, et al. (1995) Examination of the structure of the transthyretin amyloid fibril by image reconstruction from electron micrographs. J Mol Biol 254: 113–118. doi:10.1006/jmbi.1995.0604.
189. Seshadri S, Khurana R, Fink AL (1999) Fourier transform infrared spectroscopy in analysis of protein deposits. Methods Enzymol 309: 559–576.

190. Ding TT, Harper JD (1999) Analysis of amyloid-beta assemblies using tapping mode atomic force microscopy under ambient conditions. Methods Enzymol 309: 510–525.
191. Xue W-F, Homans SW, Radford SE (2009) Amyloid fibril length distribution quantified by atomic force microscopy single-particle image analysis. Prot Eng Des Sel 22: 489–496. doi:10.1093/protein/gzp026.
192. Hardy J, Selkoe DJ (2002) The amyloid hypothesis of Alzheimer's disease: progress and problems on the road to therapeutics. Science 297: 353–356.
193. Bucciantini M, Giannoni E, Chiti F, Baroni F, Formigli L, et al. (2002) Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases. Nature 416: 507–511. doi:10.1038/416507a.
194. Turnell WG, Finch JT (1992) Binding of the dye congo red to the amyloid protein pig insulin reveals a novel homology amongst amyloid-forming peptide sequences. J Mol Biol 227: 1205–1223.
195. Król M, Borowski T, Roterman I, Piekarska B, Stopa B, et al. (2004) Force-field parametrization and molecular dynamics simulations of Congo red. J Comput Aided Mol Des: 41–53.
196. Biancalana M, Makabe K, Koide A, Koide S (2009) Molecular Mechanism of Thioflavin-T Binding to the Surface of β-Rich Peptide Self-Assemblies. J Mol Biol 385: 1052–1063. doi:10.1016/j.jmb.2008.11.006.
197. Sunde M, Serpell LC, Bartlam M, Fraser PE, Pepys MB, et al. (1997) Common core structure of amyloid fibrils by synchrotron X-ray diffraction. J Mol Biol 273: 729–739. doi:10.1006/jmbi.1997.1348.
198. Colletier J-P, Laganowsky A, Landau M, Zhao M, Soriaga AB, et al. (2011) Molecular basis for amyloid-beta polymorphism. Proc Natl Acad Sci USA 108: 16938–16943. doi:10.1073/pnas.1112600108.
199. Kabat BEA, Moore DANH, Landow H (1942) An electrophoretic study of the protein components in cerebrospinal fluid and their relationship to the serum proteins. J Clin Invest 2: 571–577.
200. Seibert F, Nelson J (1942) Electrophoretic study of the blood protein response in tuberculosis. J Biol Chem 143: 29–38.
201. Ingbar SH (1958) Pre-albumin: a thyroxinebinding protein of human plasma. Endocrinology 63: 256–259.
202. Kanai M, Raz A, Goodman DS (1968) Retinol-binding protein: the transport protein for vitamin A in human plasma. J Clin Invest 47: 2025–2044.
203. Goodman DS, Peters T, Robbins J, Schwick HG (1981) Prealbumin becomes transthyretin? In Nomenclature committe of IUB (NC-IUB) IUB-IUPAC Joint Comission on Biochemical Nomenclature (JCBN). J Biol Chem 256: 12–14.
204. Quintas A, Vaz DC, Cardoso I, Saraiva MJ, Brito RM (2001) Tetramer dissociation and monomer partial unfolding precedes protofibril formation in amyloidogenic transthyretin variants. J Biol Chem 276: 27207–27213. doi:10.1074/jbc.M101024200.
205. Brito RMM, Damas AM, Saraiva MJ (2003) Amyloid Formation by Transthyretin: From Protein Stability to Protein Aggregation. Curr Med Chem Immun Endoc Metab Agents 3: 349–360.
206. Miroy GJ, Lai Z, Lashuel HA, Peterson SA, Strang C, et al. (1996) Inhibiting Transthyretin Amyloid Fibril Formation Via Protein Stabilization. Proc Natl Acad Sci USA 93: 15051–15056.
207. Hammarström P, Jiang X, Hurshman AR, Powers ET, Kelly JW (2002) Sequence-Dependent Denaturation Energetics: A Major Determinant in Amyloid Disease Diversity. Proc Natl Acad Sci USA 99 Suppl 4: 16427–16432. doi:10.1073/pnas.202495199.
208. Wiseman RL, Johnson SM, Kelker MS, Foss T, Wilson IA, et al. (2005) Kinetic Stabilization of an Oligomeric Protein by a Single Ligand Binding Event. J Am Chem Soc 127: 5540–5551.
209. Rodrigues JR, Simões CJV, Silva CG, Brito RMM (2010) Potentially amyloidogenic conformational intermediates populate the unfolding landscape of transthyretin:

insights from molecular dynamics simulations. Protein Sci 19: 202–219. doi:10.1002/pro.289.
210. Buxbaum JN, Reixach N (2009) Transthyretin: the Servant of Many Masters. Cell Mol Life Sci 66: 3095–3101. doi:10.1007/s00018-009-0109-0.
211. Munro SL, Lim CF, Hall JG, Barlow JW, Craik DJ, et al. (1989) Drug Competition for Thyroxine Binding to Transthyretin (Prealbumin): Comparison with Effects on Thyroxine-Binding Globulin. J Clin Endocrinol Metab 68: 1141–1147.
212. Baures PW, Peterson SA, Kelly JW (1998) Discovering Transthyretin Amyloid Fibril Inhibitors by Limited Screening. Bioorg Med Chem 6: 1389–1401.
213. Baures PW, Oza VB, Peterson SA, Kelly JW (1999) Synthesis and Evaluation of Inhibitors of Transthyretin Amyloid Formation Based on the Non-Steroidal Anti-Inflammatory Drug, Flufenamic Acid. Bioorg Med Chem 7: 1339–1347.
214. Oza VB, Petrassi HM, Purkey HE, Kelly JW (1999) Synthesis and Evaluation of Anthranilic Acid-Based Transthyretin Amyloid Fibril Inhibitors. Bioorg Med Chem 9: 1–6.
215. Klabunde T, Petrassi HM, Oza VB, Raman P, Kelly JW, et al. (2000) Rational design of potent human transthyretin amyloid disease inhibitors. Nat Struct Biol 7: 312–321.
216. Oza VB, Smith C, Raman P, Koepf EK, Lashuel HA, et al. (2002) Synthesis, Structure, and Activity of Diclofenac Analogues as Transthyretin Amyloid Fibril Formation Inhibitors. J Med Chem 45: 321–332.
217. Hammarström P, Wiseman RL, Powers ET, Kelly JW (2003) Prevention of Transthyretin Amyloid Disease by Changing Protein Misfolding Energetics. Science 299: 713–716. doi:10.1126/science.1079589.
218. Powers ET, Wiseman RL, Purkey HE (2003) Benzoxazoles as Transthyretin Amyloid Fibril Inhibitors: Synthesis, Evaluation, and Mechanism of Action. Angew Chem Int Ed 42: 2758–2761. doi:10.1002/anie.200351179.
219. Purkey HE, Palaninathan SK, Kent KC, Smith C, Safe SH, et al. (2004) Hydroxylated Polychlorinated Biphenyls Selectively Bind Transthyretin in Blood and Inhibit Amyloidogenesis: Rationalizing Rodent PCB Toxicity. Chem Biol 11: 1719–1728. doi:10.1016/j.chembiol.2004.10.009.
220. Adamski-Werner SL, Palaninathan SK, Sacchettini JC, Kelly JW (2004) Diflunisal Analogues Stabilize the Native State of Transthyretin. Potent Inhibition of Amyloidogenesis. J Med Chem 47: 355–374. doi:10.1021/jm030347n.
221. Johnson SM, Wiseman RL, Sekijima Y, Green NS, Adamski-Werner SL, et al. (2005) Native state kinetic stabilization as a strategy to ameliorate protein misfolding diseases: a focus on the transthyretin amyloidoses. Acc Chem Res 38: 911–921. doi:10.1021/ar020073i.
222. Petrassi HM, Johnson SM, Purkey H, Chiang KP, Walkup T, et al. (2005) Potent and Selective Structure-Based Dibenzofuran Inhibitors of Transthyretin Amyloidogenesis: Kinetic Stabilization of the Native State. J Am Chem Soc 127: 6662–6671.
223. Sparkes RS, Sasaki H, Mohandas T, Yoshioka K, Klisak I, et al. (1987) Assignment of the prealbumin (PALB) gene (familial amyloidotic polyneuropathy) to human chromosome region 18q11.2-q12.1. Hum Genet 75: 151–154.
224. Sasaki H, Yoshioka N, Takagi Y, Sakaki Y (1985) Structure of the chromosomal gene for human serum prealbumin. Gene 37: 191–197.
225. Tsuzuki T, Mita S, Maeda S, Araki S, Shimada K (1985) Structure of the human prealbumin gene. J Biol Chem 260: 12224–12227.
226. Fung WP, Thomas T, Dickson PW, Aldred a R, Milland J, et al. (1988) Structure and Expression of the Rat Transthyretin (Prealbumin) Gene. J Biol Chem 263: 480–488.
227. Costa RH, Lai E, Darnell Jr JE (1986) Transcriptional control of the mouse prealbumin (transthyretin) gene: both promoter sequences and a distinct enhancer are cell specific. Mol Cell Biol 6: 4697. doi:10.1128/MCB.6.12.4697.Updated.
228. Costa R, Grayson D, Darnell Jr J (1989) Multiple hepatocyte-enriched nuclear factors function in the regulation of transthyretin and alpha 1-antitrypsin genes. Mol Cell Biol 9: 1415. doi:10.1128/MCB.9.4.1415.Updated.

229. Schreiber G, Richardson SJ (1997) The evolution of gene expression, structure and function of transthyretin. Comp Biochem Physiol B Biochem Mol Biol 116: 137–160.
230. Aldred a R, Prapunpoj P, Schreiber G (1997) Evolution of shorter and more hydrophilic transthyretin N-termini by stepwise conversion of exon 2 into intron 1 sequences (shifting the 3' splice site of intron 1). Eur J Biochem 246: 401–409.
231. Richardson SJ (2005) Expression of transthyretin in the choroid plexus: relationship to brain homeostatsis of thyroid hormones. In: Zheng W, Chodobski A, editors. The Blood–Cerebrospinal Fluid Barrier. Boca Raton: CRC Press. pp. 275–304.
232. Chang L, Munro SL, Richardson SJ, Schreiber G (1999) Evolution of thyroid hormone binding by transthyretins in birds and mammals. Eur J Biochem 259: 534–542.
233. Yamauchi K, Nakajima J, Hayashi H, Hara A (1999) Purification and characterization of thyroid-hormone-binding protein from masu salmon serum. A homolog of higher-vertebrate transthyretin. Eur J Biochem 265: 944–949.
234. Prapunpoj P, Leelawatwatana L, Schreiber G, Richardson SJ (2006) Change in structure of the N-terminal region of transthyretin produces change in affinity of transthyretin to T4 and T3. FEBS J 273: 4013–4023. doi:10.1111/j.1742-4658.2006.05404.x.
235. Power DM, Elias NP, Richardson SJ, Mendes J, Soares CM, et al. (2000) Evolution of the Thyroid Hormone-Binding Protein, Transthyretin. Gen Comp Endocrinol 119: 241–255. doi:10.1006/gcen.2000.7520.
236. Jeanmougin F, Thompson JD, Gouy M, Higgins DG, Gibson TJ (1998) Multiple sequence alignment with Clustal X. Trends Biochem Sci 23: 403–405.
237. Duan W, Richardson SJ, Köhrle J, Chang L, Southwell BR, et al. (1995) Binding of thyroxine to pig transthyretin, its cDNA structure, and other properties. Eur J Biochem 230: 977–986.
238. Dickson PW, Howlett GJ, Schreiber G (1985) Rat transthyretin (prealbumin). Molecular cloning, nucleotide sequence, and gene expression in liver and brain. J Biol Chem 260: 8214–8219.
239. Brack CM, Duan W, Hulbert AJ, Schreiber G (1995) Wallaby transthyretin. Comp Biochem Physiol B Biochem Mol Biol 110: 523–529.
240. Duan W, Achen MG, Richardson SJ, Lawrence MC, Wettenhall RE, et al. (1991) Isolation, characterization, cDNA cloning and gene expression of an avian transthyretin. Implications for the evolution of structure and function of transthyretin in vertebrates. Eur J Biochem 200: 679–687.
241. Achen MG, Duan W, Pettersson TM, Harms PJ, Richardson SJ, et al. (1993) Transthyretin gene expression in choroid plexus first evolved in reptiles. Am J Physiol 265: R982–9.
242. Yamauchi K, Takeuchi H, Overall M, Dziadek M, Munro SL, et al. (1998) Structural characteristics of bullfrog (Rana catesbeiana) transthyretin and its cDNA - comparison of its pattern of expression during metamorphosis with that of lipocalin. Eur J Biochem 256: 287–296.
243. Santos CR, Power DM (1999) Identification of Transthyretin in Fish (Spaurus aurata) - cDNA Cloning and Characterisation. Endocrinology 140: 2430–2433.
244. Dickson P, Aldred A, Menting J, Marley P, Sawyer W, et al. (1987) Thyroxine transport in choroid plexus. J Biol Chem 262: 13907.
245. Cavallaro T, Martone RL, Dwork AJ, Schon EA, Herbert J (1990) The retinal pigment epithelium is the unique site of transthyretin synthesis in the rat eye. Invest Ophthalmol Vis Sci 31: 497–501.
246. Dwork AJ, Cavallaro T, Martone RL, Goodman DS, Schon EA, et al. (1990) Distribution of transthyretin in the rat eye. Invest Ophthalmol Vis Sci 31: 489–496.
247. Martone RL, Herbert J, Dwork A, Schon EA (1988) Transthyretin is synthesized in the mammalian eye. Biochem Biophys Res Commun 151: 905–912.
248. Jacobsson B, Carlström A, Platz A, Collins VP (1990) Transthyretin messenger ribonucleic acid expression in the pancreas and in endocrine tumors of the pancreas and gut. J Clin Endocrinol Metab 71: 875–880.

249. Ong DE, Davis JT, O'Day WT, Bok D (1994) Synthesis and secretion of retinol-binding protein and transthyretin by cultured retinal pigment epithelium. Biochemistry 33: 1835–1842.
250. Soprano DR, Soprano KJ, Goodman DS (1986) Retinol-binding protein and transthyretin mRNA levels in visceral yolk sac and liver during fetal development in the rat. Proc Natl Acad Sci USA 83: 7330–7334.
251. Mizuno R, Cavallaro T, Herbert J (1992) Temporal expression of the transthyretin gene in the developing rat eye. Invest Ophthalmol Vis Sci 33: 341–349.
252. Barron M, McAllister D, Smith SM, Lough J (1998) Expression of retinol binding protein and transthyretin during early embryogenesis. Dev Dyn 212: 413–422. doi:10.1002/(SICI)1097-0177(199807)212:3<413::AID-AJA9>3.0.CO;2-K.
253. Achen MG, Harms PJ, Thomas T, Richardson SJ, Wettenhall RE, et al. (1992) Protein synthesis at the blood-brain barrier. The major protein secreted by amphibian choroid plexus is a lipocalin. J Biol Chem 267: 23170–23174.
254. Blake CC, Geisow MJ, Swan ID, Rerat C, Rerat B (1974) Structure of human plasma prealbumin at 2-5 A resolution. A preliminary report on the polypeptide chain conformation, quaternary structure and thyroxine binding. J Mol Biol 88: 1–12.
255. Blake CCF, Geisow MJ, Oatley SJ, Rérat B, Rérat C (1978) Structure of prealbumin: Secondary, tertiary and quaternary interactions determined by Fourier refinement at 1.8 Å. J Mol Biol 121: 339–356. doi:10.1016/0022-2836(78)90368-6.
256. Hutchinson EG, Thornton JM (1990) HERA - A program to draw schematic diagrams of protein secondary structures. Proteins 8: 203–212. doi:10.1002/prot.340080303.
257. Laskowski R a (2009) PDBsum new things. Nucl Acids Res 37: D355–9. doi:10.1093/nar/gkn860.
258. Ferguson RN, Edelhoch H, Saroff HA, Robbins J, Cahnmann HJ (1975) Negative cooperativity in the binding of thyroxine to human serum prealbumin. Preparation of tritium-labeled 8-anilino-1-naphthalenesulfonic acid. Biochemistry 14: 282–289.
259. Fuller JC, Burgoyne NJ, Jackson RM (2008) Predicting druggable binding sites at the protein–protein interface. Drug Discov Today 14: 155–161. doi:10.1016/j.drudis.2008.10.009.
260. Laurie ATR, Jackson RM (2005) Q-SiteFinder: An Energy-Based Method for the Prediction of Protein-Ligand Binding Sites. Bioinformatics 21: 1908–1916. doi:10.1093/bioinformatics/bti315.
261. Goodman DS (1976) Retinol-binding protein, prealbumin, and vitamin A transport. Prog Clin Biol Res 5: 313–330.
262. Goodman DS (1984) Vitamin A and retinoids in health and disease. N Engl J Med 310: 1023–1031. doi:10.1056/NEJM198404193101605.
263. van Jaarsveld PP, Edelhoch H, Goodman DWS, Robbins J (1973) The Interaction of Human Plasma Retinol-binding Protein with Prealbumin. J Biol Chem 248: 4698.
264. Noy N, Slosberg E, Scarlata S (1992) Interactions of retinol with binding proteins: studies with retinol-binding protein and with transthyretin. Biochemistry 31: 11118–11124.
265. Monaco HL, Rizzi M, Coda A (1995) Structure of a complex of two plasma proteins: transthyretin and retinol-binding protein. Science 268: 1039–1041.
266. Woeber K a, Ingbar SH (1968) The contribution of thyroxine-binding prealbumin to the binding of thyroxine in human serum, as assessed by immunoadsorption. J Clin Invest 47: 1710–1721. doi:10.1172/JCI105861.
267. Goodman DS (1987) Retinoids and retinol-binding proteins. The Harvey Lectures Series 81 81: 2518–2525.
268. Zanotti G, Cendron L, Ramazzina I, Folli C, Percudani R, et al. (2006) Structure of zebra fish HIUase: insights into evolution of an enzyme to a hormone transporter. J Mol Biol 363: 1–9. doi:10.1016/j.jmb.2006.07.079.
269. Lee Y, Lee DH, Kho CW, Lee AY, Jang M, et al. (2005) Transthyretin-related proteins function to facilitate the hydrolysis of 5-hydroxyisourate, the end product of the uricase reaction. FEBS Lett 579: 4769–4774. doi:10.1016/j.febslet.2005.07.056.

270. Ramazzina I, Folli C, Secchi A, Berni R, Percudani R (2006) Completing the uric acid degradation pathway through phylogenetic comparison of whole genomes. Nat Chem Biol 2: 144–148. doi:10.1038/nchembio768.
271. Episkopou V, Maeda S, Nishiguchi S, Shimada K, Gaitanaris G a, et al. (1993) Disruption of the transthyretin gene results in mice with depressed levels of plasma retinol and thyroid hormone. Proc Natl Acad Sci USA 90: 2375–2379.
272. Potter MA, Luxton G (1999) Prealbumin measurement as a screening tool for protein calorie malnutrition in emergency hospital admissions: a pilot study. Clin Invest Med 22: 44–52.
273. Ingenbleek Y, Young VR (2002) Significance of transthyretin in protein metabolism. Clin Chem Lab Med 40: 1281–1291. doi:10.1515/CCLM.2002.222.
274. Robbins J (2000) Thyroid hormone transport proteins and the physiology of hormone binding. In: Braverman LE, Utiger RD, editors. The Thyroid. Philadelphia: Lippincott Williams & Wilkins. pp. 105–120.
275. Richardson SJ (2002) The evolution of transthyretin synthesis in vertebrate liver, in primitive eukaryotes and in bacteria. Clin Chem Lab Med 40: 1191–1199. doi:10.1515/CCLM.2002.209.
276. Vahlquist A, Peterson PA, Wibell L (1973) Metabolism of the viatmin A transporting protein complex. I. Turnover studies in normal persons and in patients with chronic renal failure. Eur J Clin Invest 3: 352–362.
277. Liz MA, Faro CJ, Saraiva MJ, Sousa MM (2004) Transthyretin, a new cryptic protease. J Biol Chem 279: 21431–21438. doi:10.1074/jbc.M402212200.
278. Liz MA, Sousa MM (2005) Deciphering cryptic proteases. Cell Mol Life Sci 62: 989–1002. doi:10.1007/s00018-005-4544-2.
279. Fleming CE, Saraiva MJ, Sousa MM (2007) Transthyretin enhances nerve regeneration. J Neurochem 103: 831–839. doi:10.1111/j.1471-4159.2007.04828.x.
280. Sousa JC, Grandela C, Fernández-Ruiz J, de Miguel R, de Sousa L, et al. (2004) Transthyretin is involved in depression-like behaviour and exploratory activity. J Neurochem 88: 1052–1058.
281. Nunes AF, Saraiva MJ, Sousa MM (2006) Transthyretin knockouts are a new mouse model for increased neuropeptide Y. FASEB J 20: 166–168. doi:10.1096/fj.05-4106fje.
282. Sousa JC, Marques F, Dias-Ferreira E, Cerqueira JJ, Sousa N, et al. (2007) Transthyretin influences spatial reference memory. Neurobiol Learn Mem 88: 381–385. doi:10.1016/j.nlm.2007.07.006.
283. Schwarzman a L, Gregori L, Vitek MP, Lyubski S, Strittmatter WJ, et al. (1994) Transthyretin sequesters amyloid beta protein and prevents amyloid formation. Proc Natl Acad Sci USA 91: 8368–8372.
284. Mazur-Kolecka B, Frackowiak J, Wiśniewski HM (1995) Apolipoproteins E3 and E4 induce, and transthyretin prevents accumulation of the Alzheimer's beta-amyloid peptide in cultured vascular smooth muscle cells. Brain Res 698: 217–222.
285. Choi SH, Leight SN, Lee VM-Y, Li T, Wong PC, et al. (2007) Accelerated Abeta deposition in APPswe/PS1deltaE9 mice with hemizygous deletions of TTR (transthyretin). J Neurosci 27: 7006–7010. doi:10.1523/JNEUROSCI.1919-07.2007.
286. Buxbaum JN, Ye Z, Reixach N, Friske L, Levy C, et al. (2008) Transthyretin protects Alzheimer's mice from the behavioral and biochemical effects of Abeta toxicity. Proc Natl Acad Sci USA 105: 2681–2686.
287. Roche M, Rondeau P, Singh NR, Tarnus E, Bourdon E (2008) The antioxidant properties of serum albumin. FEBS Lett 582: 1783–1787. doi:10.1016/j.febslet.2008.04.057.
288. Liz M a, Mar FM, Franquinho F, Sousa MM (2010) Aboard transthyretin: From transport to cleavage. IUBMB Life 62: 429–435. doi:10.1002/iub.340.
289. Miller SR, Sekijima Y, Kelly JW (2004) Native state stabilization by NSAIDs inhibits transthyretin amyloidogenesis from the most common familial disease variants. Lab Invest 84: 545–552. doi:10.1038/labinvest.3700059.

290. Saraiva MJ, Birken S, Costa PP, Goodman DS (1984) Family studies of the genetic abnormality in transthyretin (prealbumin) in Portuguese patients with familial amyloidotic polyneuropathy. Ann N Y Acad Sci 435: 86–100.
291. Connors LH, Richardson AM, Theberge R, Costello CE (2000) Tabulation of transthyretin (TTR) variants as of 1/1/2000. Proc Natl Acad Sci USA 7: 54–69.
292. Eneqvist T, Sauer-Eriksson AE (2001) Structural distribution of mutations associated with familial amyloidotic polyneuropathy in human transthyretin. Amyloid 8: 149–168.
293. Saraiva MJ (2001) Transthyretin mutations in hyperthyroxinemia and amyloid diseases. Human Mutat 17: 493–503. doi:10.1002/humu.1132.
294. Connors LH, Lim A, Prokaeva T, Roskens VA, Costello CE (2003) Tabulation of human transthyretin (TTR) variants, 2003. Amyloid 10: 160–184. doi:10.3109/13506120308998998.
295. Ando Y, Nakamura M, Araki S (2005) Transthyretin-related familial amyloidotic polyneuropathy. Arch Neurol 62: 1057.
296. Holmgren G, Costa PMP, Andersson C, Asplund K, Steen L, et al. (1994) Original articles Geographical distribution of TTR met30 carriers in northern Sweden: discrepancy between carrier frequency and prevalence rate. J Med Genet: 351–354.
297. Continho P (1989) Familial amyloidotic polyneuropathy. Travels of a Gene: a Tale of the Migrations of Familial Amyloidotic Polyneuropathy (Portuguese Type). Porto: ROCHA/Artes Graficas Co. pp. 1–25.
298. Ohmori H (2004) Common origin of the Val30Met mutation responsible for the amyloidogenic transthyretin type of familial amyloidotic polyneuropathy. J Med Genet 41: e51–e51. doi:10.1136/jmg.2003.014803.
299. Jacobson DR, McFarlin DE, Kane I, Buxbaum JN (1992) Transthyretin Pro55, a variant associated with early-onset, aggressive, diffuse amyloidosis with cardiac and neurologic involvement. Hum Genet 89: 353–356.
300. Jacobson DR, Pastore RD, Yaghoubian R, Kane I, Gallo G, et al. (1997) Variant-sequence transthyretin (isoleucine 122) in late-onset cardiac amyloidosis in black Americans. N Engl J Med 336: 466–473. doi:10.1056/NEJM199702133360703.
301. Jenne DE, Denzel K, Blätzinger P, Winter P, Obermaier B, et al. (1996) A new isoleucine substitution of Val-20 in transthyretin tetramers selectively impairs dimer-dimer contacts and causes systemic amyloidosis. Proc Natl Acad Sci USA 93: 6302–6307.
302. Jacobson DR, Pastore R, Pool S, Malendowicz S, Kane I, et al. (1996) Revised transthyretin Ile 122 allele frequency in African-Americans. Hum Genet 98: 236–238.
303. Afolabi I, Hamidi Asl K, Nakamura M, Jacobs P, Hendrie H, et al. (2000) Transthyretin isoleucine-122 mutation in African and American blacks. Amyloid 7: 121–125.
304. Hammarström P, Schneider F, Kelly JW (2001) Trans-suppression of misfolding in an amyloid disease. Science 293: 2459–2462. doi:10.1126/science.1062245.
305. Sunde M, Blake CC (1998) From the globular to the fibrous state: protein structure and structural conversion in amyloid formation. Q Rev Biophys 31: 1–39.
306. Goldsteins G, Persson H, Andersson K, Olofsson A, Dacklin I, et al. (1999) Exposure of cryptic epitopes on transthyretin only in amyloid and in amyloidogenic mutants. Proc Natl Acad Sci USA 96: 3108–3113.
307. Serpell LC, Goldsteins G, Dacklin I, Lundgren E, Blake CCF (1996) The “edge strand” hypothesis: Prediction and test of a mutational “hot-spot” on the transthyretin molecule associated with FAP amyloidogenesis. Amyloid 3: 75–85. doi:10.3109/13506129609014359.
308. Serag AA, Altenbach C, Gingery M, Hubbell WL, Yeates TO (2002) Arrangement of subunits and ordering of beta-strands in an amyloid sheet. Nat Struct Biol 9: 734–739.
309. Olofsson A, Ippel JH, Wijmenga SS, Lundgren E, Ohman A (2004) Probing solvent accessibility of transthyretin amyloid by solution NMR spectroscopy. J Biol Chem 279: 5699–5707.

310. Pan K-ming, Baldwin M, Nguyen J, Gasset M, Serban ANA, et al. (1993) Conversion of alpha-helices into beta-sheets features in the formation of the scrapie prion proteins. Proc Natl Acad Sci USA 90: 10962–10966.
311. Cendron L, Trovato A, Seno F, Folli C, Alfieri B, et al. (2009) Amyloidogenic potential of transthyretin variants: insights from structural and computational analyses. J Biol Chem 284: 25832–25841. doi:10.1074/jbc.M109.017657.
312. Small DH, Mok SS, Bornstein JC (2001) Alzheimer's disease and Abeta toxicity: from top to bottom. Nat Rev Neurosci 2: 595–598. doi:10.1038/35086072.
313. Hou X, Aguilar M-isabel, Small DH (2007) Transthyretin and Familial Amyloidotic Polyneuropathy. FEBS Journal 274: 1637–1650. doi:10.1111/j.1742-4658.2007.05712.x.
314. Monteiro FA, Sousa MM, Cardoso I, do Amaral JB, Guimarães A, et al. (2006) Activation of ERK1/2 MAP kinases in familial amyloidotic polyneuropathy. J Neurochem 97: 151–161. doi:10.1111/j.1471-4159.2006.03716.x.
315. Teixeira PF, Cerca F, Santos SD, Saraiva MJ (2006) Endoplasmic reticulum stress associated with extracellular aggregates. Evidence from transthyretin deposition in familial amyloid polyneuropathy. J Biol Chem 281: 21998–22003. doi:10.1074/jbc.M602302200.
316. Murphy JE, Tedbury PR, Homer-Vanniasinkam S, Walker JH, Ponnambalam S (2005) Biochemistry and cell biology of mammalian scavenger receptors. Atherosclerosis 182: 1–15. doi:10.1016/j.atherosclerosis.2005.03.036.
317. Olofsson B-O, Backman C, Karp K, Suhr OB (2002) Progression of cardiomyopathy after liver transplantation in patients with familial amyloidotic polyneuropathy, Portuguese type. Transplantation 73: 745–751.
318. Rudolph T, Kurz MW, Farbu E (2008) Late-onset familial amyloid polyneuropathy (FAP) Val30Met without family history. Clin Med Res 6: 80–82. doi:10.3121/cmr.2008.794.
319. Small DH, Mok SS, Williamson TG, Nurcombe V (1996) Role of proteoglycans in neural development, regeneration, and the aging brain. J Neurochem 67: 889–899.
320. Sobue G, Nakao N, Murakami K, Yasuda T, Sahashi K, et al. (1990) Type I familial amyloid polyneuropathy. A pathological study of the peripheral nervous system. Brain 113 (Pt 4: 903–919.
321. Dubový P, Klusáková I, Svízenská I (2002) A quantitative immunohistochemical study of the endoneurium in the rat dorsal and ventral spinal roots. Histochem Cell Biol 117: 473–480. doi:10.1007/s00418-002-0411-5.
322. Wisniewski HM, Wegiel J (1994) Beta-amyloid formation by myocytes of leptomeningeal vessels. Acta Neuropathol 87: 233–241.
323. Ancsin JB (2003) Amyloidogenesis: historical and modern observations point to heparan sulfate proteoglycans as a major culprit. Amyloid 10: 67–79.
324. Bergamaschini L, Rossi E, Storini C, Pizzimenti S, Distaso M, et al. (2004) Peripheral treatment with enoxaparin, a low molecular weight heparin, reduces plaques and beta-amyloid accumulation in a mouse model of Alzheimer's disease. J Neurosci 24: 4181–4186. doi:10.1523/JNEUROSCI.0550-04.2004.
325. Li J-P, Galvis MLE, Gong F, Zhang X, Zcharia E, et al. (2005) In vivo fragmentation of heparan sulfate by heparanase overexpression renders mice resistant to amyloid protein A amyloidosis. Proc Natl Acad Sci USA 102: 6473–6477. doi:10.1073/pnas.0502287102.
326. Kisilevsky R, Ancsin JB, Szarek WA, Petanceska S (2007) Heparan sulfate as a therapeutic target in amyloidogenesis: prospects and possible complications. Amyloid 14: 21–32. doi:10.1080/13506120601116419.
327. Richard JP (1993) Mechanism for the formation of methylglyoxal from triosephosphates. Biochem Soc Trans 21: 549–553.
328. Du Yan S, Zhu H, Fu J, Yan SF, Roher A, et al. (1997) Amyloid-beta peptide-receptor for advanced glycation endproduct interaction elicits neuronal expression of macrophage-colony stimulating factor: a proinflammatory pathway in Alzheimer disease. Proc Natl Acad Sci USA 94: 5296–5301.

329. Vitek MP, Bhattacharya K, Glendening JM, Stopa E, Vlassara H, et al. (1994) Advanced glycation end products contribute to amyloidosis in Alzheimer disease. Proc Natl Acad Sci USA 91: 4766–4770.
330. Yan SD, Chen X, Schmidt a M, Brett J, Godman G, et al. (1994) Glycated tau protein in Alzheimer disease: a mechanism for induction of oxidant stress. Proc Natl Acad Sci USA 91: 7787–7791.
331. Münch G, Lüth HJ, Wong A, Arendt T, Hirsch E, et al. (2000) Crosslinking of alpha-synuclein by advanced glycation endproducts--an early pathophysiological step in Lewy body formation? J Chem Neuroanat 20: 253–257.
332. Miyata T, Inagi R, Iida Y, Sato M, Yamada N, et al. (1994) Involvement of beta 2-microglobulin modified with advanced glycation end products in the pathogenesis of hemodialysis-associated amyloidosis. Induction of human monocyte chemotaxis and macrophage secretion of tumor necrosis factor-alpha and interleukin-1. J Clin Invest 93: 521–528. doi:10.1172/JCI117002.
333. Sasaki N, Takeuchi M, Chowei H, Kikuchi S, Hayashi Y, et al. (2002) Advanced glycation end products (AGE) and their receptor (RAGE) in the brain of patients with Creutzfeldt-Jakob disease with prion plaques. Neurosci Lett 326: 117–120.
334. Miyata T, Oda O, Inagi R, Iida Y, Araki N, et al. (1993) beta 2-Microglobulin modified with advanced glycation end products is a major component of hemodialysis-associated amyloidosis. J Clin Invest 92: 1243–1252. doi:10.1172/JCI116696.
335. Hoshii Y, Kawano H, Gondo T, Takahashi M, Ishihara T, et al. (1996) Immunohistochemical study with anti-advanced glycation end-products antibody in murine amyloidosis. Pathol Int 46: 738–742.
336. Gomes R, Sousa Silva M, Quintas A, Cordeiro C, Freire A, et al. (2005) Argpyrimidine, a methylglyoxal-derived advanced glycation end-product in familial amyloidotic polyneuropathy. Biochem J 385: 339–345. doi:10.1042/BJ20040833.
337. Costa G da, Gomes R, Guerreiro A, Mateus (2011) Beyond Genetic Factors in Familial Amyloidotic Polyneuropathy: Protein Glycation and the Loss of Fibrinogen's Chaperone Activity. PLoS ONE 6: e24850. doi:10.1371/journal.pone.0024850.
338. De Lorenzi E, Giorgetti S, Grossi S, Merlini G, Caccialanza G, et al. (2004) Pharmaceutical Strategies Against Amyloidosis: Old and New Drugs in Targeting a "Protein Misfolding Disease." Curr Med Chem 11: 1065–1084.
339. Soto C, Martin Z (2009) Therapeutic strategies against protein misfolding in neurodegenerative diseases. Expert Opin Drug Discov 4: 71–84. doi:10.1517/13543770802630455.
340. Hind CR, Collins PM, Caspi D, Baltz ML, Pepys MB (1984) Specific chemical dissociation of fibrillar and non-fibrillar components of amyloid deposits. Lancet 2: 376–378.
341. Kisilevsky R, Lemieux LJ, Fraser PE, Kong X, Hultin PG, et al. (1995) Arresting amyloidosis in vivo using small-molecule anionic sulphonates or sulphates: implications for Alzheimer's disease. Nat Med 1: 143–148.
342. Gervais F, Chalifour R, Garceau D, Kong X, Laurin J, et al. (2001) Glycosaminoglycan mimetics: a therapeutic approach to cerebral amyloid angiopathy. Amyloid 8 Suppl 1: 28–35.
343. Pepys MB, Herbert J, Hutchinson WL, Tennent GA, Lachmann HJ, et al. (2002) Targeted pharmacological depletion of serum amyloid P component for treatment of human amyloidosis. Nature 417: 254–259. doi:10.1038/417254a.
344. Zagorski MG, Yang J, Shao H, Ma K, Zeng H, et al. (1999) Methodological and chemical factors affecting amyloid beta-peptide amyloidogenicity. In: Wetzel R, editor. Amyloid, Prions, and Other Protein Aggregates. New York: Academic Press. pp. 189–204.
345. Lorenzo A, Yankner BA (1994) Beta-amyloid neurotoxicity requires fibril formation and is inhibited by Congo red. Proc Natl Acad Sci USA 91: 12243–12247.
346. Podlisny MB, Walsh DM, Amarante P, Ostaszewski BL, Stimson ER, et al. (1998) Oligomerization of endogenous and synthetic amyloid beta-protein at nanomolar

levels in cell culture and stabilization of monomer by Congo red. Biochemistry 37: 3602–3611. doi:10.1021/bi972029u.
347. Esler WP, Stimson ER, Ghilardi JR, Felix AM, Lu YA, et al. (1997) A beta deposition inhibitor screen using synthetic amyloid. Nat Biotechnol 15: 258–263. doi:10.1038/nbt0397-258.
348. Rudyk H, Vasiljevic S, Hennion RM, Birkett CR, Hope J, et al. (2000) Screening Congo Red and its analogues for their ability to prevent the formation of PrP-res in scrapie-infected cells. J Gen Virol 81: 1155–1164.
349. Kim Y-S, Randolph TW, Manning MC, Stevens FJ, Carpenter JF (2003) Congo red populates partially unfolded states of an amyloidogenic protein to enhance aggregation and amyloid fibril formation. J Biol Chem 278: 10842–10850. doi:10.1074/jbc.M212540200.
350. Li L, Darden TA, Bartolotti L, Kominos D, Pedersen LG (1999) An atomic model for the pleated beta-sheet structure of Abeta amyloid protofilaments. Biophys J 76: 2871–2878. doi:10.1016/S0006-3495(99)77442-4.
351. Wolfe MS (2002) Therapeutic strategies for Alzheimer's disease. Nat Rev Drug Discov 1: 859–866. doi:10.1038/nrd938.
352. Xia W (2003) Amyloid inhibitors and Alzheimer's disease. Curr Opin Investig Drugs 4: 55–59.
353. Morozova-Roche L, Malisauskas M (2007) A false paradise - mixed blessings in the protein universe: the amyloid as a new challenge in drug development. Curr Med Chem 14: 1221–1230.
354. Takahashi T, Mihara H (2008) Peptide and protein mimetics inhibiting amyloid beta-peptide aggregation. Acc Chem Res 41: 1309–1318. doi:10.1021/ar8000475.
355. Tomiyama T, Asano S, Suwa Y, Morita T, Kataoka K, et al. (1994) Rifampicin prevents the aggregation and neurotoxicity of amyloid beta protein in vitro. Biochem Biophys Res Commun 204: 76–83. doi:10.1006/bbrc.1994.2428.
356. Tomiyama T, Shoji A, Kataoka K, Suwa Y, Asano S, et al. (1996) Inhibition of amyloid beta protein aggregation and neurotoxicity by rifampicin. Its possible function as a hydroxyl radical scavenger. J Biol Chem 271: 6839–6844.
357. Pappolla M, Bozner P, Soto C, Shao H, Robakis NK, et al. (1998) Inhibition of Alzheimer beta-fibrillogenesis by melatonin. J Biol Chem 273: 7185–7188.
358. Ono K, Hasegawa K, Yoshiike Y, Takashima A, Yamada M, et al. (2002) Nordihydroguaiaretic acid potently breaks down pre-formed Alzheimer's beta-amyloid fibrils in vitro. J Neurochem 81: 434–440.
359. Camilleri P, Haskins NJ, Howlett DR (1994) beta-Cyclodextrin interacts with the Alzheimer amyloid beta-A4 peptide. FEBS Lett 341: 256–258.
360. Lashuel H a, Hartley DM, Balakhaneh D, Aggarwal A, Teichberg S, et al. (2002) New class of inhibitors of amyloid-beta fibril formation. Implications for the mechanism of pathogenesis in Alzheimer's disease. J Biol Chem 277: 42881–42890. doi:10.1074/jbc.M206593200.
361. Sood A, Abid M, Hailemichael S, Foster M, Török B, et al. (2009) Effect of chirality of small molecule organofluorine inhibitors of amyloid self-assembly on inhibitor potency. Bioorg Med Chem Lett 19: 6931–6934. doi:10.1016/j.bmcl.2009.10.066.
362. Cardoso I (2002) Dynamics of transthyretin fibrillogenesis: contribution to therapeutic approaches in familial amyloidotic polyneuropathy Universidade do Porto.
363. Gianni BL, Bellotti V, Gianni AM, Merlini G (1995) New Drug Therapy of Amyloidoses: Resorption of AL-Type Deposits With 4'-Iodo-4'-Deoxydoxorubicin. Blood 86: 855–861.
364. Merlini G, Ascari E, Amboldit N, Bellotti V, Arbustinit E, et al. (1995) Interaction of the anthracycline 4'-iodo-4'-deoxydoxorubicin with amyloid fibrils: Inhibition of amyloidogenesis. Proc Natl Acad Sci USA 92: 2959–2963.
365. Palha JA, Ballinari D, Amboldi N, Cardoso I, Fernandes R, et al. (2000) 4'-Iodo-4'-deoxydoxorubicin disrupts the fibrillar structure of transthyretin amyloid. Am J Pathol 156: 1919–1925. doi:10.1016/S0002-9440(10)65065-1.
366. Sebastião MP, Merlini G, Saraiva MJ, Damas a M (2000) The molecular interaction of 4'-iodo-4'-deoxydoxorubicin with Leu-55Pro transthyretin "amyloid-like" oligomer leading to disaggregation. Biochem J 351: 273–279.

367. Lansbury PT (1999) Commentary Evolution of amyloid : What normal protein folding may tell us about fibrillogenesis and disease. Proc Natl Acad Sci USA 96: 3342–3344.
368. Xue W-F, Hellewell AL, Gosal WS, Homans SW, Hewitt EW, et al. (2009) Fibril fragmentation enhances amyloid cytotoxicity. J Biol Chem 284: 34272–34282. doi:10.1074/jbc.M109.049809.
369. Xue W-feng, Hellewell AL, Hewitt EW, Radford SE (2010) Fibril fragmentation in amyloid assembly and cytotoxicity: When size matters. Prion 4: 20–25.
370. Saraiva O, Cardoso I, Merlini G, Joa M, Saraiva MJ (2003) 4'-iodo-4'-deoxydoxorubicin and tetracyclines disrupt transthyretin amyloid fibrils in vitro producing noncytotoxic species: screening for TTR fibril disrupters. FASEB J 17: 803–809. doi:10.1096/fj.02-0764com.
371. Lewis WD, Skinner M, Simms RW, Jones LA, Cohen AS, et al. (1994) Orthotopic liver transplantation for familial amyloidotic polyneuropathy. Clin Transplant 8: 107–110.
372. Stangou AJ, Hawkins PN, Heaton ND, Rela M, Monaghan M, et al. (1998) Progressive cardiac amyloidosis following liver transplantation for familial amyloid polyneuropathy: implications for amyloid fibrillogenesis. Transplantation 66: 229–233.
373. Holmgren G, Ericzon BG, Groth CG, Steen L, Suhr O, et al. (1993) Clinical improvement and amyloid regression after liver transplantation in hereditary transthyretin amyloidosis. Lancet 341: 1113–1116.
374. Ando E, Ando Y, Haraoka K (2001) Ocular amyloid involvement after liver transplantation for polyneuropathy. Ann Intern Med 135: 931–932.
375. Ando Y, Terazaki H, Haraoka K, Tajiri T, Nakamura M, et al. (2002) Presence of autoantibody against ATTR Val30Met after sequential liver transplantation. Transplantation 73: 751–755.
376. Blaney JM, Jorgensen EC, Connolly ML, Ferrin TE, Langridge R, et al. (1982) Computer Graphics in Drug Design: Molecular Modeling of Thyroid Hormone-Prealbumin Interactions. J Med Chem: 785–790.
377. Peterson S a, Klabunde T, Lashuel H a, Purkey H, Sacchettini JC, et al. (1998) Inhibiting transthyretin conformational changes that lead to amyloid fibril formation. Proc Natl Acad Sci USA 95: 12956–12960.
378. Johnson SM, Connelly S, Wilson IA, Kelly JW (2008) Biochemical and Structural Evaluation of Highly Selective 2-Arylbenzoxazole-Based Transthyretin Amyloidogenesis Inhibitors. J Med Chem 51: 260–270.
379. Choi S, Reixach N, Connelly S, Johnson SM, Wilson IA, et al. (2010) A Substructure Combination Strategy to Create Potent and Selective Transthyretin Kinetic Stabilizers That Prevent Amyloidogenesis and Cytotoxicity. J Am Chem Soc 132: 1359–1370. doi:10.1021/ja908562q.
380. Johnson SM, Connelly S, Wilson IA, Kelly JW (2009) Toward Optimization of the Second Aryl Substructure Common to Transthyretin Amyloidogenesis Inhibitors Using Biochemical and Structural Studies. J Med Chem 52: 1115–1125. doi:10.1021/jm801347s.
381. Kubinyi H (2007) Computational and Structural Approaches to Drug Discovery. Stroud R, editor Cambridge: Royal Society of Chemistry. p. doi:10.1039/9781847557964.
382. Hernandes MZ, Cavalcanti SMT, Moreira DRM, de Azevedo Junior WF, Leite ACL (2010) Halogen atoms in the modern medicinal chemistry: hints for the drug design. Curr Drug Targets 11: 303–314.
383. Cheng SY, Pages RA, Saroff HA, Edelhoch H, Robbins J (1977) Analysis of thyroid hormone binding to human serum prealbumin by 8-anilinonaphthalene-1-sulfonate fluorescence. Biochemistry 16: 3707–3713.
384. Irace G, Edelhoch H (1978) Thyroxine-Induced Conformational Changes in Prealbumin. Biochemistry 17: 5729–5733.
385. Neumann P, Cody V, Wojtczak A (2001) Structural basis of negative cooperativity in transthyretin. Acta Biochim Polonica 48.

386. Bertelli AA, Giovannini L, Giannessi D, Migliori M, Bernini W, et al. (1995) Antiplatelet activity of synthetic and natural resveratrol in red wine. Int J Tissue React 17: 1–3.
387. Demrow HS, Slane PR, Folts JD (1995) Administration of wine and grape juice inhibits in vivo platelet activity and thrombosis in stenosed canine coronary arteries. Circulation 91: 1182–1188.
388. MacCarrone M, Lorenzon T, Guerrieri P, Agrò AF (1999) Resveratrol prevents apoptosis in K562 cells by inhibiting lipoxygenase and cyclooxygenase activity. Eur J Biochem 265: 27–34.
389. Green NS, Foss TR, Kelly JW (2005) Genistein, a natural product from soy, is a potent inhibitor of transthyretin amyloidosis. Proc Natl Acad Sci USA 102: 14545–14550. doi:10.1073/pnas.0501609102.
390. Ferreira N, Saraiva MJ, Almeida MR (2011) Natural polyphenols inhibit different steps of the process of transthyretin (TTR) amyloid fibril formation. FEBS Lett 585: 2424–2430. doi:10.1016/j.febslet.2011.06.030.
391. Rickenbacher U (1986) Structurally Specific Binding of Halogenated Biphenyls to Thyroxine Transport Protein. J Med Chem 29: 641–648.
392. Petrassi HM, Klabunde T, Sacchettini J, Kelly JW (2000) Structure-Based Design of N-Phenyl Phenoxazine Transthyretin Amyloid Fibril Inhibitors. J Am Chem Soc 122: 2178–2192. doi:10.1021/ja993309v.
393. Gales L, Macedo-Ribeiro S, Arsequell G, Valencia G, Saraiva MJ, et al. (2005) Human transthyretin in complex with iododiflunisal: structural features associated with a potent amyloid inhibitor. Biochem J 388: 615–621. doi:10.1042/BJ20042035.
394. Mairal T, Nieto J, Pinto M, Rosa M (2009) Iodine Atoms: A New Molecular Feature for the Design of Potent Transthyretin Fibrillogenesis Inhibitors. PLoS ONE 4: e4124–e4137. doi:10.1371/journal.pone.0004124.
395. Gupta S, Chhibber M, Sinha S, Surolia A (2007) Design of Mechanism-Based Inhibitors of Transthyretin Amyloidosis: Studies with Biphenyl Ethers and New Structural Templates. J Med Chem 50: 5589–5599. doi:10.1021/jm0700159.
396. Gupta S, Babu P, Surolia A (2010) Biphenyl ethers conjugated CdSe/ZnS core/shell quantum dots and interpretation of the mechanism of amyloid fibril disruption. Biomaterials 31: 6809–6822. doi:10.1016/j.biomaterials.2010.05.031.
397. Johnson SM, Petrassi HM, Palaninathan SK, Mohamedmohaideen NN, Purkey HE, et al. (2005) Bisaryloxime Ethers as Potent Inhibitors of Transthyretin Amyloid Fibril Formation. J Med Chem 48: 1576–1587. doi:10.1021/jm049274d.
398. Razavi H, Powers ET, Purkey HE, Adamski-Werner SL, Chiang KP, et al. (2005) Design, synthesis, and evaluation of oxazole transthyretin amyloidogenesis inhibitors. Bioorg Med Chem 15: 1075–1078. doi:10.1016/j.bmcl.2004.12.022.
399. Green NS, Palaninathan SK, Sacchettini JC, Kelly JW (2003) Synthesis and characterization of potent bivalent amyloidosis inhibitors that bind prior to transthyretin tetramerization. J Am Chem Soc 125: 13404–13414. doi:10.1021/ja030294z.
400. Kolstoe SE, Mangione PP, Bellotti V, Taylor GW, Tennent G a, et al. (2010) Trapping of palindromic ligands within native transthyretin prevents amyloid formation. Proc Natl Acad Sci USA: 2–7. doi:10.1073/pnas.1008255107.
401. Choi S, Connelly S, Wilson IA, Kelly JW (2010) Chemoselective Small Molecules That Covalently Modify One Lysine in A Non-Enzyme Protein in Plasma. Nat Chem Biol 6: 133–139. doi:10.1038/NCHEMBIO.281.
402. Zsoldos Z, Reid D, Simon A, Sadjad BS, Johnson AP (2006) eHiTS: An Innovative Approach to the Docking and Scoring Function Problems. Curr Protein Pept Sci 7: 421–435.
403. Brooijmans N, Kuntz ID (2003) Molecular Recognition & Docking Algorithms. Review Literature And Arts Of The Americas. doi:10.1146/annurev.biophys.32.110601.142532.
404. Sperandio O (2006) Receptor-based computational screening of compound databases: the main docking-scoring engines. Curr Protein Pept Sci 7: 369–393.

405. Halperin I, Ma B, Wolfson H, Nussinov R (2002) Principles of docking: An overview of search algorithms and a guide to scoring functions. Proteins 47: 409–443. doi:10.1002/prot.10115.
406. Meng X-Y, Zhang H-X, Mezei M, Cui M (2011) Molecular docking: a powerful approach for structure-based drug discovery. Current Comput Aided Drug Des 7: 146–157.
407. Ekins S, Mestres J, Testa B (2007) In silico pharmacology for drug discovery: applications to targets and beyond. Br J Pharmacol 152: 21–37. doi:10.1038/sj.bjp.0707306.
408. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. (2000) The Protein Data Bank. Nucleic Acids Res 28: 235–242.
409. Hornberg A, Eneqvist T, Olofsson A, Lundgren E, Sauer-eriksson AE (2000) A comparative analysis of 23 structures of the amyloidogenic protein transthyretin. J Mol Biol 302: 649–669.
410. Blake CC, Oatley SJ (1977) Protein-DNA and protein-hormone interactions in prealbumin: a model of the thyroid hormone nuclear receptor? Nature 268: 115–120.
411. Murray-Rust P, Motherwell WDS (1979) Computer retrieval and analysis of molecular geometry. 4. Intermolecular interactions. J Am Chem Soc 101: 4374–4376. doi:10.1021/ja00509a056.
412. Brakoulias A, Jackson RM (2004) Towards a Structural Classification of Phosphate Binding Sites in Protein–Nucleotide Complexes: An Automated All- Against-All Structural Comparison Using Geometric Matching. Biochemistry 260: 250–260. doi:10.1002/prot.20123.
413. Gold ND, Jackson RM (2006) SitesBase: a database for structure-based protein–ligand binding site comparisons. Database 34: 231–234. doi:10.1093/nar/gkj062.
414. Cole JC, Murray CW, Nissink JWM, Taylor RD, Taylor R (2005) Comparing Protein – Ligand Docking Programs Is Difficult. Proteins 332: 325–332. doi:10.1002/prot.20497.
415. Wang R, Lu Y, Wang S (2003) Comparative Evaluation of 11 Scoring Functions for Molecular Docking. J Med Chem 46: 2287–2303.
416. Kellenberger E, Rodrigo J, Muller P, Rognan D (2004) Comparative evaluation of eight docking tools for docking and virtual screening accuracy. Proteins 57: 225–242. doi:10.1002/prot.20149.
417. Verkhivker GM, Bouzida D, Gehlhaar DK, Rejto PA, Arthurs S, et al. (2000) Deciphering Common Failures in Molecular Docking of Ligand-Protein Complexes. J Comput Aided Mol Des 14: 731–751.
418. Onodera K, Satou K, Hirota H (2007) Evaluations of Molecular Docking Programs for Virtual Screening. J Chem Inf Model 47: 1609–1618. doi:10.1021/ci7000378.
419. Warren GL, Andrews CW, Capelli A-M, Clarke B, LaLonde J, et al. (2006) A Critical Assessment of Docking Programs and Scoring Functions. J Med Chem 49: 5912–5931. doi:10.1021/jm050362n.
420. Cross JB, Thompson DC, Rai BK, Baber JC, Fan KY, et al. (2009) Comparison of several molecular docking programs: pose prediction and virtual screening accuracy. J Chem Inf Model 49: 1455–1474. doi:10.1021/ci900056c.
421. Verdonk ML, Berdini V, Hartshorn MJ, Mooij WTM, Murray CW, et al. (2004) Virtual screening using protein-ligand docking: avoiding artificial enrichment. J Chem Inf Comput Sci 44: 793–806. doi:10.1021/ci034289q.
422. Cavasotto CN, Abagyan RA (2004) Protein flexibility in ligand docking and virtual screening to protein kinases. J Mol Biol 337: 209–225. doi:10.1016/j.jmb.2004.01.003.
423. Cavasotto CN, Kovacs JA, Abagyan RA (2005) Representing receptor flexibility in ligand docking through relevant normal modes. J Am Chem Soc 127: 9632–9640. doi:10.1021/ja042260c.
424. Claussen H, Buning C, Rarey M, Lengauer T (2001) FlexE: efficient molecular docking considering protein structure variations. J Mol Biol 308: 377–395. doi:10.1006/jmbi.2001.4551.

425. B-Rao C, Subramanian J, Sharma SD (2009) Managing protein flexibility in docking and its applications. Drug Discov Today 14: 394–400. doi:10.1016/j.drudis.2009.01.003.
426. Hartmann C, Antes I, Lengauer T (2009) Docking and scoring with alternative side-chain conformations. Proteins 74: 712–726. doi:10.1002/prot.22189.
427. Davis IW, Baker D (2009) RosettaLigand docking with full ligand and receptor flexibility. J Mol Biol 385: 381–392. doi:10.1016/j.jmb.2008.11.010.
428. Meiler J, Baker D (2006) ROSETTALIGAND: protein-small molecule docking with full side-chain flexibility. Proteins 65: 538–548. doi:10.1002/prot.21086.
429. Seeliger D, de Groot BL (2010) Conformational transitions upon ligand binding: holo-structure prediction from apo conformations. PLoS Comput Biol 6: e1000634. doi:10.1371/journal.pcbi.1000634.
430. Okimoto N, Futatsugi N, Fuji H, Suenaga A, Morimoto G, et al. (2009) High-performance drug discovery: computational screening by combining docking and molecular dynamics simulations. PLoS Comput Biol 5: e1000528. doi:10.1371/journal.pcbi.1000528.
431. Nichols SE, Baron R, Ivetac AD, McCammon JA (2011) Predictive Power of Molecular Dynamics Receptor Structures in Virtual Screening. J Chem Inf Model: Epub. doi:10.1021/ci200117n.
432. Jayachandran G, Shirts MR, Park S, Pande VS (2006) Parallelized-over-parts computation of absolute binding free energy with docking and molecular dynamics. J Chem Phys 125: 084901. doi:10.1063/1.2221680.
433. Bursulaya BD, Totrov M, Abagyan R, Brooks CL (2003) Comparative study of several algorithms for flexible ligand docking. J Comput Aided Mol Des 17: 755–763.
434. Novikov FN, Zeifman AA, Stroganov OV, Stroylov VS, Kulkov V, et al. (2011) CSAR Scoring Challenge Reveals the Need for New Concepts in Estimating Protein-Ligand Binding Affinity. J Chem Inf Model. doi:10.1021/ci200034y.
435. Goodsell DS, Olson AJ (1990) Automated docking of substrates to proteins by simulated annealing. Proteins 8: 195–202. doi:10.1002/prot.340080302.
436. Sousa S, Fernandes PA, Ramos MJ (2006) Protein–Ligand Docking: Current Status and Future Challenges. Proteins 26: 15–26. doi:10.1002/prot.
437. Morris GM, Goodsell DS, Huey R, Olson AJ (1996) Distributed automated docking of flexible ligands to proteins: parallel applications of AutoDock 2.4. J Comput Aided Mol Des 10: 293–304.
438. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, et al. (2009) AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. J Comput Chem 30: 2785–2791. doi:10.1002/jcc.
439. Trott O, Olson AJ (2010) AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem 31: 455–461. doi:10.1002/jcc.21334.
440. Chang MW, Ayeni C, Breuer S, Torbett BE (2010) Virtual screening for HIV protease inhibitors: a comparison of AutoDock 4 and Vina. PloS ONE 5: 1–9. doi:10.1371/journal.pone.0011955.
441. Zsoldos Z, Reid D, Simon A, Sadjad SB, Johnson AP (2007) eHiTS: A New Fast, Exhaustive Flexible Ligand Docking System. J Mol Graph Model 26: 198–212. doi:10.1016/j.jmgm.2006.06.002.
442. Boström J, Greenwood JR, Gottfries J (2003) Assessing the Performance of OMEGA with Respect to Retrieving Bioactive Conformations. J Mol Graph Model 21: 449–462.
443. McGann M (2011) FRED Pose Prediction and Virtual Screening Accuracy. J Chem Inf Model. doi:10.1021/ci100436p.
444. Olsen L, Pettersson I, Hemmingsen L, Adolph H-W, Jørgensen FS (2004) Docking and scoring of metallo-beta-lactamases inhibitors. J Comput Aided Mol Des 18: 287–302.
445. Verdonk ML, Cole JC, Hartshorn MJ, Murray CW, Taylor RD (2003) Improved protein-ligand docking using GOLD. Proteins 52: 609–623. doi:10.1002/prot.10465.
446. Taylor RD, Jewsbury PJ, Essex JW (2002) A review of protein-small molecule docking methods. J Comput Aided Mol Des 16: 151–166.

447. Perola E, Walters WP, Charifson PS (2004) A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. Proteins 56: 235–249. doi:10.1002/prot.20088.
448. Wang R, Lu Y, Fang X, Wang S (2004) An Extensive Test of 14 Scoring Functions Using the PDBbind Refined Set of 800 Protein-Ligand Complexes. J Chem Inf Comput Sci 44: 2114–2125.
449. Krovat EM, Steindl T, Langer T (2005) Recent Advances in Docking and Scoring. Curr Comput Aided Drug Des 1: 93–102. doi:10.2174/1573409052952314.
450. Hooft RWW, Vriend G, Sander C, Abola EE (1996) Errors in protein structures. Nature 381: 272.
451. Vriend G (1990) WHAT IF: A Molecular Modeling and Drug Design program. J Mol Graph 8: 52–56, 29.
452. Laskowski RA, MacArthur MW, Moss DS, Thornton JM (1993) PROCHECK - A Program to Check the Stereochemical Quality of Protein Structures. J Appl Cryst 26: 283–291.
453. Henrick K, Thornton JM (1998) PQS: a protein quaternary structure file server. Trends Biochem Sci 23: 358–361.
454. Karypis G (2002) Cluto: A Software Package for Clustering High Dimensional Datasets. Release 1.5. Department of Computer Science, University of Minnesota.
455. Zhao Y, Karypis G (2005) Data Clustering in Life Sciences. Molecular Biotechnology 31: 055–080. doi:10.1385/MB:31:1:055.
456. The R project for Statistical Computing [online]. Accessed July 2008. Available from World Wide Web: http://www.r-project.org/ (n.d.).
457. SYBYL 7.3, Tripos International, 1699 South Hanley Rd., St. Louis, Missouri, 63144, USA (n.d.).
458. OMEGA, Version 2.2.1; Openeye Scientific Software, Inc.: Santa Fe, NM. (n.d.).
459. McGann MR, Almond HR, Nicholls A, Grant JA, Brown FK (2003) Gaussian Docking Functions. Biopolymers 68: 76–90. doi:10.1002/bip.10207.
460. Mooij WTM, Verdonk ML (2005) General and Targeted Statistical Potentials for Protein-Ligand Interactions. Proteins 61: 272–287. doi:10.1002/prot.20588.
461. Stahl M, Rarey M (2001) Detailed Analysis of Scoring Functions for Virtual Screening. J Med Chem 44: 1035–1042.
462. Humphrey W, Dalke A, Schulten K (1996) VMD: visual molecular dynamics. J Mol Graph 14: 27–81, 33–38.
463. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, et al. (2004) UCSF Chimera - a visualization system for exploratory research and analysis. J Comput Chem 25: 1605–1612. doi:10.1002/jcc.20084.
464. Kroemer RT, Vulpetti A, McDonald JJ, Rohrer DC, Trosset J-Y, et al. (2004) Assessment of docking poses: interactions-based accuracy classification (IBAC) versus crystal structure deviations. J Chem Inf Comput Sci 44: 871–881. doi:10.1021/ci049970m.
465. Yusuf D, Davis AM, Kleywegt GJ, Schmitt S (2008) An alternative method for the evaluation of docking performance: RSR vs RMSD. J Chem Inf Model 48: 1411–1422. doi:10.1021/ci800084x.
466. Lengauer T, Lemmen C, Rarey M, Zimmermann M (2004) Novel technologies for virtual screening. Drug Discov Today 9: 27–34. doi:10.1016/S1359-6446(04)02939-3.
467. Ehrlich P (1909) Über den jetzigen Stand der Chemotherapie. Ber Dtsch Chem Ges 42: 17–47.
468. Evers A, Hessler G, Matter H, Klabunde T (2005) Virtual Screening of Biogenic Amine-Binding G-protein Coupled Receptors: Comparative Evaluation of Protein- and Ligand-Based Virtual Screening Protocols. J Med Chem 48: 5448–5465. doi:10.1021/jm050090o.
469. Good AC, Cho S-J, Mason JS (2004) Descriptors you can count on? Normalized and filtered pharmacophore descriptors for virtual screening. J Comput Aided Mol Des 18: 523–527. doi:10.1007/s10822-004-4065-3.
470. Steindl TM, Crump CE, Hayden FG, Langer T (2005) Pharmacophore modeling, docking, and principal component analysis based clustering: combined computer-

assisted approaches to identify new inhibitors of the human rhinovirus coat protein. J Med Chem 48: 6250–6260. doi:10.1021/jm050343d.
471. Schneidman-Duhovny D, Dror O, Inbar Y, Nussinov R, Wolfson HJ (2008) Deterministic pharmacophore detection via multiple flexible alignment of drug-like molecules. J Comput Biol 15: 737–754. doi:10.1089/cmb.2007.0130.
472. Dror O, Schneidman-Duhovny D, Inbar Y, Nussinov R, Wolfson HJ (2009) Novel approach for efficient pharmacophore-based virtual screening: method and applications. J Chem Inf Model 49: 2333–2343. doi:10.1021/ci900263d.
473. Koes DR, Camacho CJ (2011) Pharmer: efficient and exact pharmacophore search. J Chem Inf Model 51: 1307–1314. doi:10.1021/ci200097m.
474. Schneidman-Duhovny D, Dror O, Inbar Y, Nussinov R, Wolfson HJ (2008) PharmaGist: a webserver for ligand-based pharmacophore detection. Nucleic Acids Res 36: W223–8. doi:10.1093/nar/gkn187.
475. Langer T, Hoffmann R (2006) Pharmacophores and Pharmacophore Searches. Mannhold R, Kubinyi H, Folkers G, editors Weinheim, Germany: WILEY-VCH. p.
476. Smellie A, Teig SL, Towbin P (1995) Poling: Promoting conformational variation. J Comput Chem 16: 171–187. doi:10.1002/jcc.540160205.
477. Morphy R, Kay C, Rankovic Z (2004) From magic bullets to designed multiple ligands. Drug Discov Today 9: 641–651.
478. Hopkins AL, Mason JS, Overington JP (n.d.) Can we rationally design promiscuous drugs? Current Opinion in Structural Biology. doi:10.1016/j.sbi.2006.01.013.
479. Wermuth CG (2004) Multitargeted drugs: the end of the "one-target-one-disease" philosophy? Drug Discov Today 9: 826–827.
480. Metz JT, Hajduk PJ (2010) Rational approaches to targeted polypharmacology: creating and navigating protein-ligand interaction networks. Curr Opin Chem Biol. doi:10.1016/j.cbpa.2010.06.166.
481. Joosten RP, Te Beek T a H, Krieger E, Hekkelman ML, Hooft RWW, et al. (2010) A series of PDB related databases for everyday needs. Nucleic Acids Res 39: 411–419. doi:10.1093/nar/gkq1105.
482. Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, et al. (2005) Scalable molecular dynamics with NAMD. J Comput Chem 26: 1781–1802.
483. MacKerell ADJ, Banavali N, Foloppe N (2000) Development and current status of the CHARMM force field for nucleic acids. Biopolymers 56: 257–265.
484. Zhang L, Hermans J (1996) Hydrophilicity of cavities in proteins. Proteins 24: 433–438.
485. Jorgensen WL (1982) Revised TIPS for Simulations of liquid water and aqueous solutions. J Chem Phys 77: 4156–4163.
486. Feller SE, Zhang Y, Pastor RW, Brooks BR (1995) Constant pressure molecular dynamics simulation: the Langevin piston method. J Chem Phys 103: 4613–4621.
487. Darden T, York D, Pedersen L (1993) Particle mesh Ewald. An N. log (N) method for Ewald sums in large systems. J Chem Phys 98: 10089–10092.
488. Ryckaert JP, Ciccotti G, Berendsen HJC (1977) Numerical integration of the Cartesian equations of motion of a system with constraints: Molecular Dynamics of n-alkanes. J Comput Phys 23: 327–341.
489. QuacPac, Version 1.3.1; Openeye Scientific Software, Inc.: Santa Fe, NM. (n.d.).
490. I ACA-bcc M, Jakalian A, Bush BL, Jack DB, Bayly CI (2000) Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: I. Method. J Comput Chem 21: 132–146.
491. Hurst T (1994) Flexible 3D searching: The directed tweak technique. J Chem Inf Comput Sci 34: 190–196. doi:10.1021/ci00017a025.
492. Yuan Z, Zhao J, Wang Z-X (2003) Flexibility analysis of enzyme active sites by crystallographic temperature factors. Prot Eng Des Sel 16: 109–114. doi:10.1093/proeng/gzg014.
493. Schames JR, Henchman RH, Siegel JS, Sotriffer CA, Ni H, et al. (2004) Discovery of a Novel Binding Trench in HIV Integrase. J Med Chem: 1879–1881.
494. Lin J-H, Perryman AL, Schames JR, McCammon JA (2002) Computational drug design accommodating receptor flexibility: the relaxed complex scheme. J Am Chem Soc 124: 5632–5633.

495. Varmus H, Coffin JM, Hughes SH (1997) Retroviruses. Coffin JM, Hughes SH, Varmus HE, editors New York: Cold Spring Harbor Laboratory Press. p.
496. McGaughey GB, Sheridan RP, Bayly CI, Culberson JC, Kreatsoulas C, et al. (n.d.) Comparison of topological, shape, and docking methods in virtual screening. J Chem Inf Model 47: 1504–1519. doi:10.1021/ci700052x.
497. Hajduk PJ, Bures M, Praestgaard J, Fesik SW (2000) Privileged molecules for protein binding identified from NMR-based screening. J Med Chem 43: 3443–3447.
498. Bohacek RS, McMartin C, Guida WC (1996) The art and practice of structure-based drug design: a molecular modeling perspective. Med Res Rev 16: 3–50. doi:10.1002/(SICI)1098-1128(199601)16:1<3::AID-MED1>3.0.CO;2-6.
499. ChemACX Ultra, version 11, CambridgeSoft Corporation, Cambridge, MA, EUA (2005).
500. National Cancer Institute. NCI Open Database. http://cactus.nci.nih.gov (accessed July 2009) (n.d.).
501. Teague S, Davis A, Leeson P, Oprea T (1999) The Design of Leadlike Combinatorial Libraries. Angew Chem Int Ed Engl 38: 3743–3748.
502. Lipinski CA (2000) Drug-like properties and the causes of poor solubility and poor permeability. J Pharmacol Toxicol Methods 44: 235–249.
503. Oprea TI (2000) Property distribution of drug-related chemical databases. J Comput Aided Mol Des 14: 251–264.
504. Steinbeck C, Han Y, Kuhn S, Horlacher O, Luttmann E, et al. (2003) The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics. J Chem Inf Comput Sci 43: 493–500. doi:10.1021/ci025584y.
505. CODESSA Semichem Inc., 12456 W 62nd Terrace, Suite D, Shawnee, Kansas 66216, USA (n.d.).
506. Dragon TALETE srl, Via V. Pisani, Milan 13–20124, Italy (n.d.).
507. Molconn, Hall Associates Consulting, 2 Davis Street, Quincy, Massachusetts 02170, USA (n.d.).
508. Molecular Operating Environment (MOE), Chemical Computing Group, Inc., 1010 Sherbrooke St. W, Suite 910, Montreal, Quebec H3A 2R7, Canada (n.d.).
509. Pipeline Pilot, Accelrys, Inc., 10188 Telesis Court, Suite 100, San Diego, CA 92121, USA (n.d.).
510. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev 23: 3–25. doi:10.1016/S0169-409X(96)00423-1.
511. Lipinski CA (2001) Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. Adv Drug Deliv Rev 46: 3–26.
512. Ghose AK, Viswanadhan VN, Wendoloski JJ (1999) A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. J Comb Chem 1: 55–68.
513. FILTER, Version 2.0.2; Openeye Scientific Software, Inc.: Santa Fe, NM. (n.d.).
514. Carhart RE, Smith DH, Venkataraghavan R (1985) Atom pairs as molecular features in structure-activity studies: definition and applications. J Chem Inf Comput Sci 25: 64–73. doi:10.1021/ci00046a002.
515. Willett P, Winterman V, Bawden D (1986) Implementation of nearest-neighbour searching in an online chemical structure search system. J Chem Inf Comput Sci 26: 36–41.
516. Brown R, Martin YC (1996) Use of Structure-Activity Data To Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. J Chem Inf Comput Sci 36: 572–584.
517. Brown R (1997) The information content of 2D and 3D structural descriptors relevant to ligand-receptor binding. J Chem Inf Comput Sci 2338: 1–9.
518. Martin Y, Kofron J (2002) Do structurally similar molecules have similar biological activity? J Med Chem 45: 4350–4358.

519. Sheridan R, Feuston B, Maiorov V (2004) Similarity to Molecules in the Training Set Is a Good Discriminator for Prediction Accuracy in QSAR. J Chem Inf Comput Sci 44: 1912–1928.
520. Shanmugasundaram V, Maggiora G (2005) Hit-directed nearest-neighbor searching. J Med Chem 48: 240–248.
521. He L, Jurs PC (2005) Assessing the reliability of a QSAR model' s predictions. J Mol Graph Model 23: 503–523. doi:10.1016/j.jmgm.2005.03.003.
522. Zhang Q, Muegge I (2006) Scaffold hopping through virtual screening using 2D and 3D similarity descriptors: ranking, voting, and consensus scoring. J Med Chem 49: 1536–1548.
523. Matter H (1997) Selecting optimally diverse compounds from structure databases: a validation study of two-dimensional and three-dimensional molecular descriptors. Journal of medicinal chemistry 2623: 1219–1229.
524. Schuffenhauer A, Gillet V (2000) Similarity searching in files of three-dimensional chemical structures: analysis of the BIOSTER database using two-dimensional fingerprints and molecular field descriptors. J Chem Inf Comput Sci 40: 295–307.
525. Makara GM (2001) Measuring Molecular Similarity and Diversity: Total Pharmacophore Diversity. J Med Chem 44: 3563–3571. doi:10.1021/jm010036h.
526. Sheridan R, Kearsley S (2002) Why do we need so many chemical similarity search methods? Drug Discov Today 7: 903–911.
527. Cruciani G, Pastor M, Mannhold R (2002) Suitability of Molecular Descriptors for Database Mining. A Comparative Analysis. J Med Chem 45: 2685–2694.
528. Cramer RD, Jilek RJ, Guessregen S, Clark SJ, Wendt B, et al. (2004) "Lead Hopping". Validation of Topomer Similarity as a Superior Predictor of Similar Biological Activities. J Med Chem 47: 6777–6791.
529. ChemAxon (n.d.) Chemical Hashed Fingerprints. Available:http://www.chemaxon.com/jchem/doc/user/fingerprint.html. Accessed 3 May 2011.
530. Baumann K (2002) An alignment-independent versatile structure descriptor for QSAR and QSPR based on the distribution of molecular features. J Chem Inf Comput Sci 42: 26–35.
531. von Korff M, Steger M (2004) GPCR-tailored pharmacophore pattern recognition of small molecular ligands. J Chem Inf Comput Sci 44: 1137–1147. doi:10.1021/ci0303013.
532. Flower D (1998) On the properties of bit string-based measures of chemical similarity. J Chem Inf Comput Sci 2338: 379–386.
533. Godden J, Xue L, Bajorath J (2000) Combinatorial preferences affect molecular similarity/diversity calculations using binary fingerprints and Tanimoto coefficients. J Chem Inf Comput Sci 40: 163–166.
534. Holliday JD, Hu C-Y, Willett P (2002) Grouping of coefficients for the calculation of inter-molecular similarity and dissimilarity using 2D fragment bit-strings. Comb Chem High Throughput Screen 5: 155–166.
535. Salim N, Holliday J, Willett P (2003) Combination of fingerprint-based similarity coefficients using data fusion. J Chem Inf Comput Sci 43: 435–442. doi:10.1021/ci025596j.
536. Kirchmair J, Distinto S, Markt P, Schuster D, Spitzer GM, et al. (2009) How To Optimize Shape-Based Virtual Screening: Choosing the Right Query and Including Chemical Information. J Chem Inf Model 49: 678–692. doi:10.1021/ci8004226.
537. Hert J, Willett P, Wilton D, Acklin P (2004) Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures. Org Biomol Chem 2: 3256–3266.
538. Geppert H, Horváth T, Gärtner T, Wrobel S, Bajorath J (2008) Support-vector-machine-based ranking significantly improves the effectiveness of similarity searching using 2D fingerprints and multiple reference compounds. J Chem Inf Model 48: 742–746. doi:10.1021/ci700461s.
539. Shemetulskis NE, Weininger D, Blankley CJ, Yang JJ, Humblet C (1996) Stigmata: an algorithm to determine structural commonalities in diverse datasets. J Chem Inf Comput Sci 36: 862–871.

540. Whittle M, Gillet VJ, Willett P, Alex A, Loesel J (2004) Enhancing the effectiveness of virtual screening by fusing nearest neighbor lists: a comparison of similarity coefficients. J Chem Inf Comput Sci 44: 1840–1848. doi:10.1021/ci049867x.
541. Hert J, Willett P, Wilton DJ, Acklin P, Azzaoui K, et al. (2006) New methods for ligand-based virtual screening: use of data fusion and machine learning to enhance the effectiveness of similarity searching. J Chem Inf Model 46: 462–470. doi:10.1021/ci050348j.
542. JChem, Version 5.1.4, ChemAxon (www.chemaxon.com) (n.d.).
543. Huang N, Shoichet BK, Irwin JJ (2006) Benchmarking Sets for Molecular Docking. J Med Chem 49: 6789–6801. doi:10.1021/jm0608356.
544. LibraryMCS, Version 0.7, ChemAxon (www.chemaxon.com) (n.d.).
545. Berthold MR, Cebron N, Dill F, Fatta GD, Gabriel TR, et al. (2008) KNIME: The Konstanz Information Miner. Computer 11: 319–326. doi:10.1145/1656274.1656280.
546. Berthold MR, Cebron N, Dill F, Gabriel TR, Kötter T, et al. (2009) KNIME - the Konstanz information miner: version 2.0 and beyond. ACM SIGKDD Explorations Newsletter 11: 26–31. doi:10.1145/1656274.1656280.
547. Kuhn T, Willighagen EL, Zielesny A, Steinbeck C (2010) CDK-Taverna: an open workflow environment for cheminformatics. BMC Bioinformatics 11: 159. doi:10.1186/1471-2105-11-159.
548. Larsson J, Gottfries J, Muresan S, Backlund A (2007) ChemGPS-NP: tuned for navigation in biologically relevant chemical space. J Nat Prod 70: 789–794. doi:10.1021/np070002y.
549. Rosén J, Lövgren A, Kogej T, Muresan S, Gottfries J, et al. (2009) ChemGPS-NP(Web): chemical space navigation online. J Comput Aided Mol Des 23: 253–259. doi:10.1007/s10822-008-9255-y.
550. Ertl P, Rohde B, Selzer P (2000) Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. J Med Chem 43: 3714–3717.
551. Wang R, Fu Y (1997) A New Atom-Additive Method for Calculating Partition Coefficients. J Chem Inf Comput Sci 2338: 615–621.
552. Mills JE, Dean PM (1996) Three-Dimensional Hydrogen-Bond Geometry and Probability Information from a Crystal Survey. J Comput Aided Mol Des 10: 607–622.
553. Veber DF, Johnson SR, Cheng H-Y, Smith BR, Ward KW, et al. (2002) Molecular Properties That Influence the Oral Bioavailability of Drug Candidates. J Med Chem 45: 2615–2623.
554. Martin YC (2005) A bioavailability score. J Med Chem 48: 3164–3170. doi:10.1021/jm0492002.
555. Egan WJ, Merz KM, Baldwin JJ (2000) Prediction of Drug Absorption Using Multivariate Statistics. J Med Chem 43: 3867–3877.
556. McGovern SL, Helfand BT, Feng B, Shoichet BK (2003) A Specific Mechanism of Nonspecific Inhibition. J Med Chem 46: 4265–4272. doi:10.1021/jm030266r.
557. Seidler J, McGovern SL, Doman TN, Shoichet BK (2003) Identification and Prediction of Promiscuous Aggregating Inhibitors Among Known Drugs. J Med Chem 46: 4477–4486. doi:10.1021/jm030191r.
558. Good AC, Cheney DL (2003) Analysis and optimization of structure-based virtual screening protocols (1): exploration of ligand conformational sampling techniques. J Mol Graph Model 22: 23–30. doi:10.1016/S1093-3263(03)00123-2.
559. Sperandio O, Andrieu O, Miteva MA, Vo M-quang, Souaille M, et al. (2007) MED-SuMoLig: A New Ligand-Based Screening Tool for Efficient Scaffold Hopping. J Chem Inf Model 47: 1097–1110.
560. Bashford D, Gerwert K (1992) Electrostatic calculations of the pKa values of ionizable groups in bacteriorhodopsin. J Mol Biol 224: 473–486.
561. Delano WL (2002) The PyMOL Molecular graphics system.
562. Wojtczak A, Cody V, Luft JR, Pangborn W (1996) Structures of Human Transthyretin Complexed with Thyroxine at 2.0 A Resolution and 3',5'-Dinitro-N-

Acetyl-L-Thyronine at 2.2 A Resolution. Acta Crystallogr D Biol Crystallogr 52: 758–765. doi:10.1107/S0907444996003046.
563. Marvin, Version 5.1.4, ChemAxon (www.chemaxon.com) (n.d.).
564. Nicholls A (2008) What Do We Know and When Do We Know It? J Comput Aided Mol Des 22: 239–255. doi:10.1007/s10822-008-9170-2.
565. Bergström C a S, Wassvik CM, Johansson K, Hubatsch I (2007) Poorly soluble marketed drugs display solvation limited solubility. J Med Chem 50: 5858–5862. doi:10.1021/jm0706416.
566. Shoichet BK (2004) Virtual screening of chemical libraries. Nature 432: 862–865. doi:10.1038/nature03197.
567. Tirado-Rives J, Jorgensen WL (2006) Contribution of conformer focusing to the uncertainty in predicting free energies for protein-ligand binding. J Med Chem 49: 5880–5884. doi:10.1021/jm060763i.
568. Hawkins PCD, Skillman a G, Nicholls A (2007) Comparison of shape-matching and docking as virtual screening tools. J Med Chem 50: 74–82. doi:10.1021/jm0603365.
569. Pérez-Nueno VI, Ritchie DW, Rabal O, Pascual R, Borrell JI, et al. (2008) Comparison of Ligand-Based and Receptor-Based Virtual Screening of HIV Entry Inhibitors for the CXCR4 and CCR5 Receptors Using 3D Ligand Shape Matching and Ligand-Receptor Docking. J Chem Inf Model 48: 509–533.
570. Blasi D, Pinto M, Nieto J, Arsequell G, Valencia G, et al. (2011) Drug discovery targeted at transthyretin cardiac amyloidosis: rational design, synthesis, and biological activity of new transthyretin amyloid inhibitors. Amyloid 18 Suppl 1: 55–57. doi:10.3109/13506129.2011.574354019.
571. Simões CJV, Mukherjee T, Brito RMM, Jackson RM (2010) Toward the Discovery of Functional Transthyretin Amyloid Inhibitors: Application of Virtual Screening Methods. J Chem Inf Model 50: 1806–1820. doi:10.1021/ci100250z.
572. Low CMR, Buck IM, Cooke T, Cushnir JR, Kalindjian SB, et al. (2005) Scaffold hopping with molecular field points: identification of a cholecystokinin-2 (CCK2) receptor pharmacophore and its use in the design of a prototypical series of pyrrole- and imidazole-based CCK2 antagonists. J Med Chem 48: 6790–6802. doi:10.1021/jm049069y.
573. Schuster D, Maurer EM, Laggner C, Nashev LG, Wilckens T, et al. (2006) The discovery of new 11beta-hydroxysteroid dehydrogenase type 1 inhibitors by common feature pharmacophore modeling and virtual screening. J Med Chem 49: 3454–3466. doi:10.1021/jm0600794.
574. Doddareddy MR, Choo H, Cho YS, Rhim H, Koh HY, et al. (2007) 3D Pharmacophore Based Virtual Screening of T-Type Calcium Channel Blockers. Bioorg Med Chem 15: 1091–1105. doi:10.1016/j.bmc.2006.10.013.
575. Chen X, Reynolds CH (2002) Performance of similarity measures in 2D fragment-based similarity searching: comparison of structural descriptors and similarity coefficients. J Chem Inf Comput Sci 42: 1407–1414.
576. Eckert H, Bajorath J (2006) Design and Evaluation of A Novel Class-Directed 2D Fingerprint to Search for Structurally Diverse Active Compounds. J Chem Inf Model 46: 2515–2526. doi:10.1021/ci600303b.
577. Cheeseright TJ, Mackey MD, Melville JL, Vinter JG (2008) FieldScreen: Virtual Screening Using Molecular Fields. Application to the DUD Data Set. J Chem Inf Model 48: 2108–2117. doi:10.1021/ci800110p.
578. Jain AN, Nicholls A (2008) Recommendations for evaluation of computational methods. J Comput Aided Mol Des: 133–139. doi:10.1007/s10822-008-9196-5.
579. Hawkins PCD, Warren ÆGL, Skillman AG, Nicholls ÆA, Software K, et al. (2008) How to do an evaluation: pitfalls and traps: 179–190. doi:10.1007/s10822-007-9166-3.
580. Hawkins P (2007) On how not to do an evaluation. Presentation at CUP8. February 26th-28th 2007.
581. Good AC, Hermsmeier MA, Hindle A (2005) Measuring CAMD technique performance: A virtual screening case study in the design of validation experiments. J Comput Aided Mol Des: 529–536. doi:10.1007/s10822-004-4067-1.

582. Good AC, Hermsmeier M a (2007) Measuring CAMD technique performance. 2. How “druglike” are drugs? Implications of Random test set selection exemplified using druglikeness classification models. J Chem Inf Model 47: 110–114. doi:10.1021/ci6003493.
583. Good AC, Oprea TI (2008) Optimization of CAMD techniques 3. Virtual screening enrichment studies: a help or hindrance in tool selection? J Comput Aided Mol Des 22: 169–178. doi:10.1007/s10822-007-9167-2.
584. Clark RD, Webster-Clark DJ (2008) Managing bias in ROC curves. J Comput Aided Mol Des 22: 141–146. doi:10.1007/s10822-008-9181-z.
585. Rohrer SG, Baumann K (2009) Maximum unbiased validation (MUV) data sets for virtual screening based on PubChem bioactivity data. J Chem Inf Model 49: 169–184. doi:10.1021/ci8002649.
586. Geppert H, Vogt M, Bajorath J (2010) Current trends in ligand-based virtual screening: molecular representations, data mining methods, new application areas, and performance evaluation. J Chem Inf Model 50: 205–216. doi:10.1021/ci900419k.
587. Triballeau N, Acher F, Brabet I, Pin J-P, Bertrand H-O (2005) Virtual Screening Workflow Development Guided by the “Receiver Operating Characteristic” Curve Approach. Application to High-Throughput Docking on Metabotropic Glutamate Receptor Subtype 4. J Med Chem. doi:10.1021/jm049092j.
588. Kellenberger E, Foata N, Rognan D (2008) Ranking targets in structure-based virtual screening of three-dimensional protein libraries: methods and problems. J Chem Inf Model 48: 1014–1025. doi:10.1021/ci800023x.
589. Gund P, Andose JD, Rhodes JB, Smith GM (1980) Three-dimensional molecular modeling and drug design. Science 208: 1425–1431.
590. Doucet JP, Panaye A (1998) 3D Structural Information: From Property Prediction to Substructure Recognition with Neural Networks. SAR and QSAR Environm Res 8: 249–272. doi:10.1080/10629369808039143.
591. Pepperrell CA, Taylor R, Willett P (1990) Implementation and use of an atom-mapping procedure for similarity searching in databases of 3-D chemical structures. Tetrahedron Computer Methodology 3: 575–593. doi:10.1016/0898-5529(90)90160-A.
592. Nikolova N, Jaworska J (2003) Approaches to Measure Chemical Similarity - a Review. QSAR Comb Sci 22: 1006–1026. doi:10.1002/qsar.200330831.
593. Bajorath J (2002) Virtual screening in drug discovery: Methods, expectations and reality. Curr Drug Discov: 24–27.
594. ROCS, Version 2.3.1; Openeye Scientific Software, Inc.: Santa Fe, NM. (n.d.).
595. Nicholls A, MacCuish NE, MacCuish JD (2004) Variable selection and model validation of 2D and 3D molecular descriptors. J Comput Aided Mol Des 18: 451–474. doi:10.1007/s10822-004-5202-8.
596. Weiner PK, Langridge R, Blaney JM, Schaefer R, Kollman PA (1982) Electrostatic Potential Molecular Surfaces. Proc Natl Acad Sci USA 79: 3754–3758.
597. VIDA, Version 4.0.3; Openeye Scientific Software, Inc., Santa Fe, NM. (n.d.).
598. Kinnings SL, Jackson RM (2009) LigMatch: A Multiple Structure-Based Ligand Matching Method for 3D Virtual Screening. J Chem Inf Model 49: 2056–2066.
599. Kollman P (1993) Free energy calculations: Applications to chemical and biochemical phenomena. Chem Rev 93: 2395–2417. doi:10.1021/cr00023a004.
600. Ajay, Murcko MA (1995) Computational Methods to Predict Binding Free Energy in Ligand-Receptor Complexes. J Med Chem 38: 4953–4967.
601. Jorgensen WL (1989) Free energy calculations: a breakthrough for modeling organic chemistry in solution. Acc Chem Res 22: 184–189. doi:10.1021/ar00161a004.
602. Massova I (2000) Combined molecular mechanical and continuum solvent approach (MM-PBSA/GBSA) to predict ligand binding. Perspect Drug Discov Des: 113–135.
603. Neudert G, Klebe G (2011) DSX: a knowledge-based scoring function for the assessment of receptor-ligand complexes. J Chem Inf Model In Press. doi:10.1021/ci200274q.

604. Velec HFG, Gohlke H, Klebe G (2005) Knowledge-Based Scoring Function Derived from Small Molecule Crystal Data with Superior Recognition Rate of Near-Native Ligand Poses and Better Affinity Prediction. J Mol Biol 48: 6296–6303.
605. Paleologou M, Li S, Purdy WC (1990) Liquid chromatographic retention behaviour and separation of chlorophenols on a beta-cyclodextrin bonded-phase column. Part III. Diaromatic chlorophenols. Can J Chem 68: 1208–1214.
606. Muley L, Baum B, Smolinski M, Freindorf M, Heine A, et al. (2010) Enhancement of hydrophobic interactions and hydrogen bond strength by cooperativity: synthesis, modeling, and molecular dynamics simulations of a congeneric series of thrombin inhibitors. J Med Chem 53: 2126–2135. doi:10.1021/jm9016416.
607. Baum B, Muley L, Smolinski M, Heine A, Hangauer D, et al. (2010) Non-additivity of functional group contributions in protein-ligand binding: a comprehensive study by crystallography and isothermal titration calorimetry. J Mol Biol 397: 1042–1054. doi:10.1016/j.jmb.2010.02.007.
608. Williams DH, Stephens E, O'Brien DP, Zhou M (2004) Understanding noncovalent interactions: ligand binding energy and catalytic efficiency from ligand-induced reductions in motion within receptors and enzymes. Angew Chem Int Ed 43: 6596–6616. doi:10.1002/anie.200300644.
609. Deng W, Breneman C, Embrechts MJ (2004) Predicting protein-ligand binding affinities using novel geometrical descriptors and machine-learning methods. J Chem Inf Comput Sci 44: 699–703. doi:10.1021/ci034246+.
610. Ballester PJ, Mitchell JBO (2010) A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking. Bioinformatics 26: 1169–1175. doi:10.1093/bioinformatics/btq112.
611. Zhou Z, Felts AK, Friesner R a, Levy RM (2007) Comparative performance of several flexible docking programs and scoring functions: enrichment studies for a diverse set of pharmaceutically relevant targets. J Chem Inf Model 47: 1599–1608. doi:10.1021/ci7000346.
612. Xing L, Hodgkin E, Liu Q, Sedlock D (2004) Evaluation and application of multiple scoring functions for a virtual screening experiment. J Comput Aided Mol Des 18: 333–344.
613. Kontoyianni M, McClellan LM, Sokol GS (2004) Evaluation of docking performance: comparative data on docking algorithms. J Med Chem 47: 558–565. doi:10.1021/jm0302997.
614. Nissink JWM, Murray C, Hartshorn M, Verdonk ML, Cole JC, et al. (2002) A new test set for validating predictions of protein-ligand interaction. Proteins 49: 457–471. doi:10.1002/prot.10232.
615. Hartshorn MJ, Verdonk ML, Chessari G, Brewerton SC, Mooij WTM, et al. (2007) Diverse, high-quality test set for the validation of protein-ligand docking performance. J Med Chem 50: 726–741. doi:10.1021/jm061277y.
616. Pham T (2006) Parameter estimation for scoring protein-ligand interactions using negative training data. J Med Chem 49: 5856–5868.
617. Kuntz ID, Chen K, Sharp K a, Kollman P a (1999) The maximal affinity of ligands. Proc Natl Acad Sci USA 96: 9997–10002.
618. Irwin JJ (2008) Community benchmarks for virtual screening. J Comput Aided Mol Des 22: 193–199. doi:10.1007/s10822-008-9189-4.
619. Francis G a, Fayard E, Picard F, Auwerx J (2003) Nuclear receptors and the control of metabolism. Ann Rev Physiol 65: 261–311. doi:10.1146/annurev.physiol.65.092101.142528.
620. Gronemeyer H, Gustafsson J-A, Laudet V (2004) Principles for modulation of the nuclear receptor superfamily. Nat Rev Drug Discov 3: 950–964. doi:10.1038/nrd1551.
621. Jacobsson M, Lidén P, Stjernschantz E, Boström H, Norinder U (2003) Improving structure-based virtual screening by multivariate analysis of scoring data. J Med Chem 46: 5781–5789. doi:10.1021/jm030896t.
622. Schapira M, Abagyan R, Totrov M (2003) Nuclear hormone receptor targeted virtual screening. J Med Chem 46: 3045–3059. doi:10.1021/jm0300173.

623. Yang J-moon, Shen T-wei (2005) A Pharmacophore-Based Evolutionary Approach for Screening Selective Estrogen Receptor Modulators. Proteins 59: 205–220. doi:10.1002/prot.20387.
624. Johnson L (2007) Protein kinases and their therapeutic exploitation. Biochem Soc Trans 35: 7–11. doi:10.1042/BST0350007.
625. Bogoyevitch M a, Fairlie DP (2007) A new paradigm for protein kinase inhibition: blocking phosphorylation without directly targeting ATP binding. Drug Discov Today 12: 622–633. doi:10.1016/j.drudis.2007.06.008.
626. Fedorov O, Sundström M, Marsden B, Knapp S (2007) Insights for the development of specific kinase inhibitors by targeted structural genomics. Drug Discov Today 12: 365–372. doi:10.1016/j.drudis.2007.03.006.
627. Diller DJ, Li R (2003) Kinases, homology models, and high throughput docking. J Med Chem 46: 4638–4647. doi:10.1021/jm020503a.
628. Halgren T a, Murphy RB, Friesner R a, Beard HS, Frye LL, et al. (2004) Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. J Med Chem 47: 1750–1759. doi:10.1021/jm030644s.
629. Claussen H, Gastreich M, Apelt V, Greene J, Hindle SA, et al. (2004) The FlexX database docking environment - Rational extraction of receptor based pharmacophores. Curr Drug Discov Technol 1: 49–60.
630. Kinnings SL, Jackson RM (2009) Binding site similarity analysis for the functional classification of the protein kinase family. J Chem Inf Model 49: 318–329. doi:10.1021/ci800289y.
631. Di Cera E (2009) Serine proteases. IUBMB Life 61: 510–515. doi:10.1002/iub.186.
632. Walker B, Lynas JF (2001) Strategies for the inhibition of serine proteases. Cell Mol Life Sci 58: 596–624.
633. McGovern SL, Shoichet BK (2003) Information decay in molecular docking screens against holo, apo, and modeled conformations of enzymes. J Med Chem 46: 2895–2907. doi:10.1021/jm0300330.
634. Thomson AJ, Gray HB (1998) Bio-inorganic chemistry. Curr Opin Chem Biol 2: 155–158. doi:10.1016/S1367-5931(98)80056-2.
635. Waldron KJ, Robinson NJ (2009) How do bacterial cells ensure that metalloproteins get the correct metal? Nat Rev Microbiol 7: 25–35. doi:10.1038/nrmicro2057.
636. Estabrook RW (2003) A passion for P450s (rememberances of the early history of research on cytochrome P450). Drug Metab Dispos 31: 1461–1473. doi:10.1124/dmd.31.12.1461.
637. Acharya KR, Sturrock ED, Riordan JF, Ehlers MRW (2003) Ace revisited: a new target for structure-based drug design. Nat Rev Drug Discov 2: 891–902. doi:10.1038/nrd1227.
638. David L, Amara P, Field MJ, Major F (2002) Parametrization of a force field for metals complexed to biomacromolecules: applications to Fe(II), Cu(II) and Pb(II). J Comput Aided Mol Des 16: 635–651.
639. Irwin JJ, Raushel FM, Shoichet BK (2005) Virtual screening against metalloenzymes for inhibitors and substrates. Biochemistry 44: 12316–12328. doi:10.1021/bi050801k.
640. Smith RD, Dunbar JB, Ung PM-U, Esposito EX, Yang C-Y, et al. (2011) CSAR Benchmark Exercise of 2010: Combined Evaluation Across All Submitted Scoring Functions. J Chem Inf Model Epub. doi:10.1021/ci200269q.
641. Takimoto C (1996) New Antifolates: Pharmacology and Clinical Applications. Oncologist 1: 68–81.
642. Martin TA, Jiang WG (2010) Anti-Cancer agents in medicinal chemistry (Formerly current medicinal chemistry - Anti-cancer agents). Anticancer Agents Med Chem 10: 1.
643. Kontoyianni M, Sokol GS, McClellan LM (2005) Evaluation of library ranking efficacy in virtual screening. J Comput Chem 26: 11–22. doi:10.1002/jcc.20141.
644. Schulz-Gasch T, Stahl M (2003) Binding site characteristics in structure-based virtual screening: evaluation of current docking tools. J Mol Model 9: 47–57. doi:10.1007/s00894-002-0112-y.

645. Li H, Li C, Gui C, Luo X, Chen K, et al. (2004) GAsDock: a new approach for rapid flexible docking based on an improved multi-population genetic algorithm. Bioorg Med Chem Lett 14: 4671–4676. doi:10.1016/j.bmcl.2004.06.091.
646. Mozziconacci J-C, Arnoult E, Bernard P, Do QT, Marot C, et al. (2005) Optimization and validation of a docking-scoring protocol; application to virtual screening for COX-2 inhibitors. J Med Chem 48: 1055–1068. doi:10.1021/jm049332v.
647. Dunbrack Jr R (2002) Rotamer libraries in the 21st century. Proteins: 431–440.
648. The Open Babel Package, version 2.2.3 http://openbabel.sourceforge.net/ (accessed Oct 2009) (n.d.).
649. Guha R, Howard MT, Hutchison GR, Murray-Rust P, Rzepa H, et al. (2006) The Blue Obelisk-interoperability in chemical informatics. J Chem Inf Model 46: 991–998. doi:10.1021/ci050400b.
650. Ben-Naim A (1992) Statistical Thermodynamics for Chemists and Biochemists. Statistical Thermodynamics for Chemists and Biochemists. New York: Plenum Press.
651. Sippl MJ (1990) Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. J Mol Biol 213: 859–883.
652. Sippl MJ (1993) Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. J Comput Aided Mol Des 7: 473–501.
653. Sippl MJ (1995) Knowledge-based potentials for proteins. Curr Opin Struct Biol 5: 229–235.
654. Muegge I, Martin YC (1999) A general and fast scoring function for protein-ligand interactions: a simplified potential approach. J Med Chem 42: 791–804. doi:10.1021/jm980536j.
655. Ben-Naim A (1997) Statistical potentials extracted from protein structures: Are these meaningful potentials? J Chem Phys 107: 3698–3706.
656. Koppensteiner WA, Sippl MJ (1998) Knowledge-based potentials — back to the roots. Biochemistry 63: 247–252.
657. Allen FH, Davies JE, Galloy JJ, Johnson O, Kennard O, et al. (1991) The development of versions 3 and 4 of the Cambridge Structural Database System. J Chem Inf Model 31: 187–204. doi:10.1021/ci00002a004.
658. Neudert G, Klebe G (2011) fconv: Format conversion, manipulation and feature computation of molecular data. Bioinformatics 27: 1021–1022. doi:10.1093/bioinformatics/btr055.
659. Volkamer A, Griewel A, Grombacher T, Rarey M (2010) Analyzing the topology of active sites: on the prediction of pockets and subpockets. J Chem Inf Model 50: 2041–2052. doi:10.1021/ci100241y.
660. Kinnings SL, Liu N, Tonge PJ, Jackson RM, Xie L, et al. (2011) A Machine Learning-Based Method To Improve Docking Scoring Functions and Its Application to Drug Repurposing. J Chem Inf Model Epub. doi:10.1021/ci100369f.
661. Marianayagam NJ, Fawzi NL, Head-Gordon T (2005) Protein folding by distributed computing and the denatured state ensemble. Proc Natl Acad Sci USA 102: 16684–16689. doi:10.1073/pnas.0506388102.
662. Ensign DL, Pande VS (2009) The Fip35 WW domain folds with structural and mechanistic heterogeneity in molecular dynamics simulations. Biophys J 96: L53–5. doi:10.1016/j.bpj.2009.01.024.
663. Das R, Qian B, Raman S, Vernon R, Thompson J, et al. (2007) Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home. Proteins 69 Suppl 8: 118–128. doi:10.1002/prot.21636.
664. Chang MW, Lindstrom W, Olson AJ, Belew RK (2007) Analysis of HIV wild-type and mutant structures via in silico docking against diverse ligand libraries. J Chem Inf Model 47: 1258–1262. doi:10.1021/ci700044s.
665. Kelder J, Grootenhuis PJ, Bayada DM, Delbressine LPC, Ploemen J-P (1999) Polar Molecular Surface as a Dominating Determinant for Oral Absorption and Brain Penetration of Drugs. Pharm Res 16: 1514–1519.

666. Gil-Redondo R, Estrada J, Morreale A, Herranz F, Sancho J, et al. (2009) VSDMIP: virtual screening data management on an integrated platform. J Comput Aided Mol Des 23: 171–184. doi:10.1007/s10822-008-9249-9.
667. Zhou T, Caflisch A (2009) Data management system for distributed virtual screening. J Chem Inf Model 49: 145–152. doi:10.1021/ci800295q.
668. Irwin JJ, Shoichet BK, Mysinger MM, Huang N, Colizzi F, et al. (2009) Automated docking screens: a feasibility study. J Med Chem 52: 5712–5720. doi:10.1021/jm9006966.

List of Publications

Max Pereira, Vítor S. Costa, Rui Camacho, Nuno A. Fonseca, Carlos J. V. Simões, Rui M. M. Brito. *Comparative study of classification algorithms using molecular descriptors in toxicological databases.* Advances in Bioinformatics and Computational Biology, 121-132 (2009).

Cândida G. Silva, Rui Palma, José R. Simões, Nuno Loureiro-Ferreira, Carlos J. V. Simões, Martin Swain, Werner Dubitzky, Rui M M. Brito. *P-found GRID: A Grid-enabled repository prototype for protein simulation data.* IBERGRID'2009 Conference Proceedings (2009).

J. Rui Rodrigues, Carlos J. V. Simões, Cândida G. Silva, Rui M. M. Brito. *Potentially amyloidogenic conformational intermediates populate the unfolding landscape of transthyretin: insights from molecular dynamics simulations*. Protein Science, 19 (2): 202-219 (2010).

Carlos J. V. Simões, Alejandro Rivero, Alfonso Tarancón, Fermín S. Sanz, Javier L. Romero, Pedro Abreu, Carlos Manuel, Rui Durão, João Rosa, João Pagaime, Pedro Veiga, Richard M. Jackson, Ana Noronha, Rosalia Vargas, Luís Magalhães, Gaspar Barreira, Rui M. M. Brito. *Running faster and further together: the AMILOIDE project, the onset of the Portuguese participation on the Ibercivis volunteer computing network*. IBERGRID'2010 Conference Proceedings, 404 (2010).

Carlos J. V. Simões, Trishna Mukherjee, Rui M. M. Brito, Richard M. Jackson. *Toward the Discovery of Functional Transthyretin Amyloid Inhibitors: Application of Virtual Screening Methods*. Journal of Chemical Information and Modelling, 50 (10): 1806-1820 (2010).

Carlos J. V. Simões, Alejandro Rivero, Rui Brito. *Searching for Anti-Amyloid Drugs with the Help of Citizens: the "AMILOIDE" Project on the IBERCIVIS Platform*. ERCIM News (Magazine Article), 25-26 (2010).

Rui Camacho, Max Pereira, Vítor S. Costa, Nuno A. Fonseca, Carlos Adriano, Carlos J. V. Simões, Rui M. M. Brito. *A Relational Learning Approach to Structure-Activity Relationships in Drug Design Toxicity Studies*. Journal of Integrative Bioinformatics, 8 (3): 182-201 (2011).

Catarina S. H. Jesus, Elsa S. Henriques, Carlos J. V. Simões, Cláudia V. S. Moniz, Pedro F. Cruz, Adrián Velázquez-Campoy, Rui M. M. Brito. *Binding Energetics of Four Inhibitors of Transthyretin Amyloid Formation: Guidelines for a Rational Drug Design*. To be submitted.