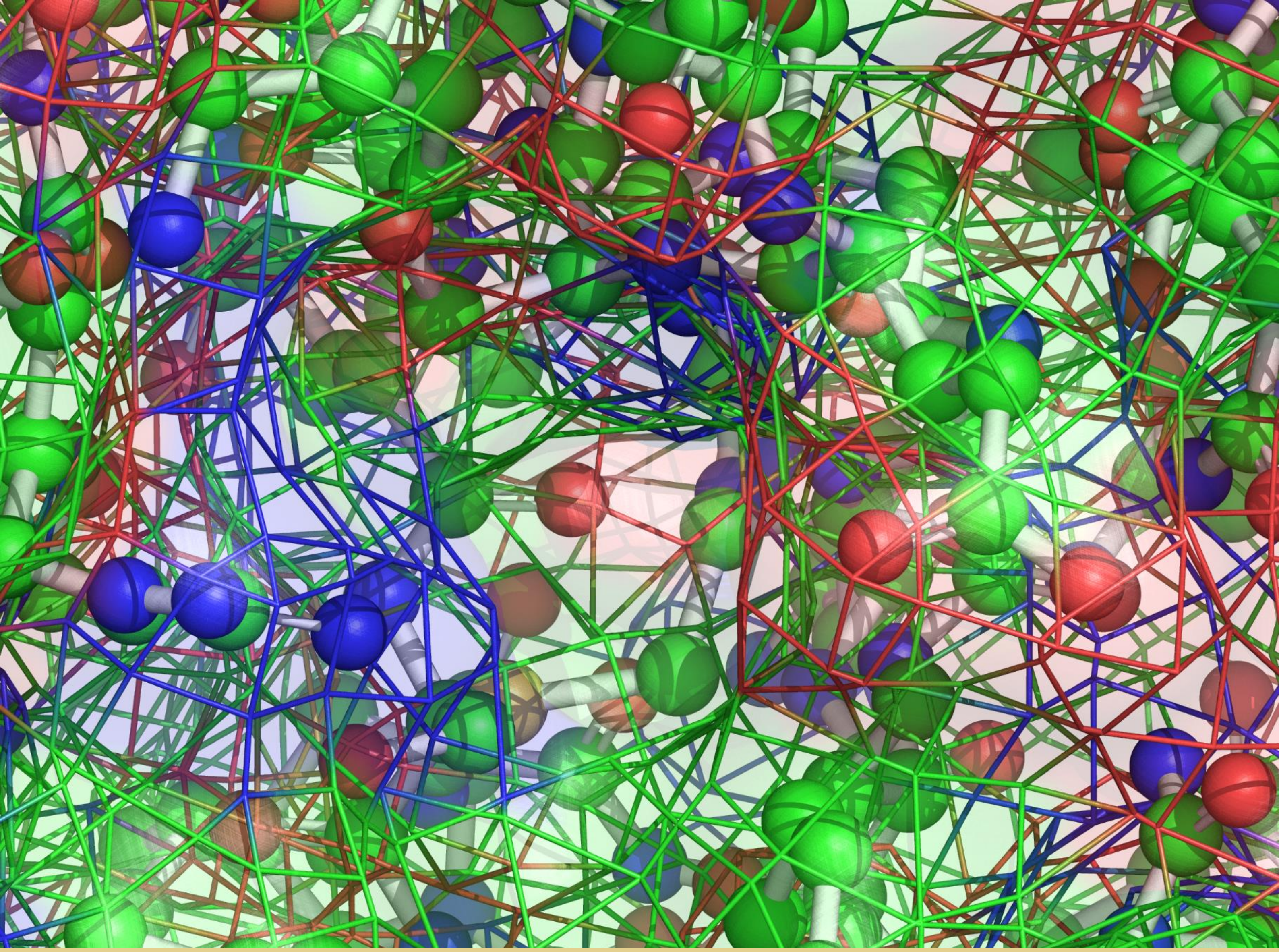


José Guilherme Coelho Peres de Almeida

Computational methods for the understanding of protein-based interactions

Dissertação de mestrado em Biologia Celular e Molecular orientada por Irina S. Moreira e Ana Luísa Carvalho e apresentada para o Departamento de Ciências da Vida da Universidade de Coimbra

Setembro/2017

· U C ·

Universidade de Coimbra

Index

Figure Index

Table Index

Annex Index

Acknowledgements

Os costumes ditam que tenho de agradecer a toda a gente que me ajudou a chegar até aqui. Prometo que, por questões meramente logísticas, não o farei da forma que devia - não há tempo nem papel que chegue.

Começo por agradecer à minha orientadora, Irina Moreira, pela constante ajuda e pela ética de trabalho focada não só nos resultados, mas também na aprendizagem.

À minha família, que foi o meu suporte e presença inabalável em todo e qualquer momento, através do encorajamento incansável e lugar seguro onde sinto que posso estar sem que nada me aconteça. Dizem que os amigos são a família que escolhemos; para mim, a família são os amigos que escolheríamos de qualquer maneira.

Aos meus amigos, que me sinto obrigado a rever da forma mais resumida e compreensiva:

- Ao Tó-Zé pelas incontáveis discussões e trocas de ideias e impressões, e pelos (contáveis, espero eu) copos e cafés - não é um termo que faça questão de usar, mas foste o meu melhor amigo;
- Aos "bioquímicos" (a todos, mas com mais cuidado a alguns: Ana Rita, Areal, Bernardo, Chicória, Valério) pelas saídas, pela AJA48 e por tudo o que me ajudou a esquecer que estava a tirar um curso;
- À RUC (nos tempos de mais e menos raios de sol) por ter sido a relação de amor-ódio que mais gostei;
- À JEST por me ter proporcionado tantas experiências e complicações novas;
- Ao resto da malta de Coimbra () pela cor que dão à cidade-a-preto-e-branco e pelo movimento que dão às paisagens estagnadas;
- A toda a malta da música (em especial ao Tiago Lopes por ter sido a presença constante) - Villa, Vilharigues Ensemble, DEL e por aí adiante. Foram o maior escape que eu tive e ensinam-me ainda hoje que nem só de pão vive o homem - a arte constrói-se com pessoas como vocês;
- Às velhas amizades de Viseu que raramente encontro por sermos todos pessoas extremamente ocupadas.

A todos vocês (e a quem me esqueci) - obrigado por tudo.

Abstract

Understanding protein-protein interfaces is at the foundation of studying molecular interactions in living organisms and in determining their relevance in high-order protein networks. However, the most detailed information on protein complexes – the three-dimensional structure – is often unavailable, mainly due to difficulties in their experimental determination. As such, computational methods rise as increasingly valid alternatives by using the information already retrieved from previous experiments as a driving key factor of protein structure study. This thesis work focuses on protein-protein interfaces and how computational tools can be used to aid in their understanding. As such, four main tasks were conducted. The first was the development of a computational pipeline for the prediction of important residues in protein-protein interfaces (Hot-spots (HS)) using an ensemble of machine-learning algorithms. The final model, SpotOn, had an Accuracy of 0.95 and a Sensitivity of 0.98 for an independent test set and is available online in http://milou.science.uu.nl/cgi/services/SPOTON/spoton/. The second task involved a global assessment of protein-protein interfaces and HS using a non-redundant database of protein complexes and an adapted version of the computational pipeline developed in the previous task. By doing so, structural features, the number of intermonomer neighbouring residues and neighbouring HS came out as essential in defining HS. The third task was the development of an algorithm capable of predicting interfacial residues from monomer structure information using deep-learning. The final model was not able to predict interfacial residues, probably due to the absence of relevant features. The fourth and final task aimed at understanding how dynamics affect protein-protein interfaces using normal mode analysis (NMA) and interhelical distance in a case study based on G-protein coupled receptor (GPCR)-partner binding. While GPCR fluctuation values from NMA were able to distinguish different G-proteins and different arrestins, they were not as useful in distinguishing G-proteins from arrestins. Interhelical distance did the opposite – G-proteins were easily distinguished from arrestins, but different G-proteins and different arrestins were undistinguishable. Results from the second and fourth tasks are available online in 45.32.153.74/spotondb and 45.32.153.74/gpcr, respectively.

Keywords: protein-protein interfaces, machine-learning, bioinformatics, structural biology

Resumo

Compreender as interfaces proteína-proteína é a base do estudo das interações molecular em organismos e na determinação da sua relevância em redes proteicas complexas. Contudo, a informação mais detalhada sobre complexos proteicos – estruturas tridimensionais – é muitas vezes inexistente, principalmente devido a dificuldades na sua determinação experimental. Como tal, métodos computacionais surgem como cada vez mais válidos e capazes de atingir o mesmo desempenho demonstrado por métodos experimentais ao utilizarem a informação disponível de experiências anteriores como um fator chave no estudo da estrutura proteica. Como tal, esta tese focou-se nas interfaces proteína-proteína e na maneira como ferramentas computacionais podem ser utilizadas para ajudar na sua compreensão. Para atingir isto, quatro tarefas foram realizadas. A primeira consistiu no desenvolvimento de um método computacional para prever resíduos importantes em interfaces proteína-proteína (*Hot-spots* (HS)) através de uma combinação de algoritmos de *machine-learning*. O modelo final, *SpotOn*, tem uma precisão de 0.95 e uma sensibilidade de 0.98 para um conjunto de dados independente e está disponível em http://milou.science.uu.nl/cgi/services/SPOTON/spoton/. A segunda tarefa envolveu uma aferição global de interfaces proteína-proteína e HS com uma base de dados não redundante de complexos proteicos e uma versão adaptada do método computacional desenvolvido na tarefa anterior. Características estruturais, o número de resíduos próximos intermonoméricos e de HS próximos foram determinantes na caracterização de HS. A terceira tarefa consistiu no desenvolvimento de um algoritmo capaz de prever resíduos interfaciais a partir de informação estrutural de monómeros com ferramentas de *deep-learning*. O modelo final foi incapaz de prever resíduos interfaciais, provavelmente devido à ausência de características relevantes para o problema. A quarta tarefa pretendia perceber como é que a dinâmica proteica afeta as interfaces proteína-proteína usando análise de modos normais (AMN) e distâncias inter-hélice num caso de estudo baseado na ligação de recetores acoplados a proteínas-G (RAPG) a parceiros intracelulares. Enquanto que valores de flutuação de RAPGs obtidos através da AMN eram capazes de distinguir diferentes proteínas-G e diferentes arrestinas, não eram úteis na distinção entre proteínas-G e arrestinas. Nas distâncias inter-hélice acontecia o oposto – as proteínas-G era facilmente distinguidas das arrestinas, mas diferentes proteínas-G e diferentes arrestinas eram indistinguíveis. Os resultados da segunda e terceira estão disponíveis online em 45.32.153.74/spotondb e 45.32.153.74/gpcr, respetivamente.

Palavras-chave: interfaces proteína-proteína, machine-learning, bioinformática, biologia estrutural

A. Introduction

1) Proteins structure fundamentals

Proteins are the fundamental units of cells and higher order organisms, representing high diversity in function, which can be closely associated with their structural and sequence composition [1]. A key aspect of protein function is their interaction with small molecules, peptides or other proteins. For example, substrate binding is an essential step for catalysis in enzymes, peptide binding is necessary to trigger signalling cascades in some receptors and protein-protein interactions are fundamental in the activation of some proteins [2]. Structural elucidation of proteins is the first approach to understand how they interact with different systems. However, experimental determination of protein structure, especially complexes, can be challenging, leading to the emergence of high-throughput but low-information methods [3].

Experimental protein structure determination is typically done through nuclear magnetic resonance (NMR), X-ray crystallography and cryo-electron microscopy (Cryo-EM). However, these techniques have distinct problems which make them much more complicated, expensive and time-consuming, than high-throughput methods. The main advantage of NMR is the ability to determine not only protein structure but also protein dynamics [4]. However, when it comes to large protein structure, spectral crowding becomes a problem due to insufficient sensitivity – while relevant motions might still be captured, structure determination becomes more challenging [5,6]. Nonetheless, some methods such as more refined data analysis [7] and solid-state NMR [8] have been developed to tackle this problem. X-ray crystallography is fairly well-developed concerning sensitivity – several structures have been resolved with subatomic resolution (< 1 Å) [9]. Its major drawback is how time-consuming and expensive it can become – the structural determination of the aspartate protease, which required 160.000 different conditions in order to achieve good, analyzable crystals [10]; the determination of the high resolution crystal structure of an engineered human β_2-adrenergic GPCR, which took 15 years [11]; and the 13 year-long structure determination of the membrane-integral diacylglycerol kinase [12], as noted by Leman *et. al.* in their 2015 review paper [13]. Cryo-EM is characterized by imaging radiation-sensitive entities – cells, viruses and macromolecules – under cryogenic conditions using a transmission electron microscope [14]. While Cryo-EM does not require protein crystal synthesis, it has relatively low resolution for membrane proteins (around 3 Å) when compared with X-ray crystallography. However, some recent cases show that Cryo-EM can provide highly informative structures of membrane proteins – the transient receptor potential channel 1 at 3.4 Å [15] and the chloride conducting (CLC) ion channel at 3.7 Å [16] are examples of the potential underlying this technique. EMDataBank [17] – publicly available at http://emdatabank.org/index.html – is a database of protein structures solved through Cryo-EM.

Despite recent advances in protein structure determination, Membrane Proteins (MPs), a rather large protein group with particular characteristics, stand out as one of the biggest challenges in protein structure determination [18,19]. This aspect is largely derived from the influence exerted by the membrane on the proteins and vice-versa [20], which might be a consequence of the membrane's cholesterol content [21,22] and/or thickness of the lipid bilayer's hydrophobic region [20,23-25]. This leads to a relatively low abundance of structural information on MPs when compared with soluble proteins in protein databases even though MPs represent a considerable proportion of the human proteome [26]. Furthermore, deriving information on how proteins interact is much harder when considering MPs and the amount of structural information available.

Luckily, computational methods – the alternative to experimental methods – are a particularly advanced field, concerning both variety and efficacy [27-41]. They rely on several different approaches, such as homology modelling and Molecular Dynamics (MD) [42] and *de novo* protein structure determination [43,44]. As such, considering the available computational tools, methods for the prediction of protein structure and interaction should be developed, and they should be utilized to better understand the complex structure-function relationships in MPs.

Concerning what has been described previously, this thesis will focus on:

i. Describing MPs, the biggest challenge in the determination of protein structure and protein-protein interactions;
ii. Reviewing the state-of-the-art on protein structure and interaction prediction;
iii. Describing a new method for the prediction of Hot-Spots (HS) in protein complexes;
iv. Understanding the patterns underlying protein-protein interactions through a big-data high-throughput analysis of protein interfaces;
v. Attempting to develop a new method to predict protein interfacial patches using Deep-Learning;
vi. Understanding how the interface might be affected by receptor dynamics using a GPCR-partner case study.

2) Membrane Proteins

MP are key representative of the challenges associated with protein structure determination and mechanistic understanding. Here, we also focused on GPCR-partner case study. MPs are proteins associated to lipid domains involved in communication, regulation and structural coherence. In fact, proteins that entirely or partially span the membrane (intrinsic/trans-membrane (TM) proteins), as well as proteins that are peripherally membrane-bound (peripheral MPs – PMPs), can carry out these functions. Considering that most computational methods for MPs are focused on TM proteins, the literature review presented in this thesis work will focus heavily on TM proteins and MPs and TM proteins will be used interchangeably. If the reader is interested on PMPs, specialized reviews covering this class of membrane proteins [45], their interaction with the membrane [46] and a review on experimental and computational methods for their study [47] can be consulted.

Understanding protein structure-function relationships is essential to understand common pathologies at a molecular level and to develop improved pharmacological approaches [48, 49, 50]. One of the most functionally relevant MP type are membrane receptors [51,52]. Membrane receptors, comprising GPCRs – which will be considered and reviewed further ahead –, olfactory receptors and nuclear receptors [53] play many roles in biochemical and signaling pathways, and in triggering environment, immune, hormonal and neurological responses, making them highly interesting targets for therapeutic investigation. They often share common structural traits among different functional protein groups, allowing for their classification into protein families or superfamilies.

MPs typically consist of several domains: extracellular (typically involved in cell-cell signaling and/or interactions), intracellular (performing a wide range of functions such as activating signaling pathways and anchoring cytoskeletal proteins) and intramembrane (such as pores and channels) [54]. TM proteins generally have different electronegativity and hydrophobicity profiles along their structure – they are amphipathic –, allowing them to be in contact with both water (hydrophilic) and the membrane (hydrophobic). The structure and function of many proteins – including TM proteins – depend on post-translational modifications (PTM) such as phosphorylation and glycosylation. The two major recurrent protein structure motifs in MPs are TM α-helices [55], repeatedly crossing the membranes in α-helical bundles and β-strands arranged into super-secondary structures known as β -barrels [56].

Despite their functional importance, only 4012 MP structures can be found among the 131205 determined protein structures deposited at the Protein Data Bank (PDB) [7] (statistics from June 27th 2017) – less than 1%, including multiple submissions of the same protein under a variety of experimental conditions and relatively small domains of MPs. In contrast to the number of available MP structures, there are 199.322 MP sequence clusters according to UniProt's UniRef (June 27th 2017).

Two major factors can explain this discrepancy: i) difficulties in both expression – done in several organisms [57] but mostly in Escherichia coli [58] – and purification processes and ii) challenges associated with 3D structure determination of purified MPs. Concerning the first factor, overexpression of MPs usually leads to cytoplasmic aggregates and changes in the cell metabolism. A few methods have been devised to avoid the associated cytotoxicity, such as using and tuning *E. coli* strains that are not as affected by the protein overexpression – a well-known example being "Walker strains" [59]. Protein extraction and purification is also challenging as different protein extraction conditions provide different outcomes when it comes to protein stability, state and viability for structure determination [60] (these conditions may come down to something as apparently simple as choosing the right detergent for MP isolation [61]).

MP structures solved by X-ray crystallography are often the result of a high amount of time invested in fine tuning the best experimental conditions possible. After the optimal initial conditions have been determined, further optimization is required [62] – detergent addition, utilization of different 3D continuous lipid phases (allowing the protein to freely flow) [63] or antibody fragments addition for the stabilization of protein structure [64] are examples of what might be needed to determine a MP 3D structure. Even data collection *per se* can be problematic, as the variability of crystals and their conditions (i. e. hydrophobic protein regions are camouflaged by hydrophobic solvent, making it difficult to assess the transmembrane MP structure) are responsible for preventing automated and stable data acquisition and processing [62].

NMR spectroscopy has progressed steadily, but some major drawbacks can still be identified: low sensitivity, relatively small protein size cap and the intrinsic motions of the system under investigated. Considering MPs, challenges such as sample preparation and spectral crowding arise besides those already mentioned [65]. Despite this, NMR has been somewhat efficient when it comes to studying the dynamics (e.g. relative population and conformation of different states, and exchange rates) of MPs with intrinsic conformational changes, such as channels, transporters and receptors [66]. Recently, solid state NMR have provided much better results when compared to liquid phase NMR by preventing a molecular weight cap. Unfortunately, spectral crowding is still a problem in solid state NMR. These techniques are crucial as they enable the determination of MP structure in an actual membrane and not in a "detergent simulation" of a membrane as in X-ray crystallography [67-69]. Recently, MPs have been studied by solid state NMR in their native cellular environment [70].

i) G-protein coupled receptors

GPCRs are involved in a myriad of important functions, such as vision, taste and mood regulation [71,72] and their ligands can range from a single photon to a protein [73]. While several different extracellular ligands can be identified, they share a restricted number of intracellular ligands – G-proteins, arrestins and GPCR-interacting proteins (membrane-inserted GPCR-binding proteins) [74]. G-proteins are composed of three different subunits (α, β and γ) and the α subunit acts by activating ($G_{\alpha s}$) [75] or inhibiting ($G_{\alpha i}$) [76] the cyclic adenosine monophosphate (cAMP) pathway, stimulating the membrane bound phospholipase C-beta ($G_{\alpha q/11}$) [77] or modulating the Rho family GTPase signaling ($G_{\alpha 12/13}$) [78]. The β and γ subunits constitute the βγ complex whose function stood unknown for several years. However, some roles have been identified for this complex, such as modulating the function of phospholipase C-beta [79] and activating G-protein coupled inwardly rectifying potassium channels [80]. A key aspect in G-proteins is that binding is preferential – some complexes are much more frequent than others and, while the binding interface might be similar, some key differences are likely to give rise to this differential binding. Arrestins are mostly responsible for stopping G-protein signaling through direct competition for the binding site, receptor internalization and resensitization [81]. They are, however, involved in other roles such as stress responses [82] and have, recently, gained some interest as drug targets [83,84]. The lipid environment also has an active role in modulating GPCR structure and function. For example, interaction with cholesterol significantly changes GPCRs conformational flexibility [85] and modulates their interactions. As such, it was suggested that rather than "binding sites", GPCRs, many times, have "high occupancy sites", when associated to these cholesterol "hot-spots" in the membrane.

Structurally, GPCRs share a typical pattern consisting of seven TM helixes (TMH – TMH1-7) and a perimembranar intracellular helix (HX8), and similar intracellular binding partners. Intracellular loops (ICL) and extracellular loops (ECL), which interact with different intracellular and extracellular partners and ligands, respectively, bridge TMHs in the following order: TMH1-ICL1-TMH2-ECL1-TMH3-ICL2-TMH4-ECL2-TMH5-ICL3-TMH6-ECL3-TMH7-HX8. Three High Variability Regions (HVR) have been identified in ICL3 and at the N- and C-terminal regions [8,86]. All ICLs and

HX8 have been involved in different GPCR-associated roles, such as ICL1 in receptor export from the endoplasmatic reticulum [87,88] to the cellular membrane, ICL2 in modulating dimerization and partner interaction [89] and ICL3 – the most disordered ICL – in G-protein interaction [90], and HX8 in chemokine interaction [91] and in PDZ-domain interaction [92]. As such, these should be held in high regard when studying GPCR-intracellular partner interaction, as they must be responsible for propagating the signal generated by GPCRs. Figure 1 displays the 3D structure of a GPCR concerning the aforementioned structural features.

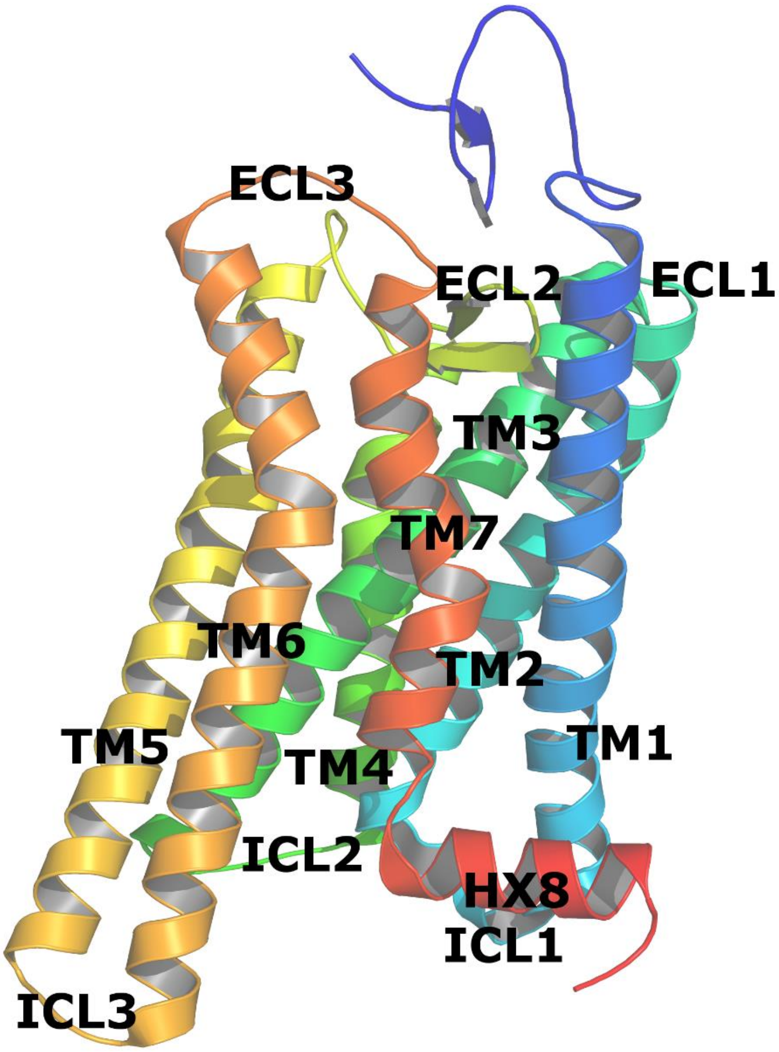


Figure 1 - The 3D structure for opsin (PDBID: 4J4Q [86]). The main structural features are easily observable and labelled. Legend: TM1 – TransMembrane Helix 1; ICL1 – IntraCellular Loop 1; TM2 – TransMembrane Helix 2; ECL1 – ExtraCellular Loop 1; TM3 – TransMembrane Helix 3; ICL2 – IntraCellular Loop 2; TM4 – TransMembrane Helix 4; ECL2 – ExtraCellular Loop 2; TM5 – TransMembrane Helix 5; ICL3 – IntraCellular Loop 3; TM6 – TransMembrane Helix 6; ECL3 – ExtraCellular Loop 3; TM7 – TransMembrane Helix 7; HX8 – Perimembranar Helix.

Considering their structure conservation and high representation and ubiquity in the human organism in both physiological functions and disease, it is no surprise that GPCRs sprouted a myriad of studies to understand their structure, especially at a computational level. Parkinson's disease (PD), for example, has harnessed comprehensive interest for the development of new drugs, and computational methods became key in doing so. GPCRs-targeted agents represent approximately 30-40% of current marketed drugs for human therapeutics and these receptors have been subjected to a substantial number of computational studies [93], namely as PD targets. Drug discovery efforts targeting GPCRs have focused on the development of orthosteric agonists/antagonists to modulate receptor activity. However, the high structural conservation of orthosteric binding sites across subtypes of GPCR subfamilies stands as the limiting factor for the design of orthosteric therapeutic agents with high receptor subtype selectivity. Additionally, orthosteric ligands interacting with some GPCRs, namely peptide or protein receptors, have physicochemical and pharmacokinetic properties that render drug discovery of small-molecule ligands unsuitable. This sparked interest in novel therapeutic agents acting as allosteric modulators of GPCRs, providing an alternative approach for subtype selectivity in treating disorders such as PD. Allosteric modulators interact with topographically distinct binding sites - allosteric sites – from the orthosteric sites of the endogenous ligands, increasing (positive allosteric modulators) or reducing (negative allosteric modulators) receptor responsiveness to ligands or the activity of the receptor. Some allosteric sites do not present high structural conservation, enabling higher subtype selectivity. Overall, exploring allosteric sites of GPCRs for drug design is of utmost importance in medicinal chemistry, enabling the possibility of targeting selective GPCR-signaling pathways without modulating others that may lead to adverse effects and to search for considerable diversity of chemical scaffolds for the optimization of the pharmacological profile of likely drug candidates [94,95]. Considering the importance of this topic, my group did a highly comprehensive review on *in silico* methods for GPCR therapy in PD [96].

Dopamine receptors, which are also involved in PD, are an important GPCR family and are divided into two major subclasses – D_1-like receptors (D_1R and D_5R) and D_2-like receptors (D_2R, D_3R and D_4R) – based on their intracellular ligands, anatomical distribution and physiological effects. They are involved in a wide array of neurological functions such as voluntary movement, sleep, attention and learning, and other roles such as mediating hormone and immune system regulation [97-99]. Furthermore, its deregulation has been implied in several diseases, namely Parkinson's disease [100,101], by the far the best studied disease involving dopamine receptors, Huntington's disease and schizophrenia [102]. As such, studying these GPCRs and, more importantly, their interaction with intracellular partners, as many other MPs, is crucial for the understanding of the molecular mechanisms underlying cellular functions and disease.

Considering that structure is of crucial importance to understand mechanisms from a molecular perspective and that experimental methods are not capable of handling a high demand, fast and accurate computational methods should be developed to make sense of the high amounts of data being generated about protein complexes.

3) Machine-learning – general aspects

Most of the work I developed in my thesis work was based on Machine-Learning (ML) techniques. ML, fundamentally, gives the computer, through the use of algorithms, the ability to learn a pattern or a stream of actions without being explicitly programmed to do so [103]. Canonically, it is has become an iterative optimization process – by repeating the same step several times (epochs), it can achieve the best possible solution. However, some more simple models, such as linear regressions, can be considered ML techniques and require only one simple mathematical function with no apparent repetition. A simple scientist/ML parallel can be established: to solve a problem using ML methods, a model first faces a problem without the necessary knowledge on how to solve it. To harness it, repeated experience is used, adapting itself – training the algorithm using a training set – and scoring its own performance iteratively (using a validation set) until the best possible predictions are reached. The validation set is used to assess whether the training of the model is progressing in the right direction. After this step, the model is tested using an independent test set. All three mentioned sets – training, testing and validation – are extracted from the original dataset. Figure 2 summarizes ML as a simple, understandable workflow.

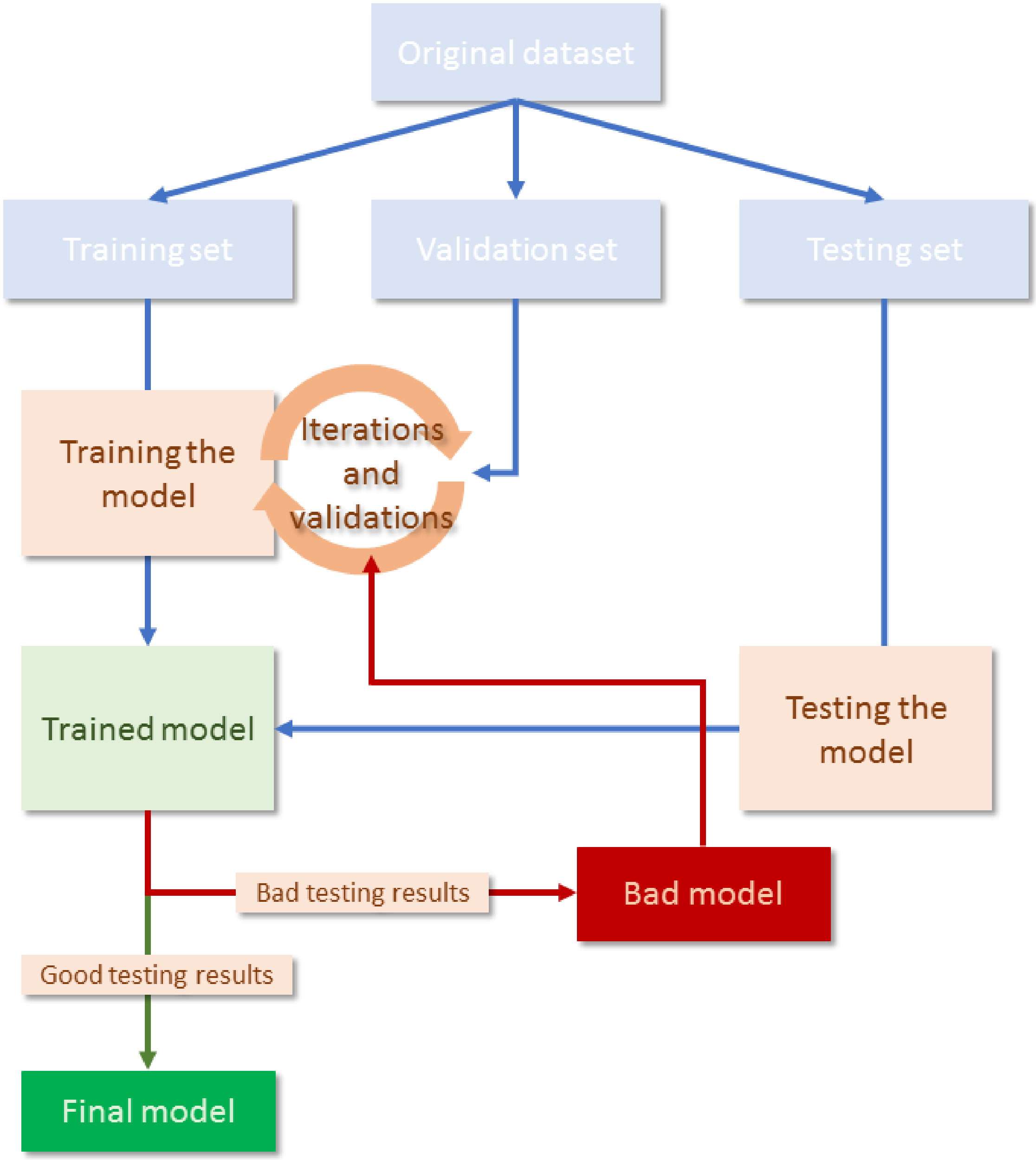


Figure 2 - The typical machine-learning workflow. The dataset is first split into training, validation and testing. Then, the algorithm is trained using the training set and after each iteration it is validated using the validation set.

ML can use either regression or classification algorithms and both are able to handle both categorical and continuous data [104]. Regression algorithms – such as linear regressions – try to find a relation between one (univariate) or more (multivariate) independent variables and a dependent variable. As such, input and output correlate with each other in a continuous fashion. While a univariate linear regression can be easily calculated using the least-squares method, for example, more complex methods might involve random forests (RFs, which can also be utilized for classification), which use typically several data subsets drawn randomly with replacement into several decision trees to train a model to provide an optimal output [105]. This process of using random data subsets with replacement to train the model is known as bootstrap aggregating or bagging. Furthermore, techniques such as tree pruning (eliminating or greatly reducing the impact of some variables) render RFs as a very powerful algorithm in machine-learning. Classification algorithms use data to provide categorical outputs. This includes binary outputs, such as classifying a surface residue as interfacial or non-interfacial based on several of its and the protein's characteristics as it will be demonstrated further ahead, or multiclass outputs, such as identifying the intracellular localization of a protein from several protein characteristics [106]. To do so, algorithms such as Support Vector Machines (SVMs), which separate classes using an artificial hyperplane, or the k-nearest neighbours algorithm, which determines the class of a new object by identifying the classes of the nearest k neighbours [107].

The previously described methods – both for regression and classification – are considered supervised learning. Supervised learning is a type of ML in which the output is known and the task is to predict that output. Unsupervised learning, on the other hand, is a bit more esoteric – without any output, the algorithm creates an artificial division, sorting the data into different groups known as clusters based on the similarity of the different samples (clustering) or attempts to identify underlying patters to simplify the dataset (dimensionality reduction). For clustering, several methods can be utilized. The most popular approach is likely to be the k-means algorithm, which assigns the data to one of k user-determined clusters. It iteratively determines the best mean for each cluster, typically defined as the centroid (the mean position) for the data points belonging to a cluster. New observations are then assigned to the nearest cluster (centroid) [108]. As for dimensionality reduction, the most commonly used approach is Principal Component Analysis (PCA), which will be described in the methods section. Figure 3 is a representation that aims to simplify these ML algorithm classifications. Annex 1 contains all ML algorithms mentioned along this thesis work, as well as a short explanation on each.

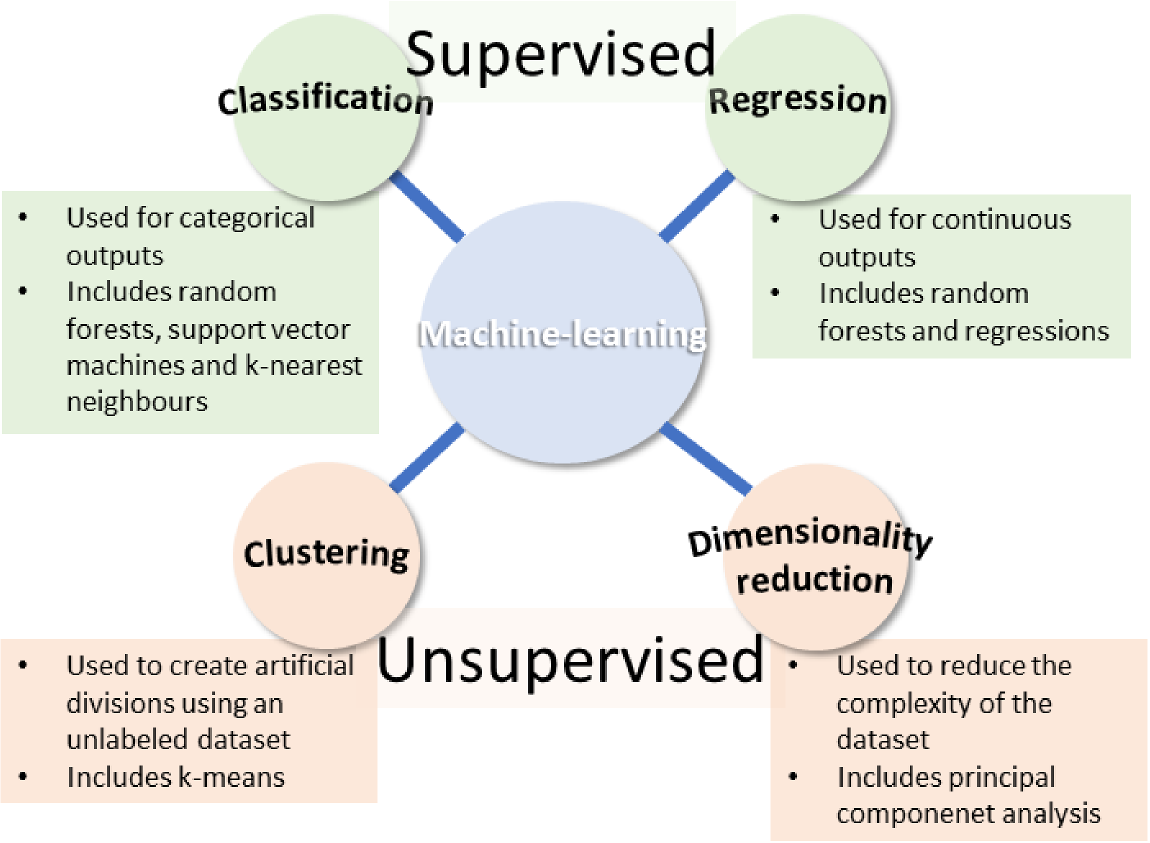

Supervised
Classification
Regression
Used for categorical outputs
Includes random forests, support vector machines and k-nearest neighbours
Machine-learning
Used for continuous outputs
Includes random forests and regressions
Clustering
Dimensionality reduction
Unsupervised
Used to create artificial divisions using an unlabeled dataset
Includes k-means
Used to reduce the complexity of the dataset
Includes principal componenet analysis

Figure 3 - Typical classification of machine-learning techniques: classification and regression (supervised), and clustering and dimensionality reduction (unsupervised).

4) Computational methods for membrane protein structure

To understand protein structure and interaction using computational tools, it was necessary to first know and understand the available methods. As such, this Introduction section will highlight important computational methods for protein structure study. It will address the relevance of sequence information for protein structure prediction and will be covering bioinformatics tools to study: soluble proteins, membrane proteins, protein-protein interactions, MP-partner interaction and GPCR-related computational methods.

i) Sequence importance for protein structure prediction

Two key concepts are necessary to understand how protein sequence can determine protein structure – Position-Specific Scoring Matrices (PSSMs) and Multiple Sequence Alignmens (MSA). PSSMs provide an easy way of determining how likely an amino acid is to be represented at a position or as having a functional property by constructing MSAs, which can, ideally, match residue pairs or vectors and reveal common sequence patterns underlying protein function and structure. To do so, they use three different classifications for each aligned residue pair – match (when the residues are the same), mismatch (when residues are different) and gap (when there is no corresponding residue). Given their central role, it is important to consider the general aspects underlying MSAs and PSSMs.

MSAs started off as techniques performing global alignment [109], which tries to match sequences using their full length. This leads to some problems, since some sequences might share homology only on some regions and, even if there are several highly homologous regions, these can be shuffled, distant or repeated [110,111]. To address this problem, local alignment techniques were developed [112], which do not require the full length of the protein. Instead, they focus on finding only the common subsequences across different proteins. Methods capable of finding subsequences with common residue pairs [113] or using only "exclusive" – nonintersecting – residue pairs [114] were developed. Even though the theory underlying local alignment makes it seem as a more capable algorithm when finding common subsequences in different protein sequences, these methods are often incapable of dealing with highly gapped common subsequences [115], making them a poor alternative.

Simossis *et al.* [116] consider three essential steps when performing MSAs: i) selecting sequences (building a database of sequences to be aligned and compared), ii) selecting an adequate scoring function that allows the comparison of sequences or subsequences and iii) iteratively applying this scoring function to build and optimize the alignment. When comparing already known proteins, sequence selection is typically not needed, unless some sequences are detrimental for the result of the final MSA due to clear differences. However, when using MSAs in order to calculate PSSMs, for example, databases comprehending thousands or millions [117] of sequences can be used to search for protein sequences. This search is usually done considering homology, using methods such as Basic Local Alignment Search Tool (BLAST) [118]. Selecting the appropriate scoring function is key in constructing the optimal MSA. These typically work column-wise (analysing each column of aligned residues at a time) and are usually the summation of all pair-wise scores.

Several scoring functions are available to consequentially evaluate the MSA through iterations – to do so, one can use scoring matrices, which either quantify the likelihood of a residue to show at a given position or use pre-calculated likelihoods, both of which are used to assess the global score of the MSA. Possibly, the best-known pre-calculated scoring matrices are substitution matrices, which are based on the observed substitution frequencies in sequence alignments. For example, mutations occurring between residues with identical nature can be considered – hydrophobic-hydrophobic mutations (leucine and isoleucine, for example) – are more likely to occur than those not keeping the position's nature. One of the earliest pre-calculated scoring matrices is the Point Accepted Mutation (*PAM*) matrix [119], which generates each residue pair score considering the probability of one residue mutating to a different one over a course of some mutations (alanine can mutate directly to arginine or it can first mutate to isoleucine and only then to arginine) – this led to the creation of several *PAM* matrices, such as *PAM250*, which considers 250 mutations for 100 residues in the sequence, or *PAM1* which considers only 1 mutation for the same sequence length. BLOcks of amino acid Substitution Matrix (*BLOSUM*) [120] is a different pre-calculated substitution matrix, used in BLAST. Instead of using global alignments, local alignments are used using several different sequence databases with different homology percentages – generating different *BLOSUMs*, such as *BLOSUM80*, with a database composed of sequences with 80% homology, and *BLOSUM52*, with a database composed of sequences with 52% homology. To calculate likelihoods from the MSA itself and combine it with the information from the pre-calculated scoring matrices, the frequency or count of a given residue in a column can also be calculated and combined with *BLOSUM* scores. The combination of both calculated and pre-calculated counts or likelihoods renders what is known as pseudocounts or pseudolikelihoods, respectively. This enables the combination of context information (from the MSA itself) and previously obtained knowledge. Additionally, it prevents scores from being 0 when the MSA-derived counts and likelihoods are 0 (for example, if at a given position in a MSA no leucine residues are observed, its count and likelihood is 0, but its pseudocount and likelihood is never 0). This is important as it would be extremely unlikely for a residue to never be represented at a given position of a MSA if enough sequences were presented and can prevent mathematical complications (such as accidentally dividing something by 0). Hidden Markov Models (HMM) are increasingly popular algorithms in bioinformatics that can also be used to derive MSA profiles. They offer great advantage – theoretically, HMMs can work with both aligned and unaligned data, and provide a solid statistical basis to sequence alignment. To generate profiles, HMMs are trained with a set of sequences to determine how likely a transition (passing from a residue to the next) is in an MSA, considering its current state and next state. The available states are deletion – a position is skipped in the MSA for a single/minority of available sequences – insert – a position is skipped in the MSA for most available sequences – and match – all sequences have a residue in that position. By deriving these probabilities for each position and for each possible residue at that position, a HMM can be built to score a sequence and build its profile. Furthermore, if the HMM is good enough, it can also be used to actually build the alignment for new sequences, considering the transition probabilities for each state [121]. After scoring all residues, techniques such as the sum-of-pair score [120], which sums every residue pair score to render a global MSA score, can be used to obtain a global MSA score that effectively evaluates the fitness of an MSA.

Upon selecting or constructing an appropriate scoring function, the MSA algorithm will then iteratively construct and improve a MSA. To build an order according to which sequences are aligned, phylogenetic trees can be used, placing more homologous sequences closer to each other and sequentially aligning each sequence according to the previous [116]. Besides doing this, it can focus on performing local alignments and considering those subsequences as starting points in global alignments. After having constructed this preliminary MSA, other techniques can be used, most of which function iteratively. To better illustrate what has been discussed, a practical example will be described – the Tree-based Consistency Objective Function for alignment Evaluation (T-Coffee) algorithm. First, T-Coffee will retrieve pairwise global and local alignments from ClustalW [122] and Lalign [123], respectively, and alignments are weighted according to pairwise sequence identity. To combine both libraries, the scores for two identical residue pairs from ClustalW and Lalgin for the same position are summed and considered as a single entry, while unique residue pairs for the same position are considered as separate entries. This will create a series of constraints, which will provide better MSAs overall. Then, it performs what the authors refer to as library extension, a heuristic process which calculates the likelihood of a pair based on triplets of matched residues – if two sequences share the same residue at a given position and, if other sequences have the same residue in that position, the weight for this residue pair will as high as the number of triplets considering the initial residue pair. Residue pairs which do not occur are given a weight of zero. By using a tree to calculate sequence similarity, the two most similar sequences are selected and the weights calculated during library extension are used to maximize the MSA score. Then, sequence pairs are added and residues are shifted until the final MSA is constructed. During this process, no gaps are removed after being added to the MSA.

Upon performing an MSA, a convenient and comprehensive way to represent is created – typically called "profiles". Numerically, this can be done using PSSMs, which constructs vectors with 20 elements for each single residue in a specific position of the MSA, resulting in a 20x20 matrix *per* position. This comprehends a heavy amount of information – 20x20xN, with N as the length of the MSA (for a MSA with 200 residues, 80000 likelihoods are determined). Visually, platforms such as Consurf [124] use colour to represent the conservation of a residue at a specific position. Although one can argue some correlation between both, PSSMs are not the same as conservation – while a PSSM, for a specific residue at a specific position, comprehends 20 different values, one for each residue, a conservation score – such as Rate4Site [125] – is a single value for each residue. Given its utilization during my thesis project, the Rate4Site algorithm for functional conservation calculation will be briefly described to provide a clearer understanding of how computational calculation of residue conservation can be done. Starting with an MSA, Rate4Site calculates phylogenetic trees using the Neighbour-Joining (NJ) algorithm. This algorithm sequentially joins sequences that are closer and therefore more similar, with each other by creating a new node in each iteration. This node always connects to the tree constructed up that point in the algorithm. Considering that a single position in the MSA has the same evolutionary rate, Rate4Site determines the maximum rate that would explain the conditional probability of the data given that rate. All rates are then normalized and standardized, resulting in a rate distribution with mean = 0 and standard deviation (sd) = 1. It becomes easy to understand why PSSMs and conservation scores are used extensively in different ML methods and how MSAs became key players in bioinformatics and computational biology.

ii) Computational methods for soluble protein structure

Computational methods for the study of soluble protein structure are considerably more developed than those for MP structure. This arises due to the much higher availability of data for both result validation and method development. For this subsection, I focused on soluble protein structure prediction techniques, which can be spread across four classes: (i) knowledge-independent ab initio methods, (ii) knowledge-dependent ab initio methods (machine-learning methods), (iii) fold recognition via threading and (iv) comparative modelling methods and sequence alignment techniques – homology modelling [126].

Theoretically, Molecular Dynamics (MD) simulations are the best possible solution when considering solely accuracy. It is a knowledge-independent *ab initio* method that simulates a physical atomic/molecular system, rendering the best structure prediction of a protein by using its sequence. However, even though MD is widely used and have proven to be effective in understanding protein motion [127,128] and other molecular mechanisms [129-133], they should be disregarded when predicting most protein structures as they require immense computational power to correctly predict the structure of a protein based solely on its sequence [134]. The first MD software to be developed, Chemistry at HARvard Macromolecular Mechanics (*CHARMM*) [135], is regularly updated, making it a highly capable program in terms of both MD analysis and model building tools. *CHARMM* has several energy functions, comprising quantum-mechanical force fields and all-atom classical potential energy functions with explicit solvent, among others

[135]. Furthermore, the conditions in which molecular systems are simulated are highly customizable, enabling both soluble and membrane-bound systems with different types of biomolecules. Assisted Model Building with Energy Refinement (*AMBER*) [136], another example of a MD simulation package, is widely used by other packages as force field. GROningen MAchine for Chemical Simulations (*GROMACS*) [137] is one of the packages that uses force fields provided by *AMBER* and other MD software programs. It is currently capable of functioning with limited computational power, through the use of several algorithmic optimizations and other methods, such as the united-atom model, reducing the complexity of the representation of the molecular system and removing some degrees of freedom [126,138].

Knowledge-dependent a*b initio* methods uses data retrieved from experimental studies to predict 3D protein structures. To do so, it relies heavily on the assumption that structural domains, subdomains or motifs are conserved across proteins and are highly related with protein sequence information [139]. A 2014 review by Dorn *et al.* on computational strategies for three-dimensional protein structure prediction [126] summarizes knowledge-dependent ab initio in 5 essential steps: i) dividing the target sequence into fragments, ii) searching for similar sequences to each fragment in a known structure database, iii) scoring the different fragments, iv) assembling the three-dimensional structure from the fragments, and v) refining the final structure. The major drawback for this process is step iv) – combining different protein fragments has to satisfy several physicochemical constraints and can be a heavy computational burden. *ROSETTA* [140] is a computer software initially created for the determination of some protein structures as a part of the third edition of the Critical Assessment of protein Structure Prediction (*CASP3*), a worldwide effort to determine new methods for protein structure determination, as well as new proteins structures [141]. A Monte Carlo simulated annealing strategy is utilized to assemble short protein fragments from known proteins with typical accuracies of 3-6 Å root mean square deviation from the experimentally determined structures. Assembling all fragments is made possible by the influence local sequence preference has on the local structure of a protein [140]. Several interactions are considered, such as solvation, electrostatic interactions and disulphide bonds, hydrogen bonds (HB), arrangement of sheets and packing into helixes. The final score for the prediction is then obtained by combining all functions into a single value. The scoring function – the potential energy surface – separates the total energy into several Bayesian components describing the likelihood of a particular structure independent of the sequence [140,142]. *ROSETTA* has been utilized for other functions as well, such as protein-protein docking [142] and loop modelling [143], among others.

FRAGFOLD [144] is another knowledge-dependent *ab initio* method for protein structure prediction assessed on the fourth edition of CASP program, *CASP4* [145]. It is based on the assembly of supersecondary structural fragments taken from highly resolved proteins. First, there is a folding simulation to identify favourable supersecondary structure fragments. *PSIPRED* [146] is used to determine secondary structures, which will then be needed to predict supersecondary structures. Then, several random conformations are calculated for the fragments until all atoms satisfy a minimum distance between different amino acid residue atoms. These conformations are then sampled by randomly selecting different fragment conformations. The component energy terms and their standard deviations are then calculated and used as weights. The optimization process is then followed, through a simulated annealing approach, which aims to minimize all weights. To keep the structure from getting stuck on local optima, 50% of all simulations are performed on a free (non-optimum) conformation.

Fold recognition via threading is a technique relying on the existence of conserved and sequence-independent protein regions and motifs [147,148]. As such, a limited number of possible protein regions and motifs have been discovered [149]. This method is particularly useful when not enough sequence information is available, thus complementing sequence-dependent methods. According to an assessment of protein folding methods done by Abual-Rub and Abdullah in 2008 [150], four steps characterize the method of fold recognition via threading: i) construction of a protein structure template library, ii) constructing a scoring function that measures the fitness between the target sequences and the templates, iii) devising an efficient, computational cost-effective algorithm to search for the best possible template(s) and iv) minimizing the scoring function in order to get the best fit possible for the selected template. Steps iii) and iv) might have to be repeated several times in order to achieve the best possible solution, making this an iterative process. *GenTHREADER* is a fold recognition via threading platform which used only Class, Architecture, Topology, and Homologous superfamily (*CATH*) [151] as its scoring function, and alignment score, length information and energy potentials as its features [152]. It comprehended a simple artificial neural network (ANN) trained to combine these three factors into a single score. As of 2003 [153], it started to include Position-Specific Iterative Basic Local Alignment Search Tool (*PSI-BLAST)* [154] scores – a protein sequence feature – with the help of Families of Structuraly Similar Proteins (*FSSP*) [155] – a database of known and aligned protein folds – and *PSIPRED* [146]. The sequence alignments are utilized to calculate pair wise potentials and solvation potentials as inputs to a multi-layer feed-forward ANN with the previous score. Two novel components were introduced in 2013 [156] – *pGenTHREADER* and *pDomTHREADER*. The first is recommended for fold recognition and is used to identify distant homologues, while the second is useful in discriminating superfamilies. The latter uses both sequence and structural data, making it a much more sensitive and selective method.

Homology modelling is a method which predicts the structure of a protein sequence using a high-sequence homology template. The main and most utilized databank for this is the *PDB* [157]. A good example of one relatively well known and used software is *MODELER* [158], which, through comparative protein structure modelling satisfies special restraints, enabling several of the processes employed currently in homology modelling, such as sequence search in databases and optimization of various protein structures using an objective flexibility function. One of the main aspects to consider in homology methods and other computational methods is that amino acids have physical properties which can be divided across two categories: *(i)* properties that favour sequentially localized interaction clusters and *(ii)* properties that favour globally distributed interactions [159] – this means that for protein structure prediction tasks it is not enough to consider the amino acid residue *per se*, but also its neighbouring residues and other interactions with remotely positioned amino acids.

Even though there have already been developed several methods for the determination of protein structures, most of these cannot be considered for membrane proteins. Knowledge-independent *ab initio* methods require the modelling of membrane, which is completely different from a cytosolic environment – while the former is an amphiphilic layer where the protein is inserted, the latter is generally a strictly hydrophilic environment. As for all knowledge-dependent methods, membrane protein-specific databases have to be constructed due to the highly different structures occurring in membrane proteins (for example, while most cytosolic proteins have a hydrophobic core and hydrophilic residues in their surface, hydrophobic amino acid residues are likely to be found in the membrane protein surface. This elicits the need to create different tools to determine the structure of membrane proteins, which will be further addressed in the next chapter.

iii) Computational methods for membrane protein structure

To understand how to adapt the methods used for cytosolic proteins to MP, a short summary is presented for some method categories:

i. Knowledge-independent *de novo* methods should consider both cytosolic and membrane environment, which makes MD simulations more computationally expensive. Since MD was not used in the vast majority of this thesis project, the reader can access a review which covers some uses of MD in MP structure prediction and contact interaction study [160];

ii. Knowledge-dependent *de novo* methods can be developed considering information on secondary structures and trans-membrane segments, for example, and how to assemble each region into a MP structure. The number of MP structures available to construct a database is, however, a limiting factor for these methods;

iii. Threading methods and homology modelling are the best techniques if homologs are to be found, since it provides good results without requiring much time. MP homologues are, however, much less abundant than those for cytosolic proteins, which complicates these methods.

As mentioned above, methods for the prediction of membrane protein structure should consider most steps mentioned for cytosolic protein structure with some key adaptations, such as the prediction of the MP's transmembrane segments and residue hydrophobicity. For example, the already mentioned *PSIPRED* [146] is an online platform, using PSSMs and artificial neural networks (ANN) to predict secondary protein structure and orientation – also known as protein topology. The aforementioned methods for topology prediction are not ideal for membrane proteins. As such, the laboratory that developed *PSIPRED* also developed *MEMSAT-SVM* [161], a topology prediction tool based on SVMs specific for MPs and which is able to discriminate between cytosolic and membrane proteins, resulting in much better predictions overall. Combining hydrophobicity scales with the prediction of secondary structures is also recommended. The most well-known scale is the White hydrophobicity scale [162], but several others have been recently developed, such as the Unified Hydrophobicity Scale [163].

Further addressing topology prediction, *OCTOPUS* [164], one of the best transmembrane α-helical segment predictors, combines four different ANN, each focused on predicting membrane, interface, loops and globular residues, through a HMM. Concerning β-barrels, an important supersecondary structure in MPs, *BOCTOPUS* [165] is a highly popular method by the same group that developed *OCTOPUS. BOCTOPUS* combines local predictions through SVMs and an HMM to combine all local SVM predictions. *SPOCTOPUS* [166], yet another method by the same group, can distinguish signal peptides from membrane proteins and predict their topology. These three different methods for MP topology prediction – *OCTOPUS, BOCTOPUS* and *SPOCTOPUS* – are ideal to highlight the rich variability of approaches and algorithms used to deal with specific problems. The methods mentioned in this paragraph are knowledge-dependent and, as such, must have access to consistent and robust databases. *ExTopoDB* [167] is a comprehensive database with topology information on 2143 MPs across 158 organisms (*accessed in* July 28th).

To combine topology prediction with hydrophobicity, Light Interfaces of High Polarity (*LIPS*) [168] uses hydrophobicity scales to predict helices and amino acid residue orientation, *PRIMSIPLR* [169], a knowledge-dependent method which uses PSSMs, hydrophobicity scales, flexibility scales and evolutionary conservation to train a SVM to identify central pore residues, and *MemBrain* [170], an online server capable of predicting transmembrane α-helices and how they interact by predicting their contacts.

While methods to predict α-helices and β-sheets and their association in MP seem to be present, some secondary structures with lower occurrence are harder to predict. According to Leman *et al.* [13], these are re-entrant helices (sometimes mentioned as "P-loops"), half-helices (α-helices that do not span the entire membrane), amphipatic helices (α-helices that lie on the surface of the membrane), trans-membrane helix kinks and β-barrels composed by more than one protein chain. The methods presented for topology prediction typically present poor results for these situations. There are, however, some methods which address some of the mentioned problems, such as *TMkink* [171] for the determination of helix kinks. The main problem with these sorts of methods is that they depend on the available information to make accurate predictions, which might not be enough to develop a reliable tool.

To calculate MP tertiary structure, the most popular methods are with no doubt homology modelling and ML methods (knowledge-dependent *de novo* methods). While the previous paragraphs focused more on the latter, the following remainder of this chapter will review homology modelling tools for MPs. The availability (or rather scarcity) of homologues is particularly relevant for homology modelling in MPs since the number of unique MP 3D structures is significantly lower than that of cytosolic proteins. Some methods have been developed specifically for membrane protein modelling, namely MEMOIR (Membrane protein modelling pipeline), [172] which can model the 3D structure of a protein of known sequence provided there are available homologous MPs with determined 3D structures, and *MEDELLER* [173], which has provided interesting results thanks to its tailor-made MP structure prediction – a sequential prediction of protein core and loops. *MEDELLER* will not generate 3D coordinates for regions for which the prediction is uncertain. This has the advantage of rendering more accurate models. On the other hands, these models are also slightly more incomplete. Structural homology modeling (threading) can overcome the lack of homologues for given sequences, however, as already mentioned, the small number of experimentally available MP structures can lead to insufficient sampling. An example of a pipeline using threading is *TMFoldWeb* [174], a web implementation of *TMFoldRec* [175]. Upon topology prediction, systematic sequence to structure alignment is performed, resulting in the selection of several templates which are ordered according to energy and reliability. Rosetta has also been widely applied to MP prediction [176]. The main improvement over soluble protein prediction was the implementation of a new membrane-specific version of the original Rosetta energy function, which considers the membrane environment as an additional variable next to amino acid identity, inter-residue distances and density [176]. Rosetta has been used to reveal important structural details in voltage sensor MPs, namely the potassium channels K(v)1.2 and KvAP channels [177], and gain insight into voltage-dependent gating [178]. Recently, *RosettaMP* was developed as a general framework for membrane protein modeling, featuring modeling tools developed in the past few years [179].

iv) Computational methods for protein-protein interactions and complexes

To understand how proteins interact and what drives the formation of a protein-protein complex from a computational perspective, it is first necessary to know what methods are available to study interface-related features in protein complexes. As such, this Introduction subsection focus on: i) interface prediction, ii) interface-related properties prediction and iii) docking.

Concerning interface prediction, an aspect that is widely regarded as critical for the identification of interfacial residues is evolutionary conservation since interfacial residues are typically more conserved when compared with non-interfacial surface residues [180]. To understand evolutionary conservation, one has to consider how evolution shapes protein sequence and structure – while residues which favour the protein structure/function are present in several proteins at specific motifs/positions, those that are detrimental are usually more variable [181]. As such, some tools that can calculate evolutionary conservation can be considered as extremely helpful when seeking to predict interfacial residues, such as *Consurf* [124] and *Scorecons* (*available in*: https://www.ebi.ac.uk/thornton-srv/databases/cgi-bin/valdar/scorecons_server.pl). Both these tools are based heavily on MSA [118,182-184]. *Consurf* provides a structural perspective for this problem, as it models the provided sequence using homology modelling and further determines residue conservation. A method which combines MSA with interface prediction is *EVComplex* [185]. Using sequences from two interacting proteins, *EVComplex* will use *EVCouplings* [186] to predict intramonomer and intermonomer interactions, using this information to predict the structure of a protein complex using only the sequence of both interacting proteins. A key aspect of *EVComplex* and *EVCouplings* is that it considers co-occurring mutations as accurate predictors for intramonomer and intermonomer contacts. They happen when mutations in an apparently conserved region are accompanied by a mutation in another protein region or in a different complex – since evolution drives mutation permanence, if two residues are considered to mutate together, it is likely that they interact with each other. *PS-HomPPI* and *NPS-HomPPI* [187] uses only sequence as well, and predicts interfacial residues based on the interfacial residues of homolog proteins with data on interfacial residues. This method can depend be partner-specific (*PS-HomPPI*) or non-partner-specific (*NPS-HomPPI*), but data suggests that partner specificity increases the accuracy of the results. In fact, an important note made by Xue *et al.* in their 2015 review [188] is that partner information is very valuable for protein interface prediction, which is often overlooked. A comparison of the results obtained through *PPIPP* and *PAIRpred* – with partner information – with the ones from *PSIVER* [189] (sequence-based) and *SPPIDER* [190] (structure-based) – which will discussed briefly – proved that partner information greatly improves the predictions made. As for methods using monomer structure, *SPPIDER* [190] is a machine-learning approach that predicts interfacial residues based on the predicted relative solvent accessibility (which uses the unbound monomer solvent accessibility and other structural features), WHat Information Does Surface Conservation Yield? (*WHISCY*) [191] uses structure to define surface residues and to smooth the prediction and calculates conservation for all surface residues, and Consensus Prediction Of interface Residues in Transient complexes (*CPORT*) [192] uses several other methods that predict interfacial residues to make its own predictions, making it a meta-server. Evolutionary features, regardless of their popularity, have a considerable disadvantage – they are quite successful only when a high number of homologs are available [193,194]. As such, methods which are able to abstract from evolutionary conservation are bound to be more robust across all sorts of protein sequences. One example is the method developed by Wang *et al.* [194] for intramonomer contact prediction. It processes sequences using an ultra-deep ANN and convolutional neural networks (CNN) to accurately predict protein contacts without using evolutionary conservation.

Interface-related characteristics are plentiful and necessary to describe the interface – these can range from H-bonds, salt bridges, hydrophobic interactions, solvent accessible surface area (SASA), number of nearby atoms, total number of interface atoms, polar and apolar energy in the interface, hot-spots (HS), hot-regions and so forth. To predict H-bonds, salt bridges and hydrophobic interactions, one can use software programs such as Visual Molecular Dynamics (VMD) [195] and PyMol [196], which provide several built-in tools to do so and are easily scriptable, making these characteristics easy to determine at a high-throughput level. SASA can also be calculated using VMD, but some servers are able to provide better results, such as bioCOmplexes

COntact Maps (*COCOMAPS*) [197], which calculates several interface features, such as total SASA, SASA in terms of independent residues, total interface polar and apolar area, as well as H-bonds and number of nearby atoms, making it a relatively comprehensible and quick method for interface characterization. HS, defined as residues which, upon alanine mutation, cause an alteration of over 2 kcal/mol in the free binding energy difference ($\Delta\Delta G \geq 2$), represent residues which are highly important for the protein-protein interface [198], while null-spots (NS) are all interfacial residues which are not HS. As such, to understand which protein-protein interface regions or residues are important without using experimental methods, computational tools can be used. *SpotOn* [199], a tool developed during this thesis project which will be described shortly, representing the highest sensitivity and accuracy for HS prediction so far, makes use of several features such as the ones previously described for the characterization of the protein-protein interface to predict which residues should be considered HS or NS[200]. As for hot-regions – characterized as HS clusters (HS are not randomly distributed across the protein-protein interface but rather clustered [201]) – *HotRegion* [202] can be used, which uses *HotPoint* [203] to predict HS across all PDB entries.

Docking is the process focused on finding the best possible conformation for two proteins interacting with each other. Typically, it starts by finding the best orientation possible through rigid docking – both monomers remain unaltered as several orientations are sampled. Then, it refines the structure by semi-flexible or flexible docking, which enables some coordinate fluctuations on one or both proteins, respectively [204]. The key aspect that typically differentiates docking algorithms is how they perform the rigid docking (search phase) and how they rank each of the complex structures (scoring phase). As for other structural tasks, several computational methods have been developed. For example, *ZDOCK* [205] is a docking algorithm with a search phase based on the Fast Fourier transformation and a scoring phase based on shape complementarity, electrostatics and statistical potential terms, and High Ambiguity Driven protein-protein DOCKing (*HADDOCK*) [206] is a docking algorithm whose main feature is the integration of information on interfacial residues, greatly reducing the search phase and improving the scoring phase, rendering better, more realistic complexes.

While some of the methods described – especially those that use previous information on the interface, such as interface characterization methods and HADDOCK [206] – work in MPs without the need for any major adaptations, interface prediction and standard docking tools are much more challenging endeavours in MP study. As such, the following chapter will focus on methods developed specifically for this.

v) Computational methods for membrane protein-protein partner interactions

As with MP structure prediction, MP interface prediction typically requires tools different from those for cytosolic proteins. Methods for MP-MP complexes and MP-cytosolic protein complexes will be considered in the following paragraphs.

Concerning MP-MP complexes, Some algorithms have been developed for this purpose, such as SVMs [207] and RFs [208] with residue type distribution and evolutionary conservation features using a MP structure as input and residue averaging (areas with several residues classified as interfacial are more likely to be predicted as interfacial). The number of contacts can also be decisive in identifying protein-protein interfaces (PPI) – by knowing how many contacts exist in a protein-protein interface, excluding regions incorrectly classified as interfacial becomes much easier. *TMH-Expo* [209] is method developed for intramonomer contact prediction in multi-spanning helical MPs using ANNs. Even though this method is used for intramonomer contacts, an adaptation of this algorithm for MP interfaces could be of great use in MP complex prediction.

Concerning MP-cytosolic protein interactions, *ProMate* [210] is an interesting example of a structure-based method, which uses several features such as secondary structure, length of non-secondary structure protein regions and pairwise amino acid residues distribution to calculate an interface propensity value for each residue. Part of the development of *ProMate* involved the elimination of redundant or highly correlated features, which reduces computation and search space. Protein-Protein Interaction Prediction Platform (*PPIPP*) [211] is a good example of a sequence based method by using propensity scores based on the presence of a given residue compared to any other residue at the interface. To solve the lack of partner information, the model was trained by comparing residues in intermolecular protein-protein interface with intra-protein contacts. *PPIPP* is built on 24 ANN and returns the average score as final score, using PSSMs as one of its main features. *PAIRpred* [212] is an hybrid approach, using both sequence and structure-based features: the structure-based features consist of relative Surface Accessible Surface Area (SASA), residue depth, half sphere amino acid composition and a protrusion index, while the sequence-based features are based on PSSMs and predicted relative accessible surface area. All these are combined through a SVM to predict protein-protein interactions.

Docking in MPs has also been improved in recent years. *Memdock* [213] (specific for α-helical MPs) and docking tools in *RosettaMP* [179] take into consideration the lipid bilayer environment for both search and scoring phases. The key aspect of these algorithms is how they reduce the search space by considering that MP complexes are structurally confined and do not have the same conformational flexibility as cytosolic protein complexes. These assumptions are key when dealing with MP complex structure prediction as they allow researchers to overcome the lack of experimental knowledge on the subject by using trustworthy assumptions that eliminate several incorrect conformations. Scoring functions are also adapted considering these structural restrictions – while exposed surface hydrophobic residues/regions might be favoured in membrane-inserted regions, they are considered detrimental in solvent-facing protein regions, and hydrophilic residues/regions in membrane inserted regions are more likely to be considered for interfaces in MP complexes.

This particular subsection should alert the reader to a particularity – computational methods become scarcer as MPs are further considered. However, some MP groups are highly studied due to their relevance and specific methods for these MP groups became widely available. One such example is GPCRs, for which several computational methods have been developed. Some of these methods are reviewed in the following subsection, highlighting *GPCRdb* [214], one of the most ambitious bioinformatics efforts to date.

vi) Computational methods for GPCRs

GPCRs have been extensively considered from a computational perspective, leading to one of the most well-known online interfaces for computational biology – *GPCRdb* [214]. Some of the most relevant and tailor-made tools are reviewed in this chapter.

Alignments are one of the key tools in studying GPCRs, particularly due to the Ballesteros-Weinstein (BW) nomenclature [215], a residue identification system based on the most conserved residue of each TMH. This makes comparisons across several different GPCRs much easier as common and conserved residues are much more easily identified when analysing heavy amounts of data. *GPCRdb* provides a platform for automatic residue numbering, including the BW nomenclature. One other key aspect when studying GPCRs through computational methods is, upon the absence of a resolved 3D structure, to find a template sufficiently adequate to accurately model GPCRs. *GPCRdb* provides tools for high homology template selection, as well as a database of known homology models to aid in this task. Nonetheless, some tools such as *GPCR-ModSim* [216] have been developed. *GPCR-ModSim* models GPCRs considering active/partially active/inactive states using homology modeling and includes a MD refining step using the membrane-inserted GPCR.

Concerning GPCR partners, *GPCRdb* provides several tools to study GPCR-G-protein coupling and G-protein alignments. Phylogenetic maps are also provided to help the user understand how evolution drove this sort of interaction. Another interesting tool available in the website is the Interface Mapping tool, which uses the resolved structure of the β2-adrenergic receptor upon complex formation with a Gs-protein (PDBID: 3SN6 [217]) to infer the GPCR-Gs-protein interaction site for all G-proteins. This approach, however, is not reliable – using a single template to derive interacting residues across several different G-proteins is not robust as key differences will probably arise and explain how differently the Gs-protein interacts with different GPCRs. Furthermore, it leaves other GPCR-G-protein interactions unexplained.

While computational tools for cytosolic protein and MP structure have been widely regarded, studying protein interfaces is still challenging. As such, this project focused on some key aspects of protein interface computational study, namely the development of SpotOn [199], a method for the prediction of HS, which are thought to be essential in interfacial characterization, building the SpotOnDB, a comprehensive overview of PPI concerning structural and evolutionary features using the PPI4DOCK non-redundant complex dataset [218], using ML to predict interfacial residues in monomer structures, and using dopamine receptor-partner interactions as a case study to understand how receptor dynamics affects partner differential binding. By doing so, this project covers method development, utilization of high-throughput methods to make sense of high amounts of data, and focusing on a set of complexes between similar proteins to understand how differential binding is affected by what are apparently small nuances.

B. Methodology

1) SpotOn – prediction of Hot-Spots from complex structural information

During this thesis I participated in the development of SpotOn [199], a method for HS prediction from 3D complex structure It relied on the collaboration by researchers from the Center for Neuroscience and Cell Biology from the University of Coimbra, the Bijvoet Center for Biomolecular Research from the Utrecht University, the Centro de Ciências e Tecnologias Nucleares from the University of Lisbon, Department of Genetics and Genomics and Icahn Institute for Genomics and Multiscale Biology from the Icahn School of Medicine at Mount Sinai and Centro de Matemática da Universidade do Porto.

As described previously, all ML based methods comprehend some essential steps:

i. Compiling all known cases into a comprehensive dataset – for this case, all known complexes with experimentally determined structure and experimental information on HS were gathered in a non-redundant database;

ii. Training the prediction model – this step should use a fraction of the dataset to determine what are the best parameters of a ML algorithm, enabling it to perform the best for the prediction of novel cases;

iii. Testing the prediction model – after the model training step, an independent test set should be used to confirm the model's best parameters.

i) Dataset construction

By combining the Alanine Scanning Energetics database (ASEdb) [219], the Binding Interface Database (BID) [220], Protein-protein Interactions Thermodynamic database (PINT) [221] and Structural database of Kinetics and Energetics of Mutant Protein Interactions (SKEMPI) [222] databases to construct a non-redundant dataset of mutations, 534 mutations were compiled across 53 different non-redundant complexes. These databases gather information on ΔΔG values and residues were considered as HS if, upon alanine mutation, ΔΔG ≥ 2.0 kcal/mol. 3D structures were retrieved from the PDB [223]. To ensure maximum variability across all complex interfaces, all sequences were filtered to ensure at most 35% sequence homology in each interface. Hydrogens were added by an in-house VMD [195] script and only protein atoms were considered.

ii) Structural/sequence features

Concerning structural features, ten SASA-related features were calculated from unbound, bound and standard SASA values (${}_{mon}SASA_i$, ${}_{comp}SASA_i$ and ${}_{res}SASA_i$, respectively) in Equations 1-10 and ${}_{mon}SASA_i$ and ${}_{comp}SASA_i$ were used as well. Twenty features corresponding to interfacial residue count were also added, one for each residue. Intermolecular atomic contacts within 2.5 Å and 4.0 Å, and the number of intermolecular hydrophobic interactions were also used as features. These were calculated using in-house VMD software [195] scripts, which are incorporated in the SpotOn pipeline.

$$\Delta SASA_i = |{}_{comp}SASA_i - {}_{mon}SASA_i| \tag{1}$$

$$_{rel}SASA_i = \frac{\Delta SASA_i}{_{mon}SASA_i} \tag{2}$$

$$_{comp/res}SASA_i = \frac{_{comp}SASA_i}{_{res}SASA_r} \tag{3}$$

$$_{mon/res}SASA_i = \frac{_{mon}SASA_i}{_{res}SASA_r} \tag{4}$$

$$_{\Delta/res}SASA_i = \frac{\Delta SASA_i}{_{res}SASA_r} \tag{5}$$

$$_{rel/res}SASA_i = \frac{_{rel}SASA_i}{_{res}SASA_r} \tag{6}$$

$$_{comp/ave}SASA_i = \frac{_{comp}SASA_i}{_{ave}SASA_r} \tag{7}$$

$$_{mon/ave}SASA_i = \frac{_{mon}SASA_i}{_{ave}SASA_r} \tag{8}$$

$$_{\Delta/ave}SASA_i = \frac{_{\Delta}SASA_i}{_{ave}SASA_r} \tag{9}$$

$$_{rel/ave}SASA_i = \frac{_{rel}SASA_i}{_{ave}SASA_r} \tag{10}$$

As for sequence features, PSSMs and the corresponding weighted observed percentages were computed using BLAST [118,224], providing forty additional features. Sequence related features were extended to include those 805 extracted from the *protr* [225] package from R:

i. Amino Acid Composition (AAC) of the protein (fraction of each amino acid type within the protein);
ii. Pseudo Amino Acid Composition (PAAC) [226] (adds up to the standard 20 amino acid definition with information on residue motifs),
iii. Amphiphilic PAAC (a set of the twenty original amino acids, plus descriptors regarding the hydrophobicity/hydrophilicity of the sequences that have often displayed positive effects regarding protein-protein interaction prediction algorithms);
iv. BLOcks Substitution Matrix (BLOSUM) (evolutionary features in the form of a scoring matrix upon sequence alignment taking into account amino acid substitution at a 62% level of similarity);
v. Protein Fingerprinting (the identification and differentiation of proteins by unique characteristics using the amino acid index and PCA (which will be described ahead));
vi. ProteoChemometric Modeling (PCM) [227] (PCA of 2D and 3D descriptors to describe protein dynamics and ligand interaction).

PAAC is highly informative as it does not include residue composition solely, but also long-range correlations of the physicochemical properties between two residues, with valuable results in protein classification tasks [228-232] . A final number of 881 features were therefore calculated for all 534 observations, composed of 127 HS and 407 NS. 55 features are based on the amino acid residue, while the remaining are based on the whole protein.

As observable, this dataset has more features than observations and more negative cases (NS) than positive cases (HS) – this is known as an unbalanced dataset, since there is more information on one class than the other. As such, overfitting becomes a problem. Overfitting happens when the trained model is only able to explain the data used to train it. By having too many features characterizing each observation or too many of one class, overfitting becomes much more likely. The following subsection will explain how to deal with these problems, as well as some considerations on ML.

iii) Machine-learning techniques

The R programming language [233] was used to perform ML and most statistical analysis, together with the Classification And Regression Training (*caret*) [234] package, which provides an elegant and high-throughput way of doing machine-learning. The dataset was randomly split into training (70% (374) of all observations) and testing (30% (160) mutations/observations) sets. An equal proportion of positive/negative cases is maintained across all subsets. 51 different algorithms in the *caret* package were tested: Boruta, C5.0, C5.0Rules, C5.0Tree, LogitBoost, ORFlog, ORFpls, ORFridge, ORFsvm, RRF, RRFglobal, ada, adaboost, amdai, avNNet, bagEarth, bagEarthGCV, bagFDA, bagFDAGCV, ctree, ctree2, dwdPoly, dwdRadial, evtree, fda, gamboost, glm, glmboost, hdda, knn, lda, lda2, loclda, multinom, nb, pda, plr, qda, ranger, rda, rf, stepLDA, stepQDA, svmLinear, svmLinear2, svmPoly, svmRadial, svmRadialCost, svmRadialSigma, svmRadialWeights and wsrf.

Overcoming overfitting can be done using several techniques. Down- and up-sampling were both used – in the first a random subset of all classes in the training is generated so that each class size matches the size of the least prevalent class, while in up-sampling, random sampling of the minor class with replacement is performed so that the size of the minor class (HS) matches that of the major class (NS). Both down- and up-sampling are techniques that help in dealing with overfitting by equalling the number of positive and negative classes. To further prevent overfitting, 10-fold cross validation repeated 10 times was used. This technique splits the training subset 10 times into training (80%) and validation (20%) subsets. The latter is used to assess how the model is performing across each iteration. By using several different validation subsets to validate the model, it is guaranteed to fit no particular subset of the data.

A dimensionality reduction technique – PCA – was also used to prevent overfitting. To do so, it first characterizes the first principal component (PC) by determining the data projection with the largest variance. Consequently, it determines the following PCs by considering the orthogonal data projections relatively to the already calculated PCs with the greatest variance. By doing so, it is able to describe the data using a greatly reduced number of dimensions [235]. Apart from PCA, all data was scaled according to Equation 11, in which N is the normalized vector (feature), V is the original vector, and m and σ are the vector's mean and sd, respectively. This allows us to consider each feature has having a similar distribution, with mean = 0 and sd = 1.

$$N = \frac{V - m}{\sigma} \tag{11}$$

Considering the previously described data treatments, six pre-processing conditions were tested to explore the best possible results:

a. *Scaled* – normal scaling of the variables;

b. *ScaledDown* – normal scaling of the variables with down-sampling of the negative class;

c. *ScaledUp* – normal scaling of the variables with up-sampling of the positive class;

d. *PCA* – normal scaling of the variables and PCA of the dataset with down-sampling of the negative class;

e. *PCADown* – normal scaling of the variables and PCA of the dataset;

f. *PCAUp* – normal scaling of the variables and PCA of the dataset with up-sampling of the positive class.

The best pre-processing condition, according to average Area Under Receiver Operating Curve (AUROC) and sensitivity (which will be explained shortly ahead) was chosen to further develop the method. Validity and performance in machine-learning was done using several methods, namely the AUROC, the Accuracy (equation 1.1), the Sensitivity, the Specificity, Positive Predictive Value (PPV), Negative Predictive Value (NPV), False Discovery Rate (FDR), False Negative Rate (FNR), F1-score and Mathew's Correlation Coefficient (MCC). Except for AUROC, all values are calculated using the model's predictions, considering four distinct cases: true positives (TP – the model's prediction for the positive class is correct), true negatives (TN – the model's prediction for the negative class is correct), false positive (FP – the model predicts an observation as belonging to the positive class while it is in fact a negative class) and false negative (FN – the model predicts an observation as belonging to the negative class while it is in fact a positive class), are described in Equations 12 – 20 and, just as the AUROC, were calculated using R. AUROC uses a plot with 1 – Sensitivity in the x-axis and the Specificity in the y-axis. Three points are considered: (0,0), (1 – Specificity, Sensitivity) and (1,1) and the area underneath this "curve" is calculated, thus rendering a metric that considers both Sensitivity and specificity.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{12}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{13}$$

$$Specificity = \frac{TN}{TN + FP} \tag{14}$$

$$PPV = \frac{TP}{TP + FP} \tag{15}$$

$$NPV = \frac{TN}{TN + FN} \tag{16}$$

$$FDR = \frac{FP}{FP + TP} = 1 - PPV \tag{17}$$

$$FDR = \frac{FN}{FN + TN} = 1 - NPV \tag{18}$$

$$F1 - score = \frac{2TP}{2TP + FP + FN} \tag{19}$$

$$MCC = \frac{TP * TN - FN * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} = 1 - NPV \tag{20}$$

The final algorithm used an ensemble of the 5 best performing methods to further reduce overfitting and improve performance by combining the prediction of several different techniques [236] using a logistic regression function. To do so algorithms were clustered into 5 groups according to *caret*'s tags [234], and a detailed explanation can be found at Moreira *et al.* [199]. A logistic regression is handy when it comes to probabilities: its output equals 0 or 1 as its input approximates negative or positive infinity, respectively, and provides a smooth transition between 0 and 1. The method is publicly available in http://milou.science.uu.nl/cgi/services/SPOTON/spoton/ [199].

2) SpotOnDB – understanding protein-protein interfaces through big data approaches

SpotOnDB made extensive use of the SpotOn pipeline with some additional being calculated with bioinformatics and programming tools. It aimed to further characterize and understand PPIs using high-throughput data analysis in the PPI4DOCK [218] non-redundant dataset.

i) Protein-protein Dataset

Our dataset was composed by 1403 protein-protein complexes retrieved from the PPI4DOCK database, totalizing 66.710 interfacial residues [218]. Even though 1.417 PDB entries are present in the original database, the remaining 14 complexes do not have sufficient homologs to calculate PSSM values, an important feature of the SpotOn classifier. The complexes are non-redundant (at the 70% sequence identity level) heterodimeric complexes with resolutions lower than 3.5 Å, with a biological assembly containing no more than 20 chains, classified as a biological entity, and with an interface size greater than 300 Å^2.

ii) Hot-spot classification

SpotOn [199] is based on the previously described machine-learning ensemble which uses the 3D structure of a complex as input and determines several features for interfacial residues with the ultimate goal to classify them as HS or NS. Apart from the previously described features calculated for the SpotOn classification algorithm, some extra-features were calculated using this pipeline, namely Hydrogen-Bonds, Salt-Bridges [237] and hydrophobic interactions [238,239].

iii) Protein-protein interfacial characterization

All features calculated were compared in residues classified as HS and NS. A few additional evolutionary- and structure-based features were calculated for a better characterization of PPIs. These were:

<u>*Enrichment factors:*</u> Enrichment factors or values ($E.Factor$) are an easy way of comparing the frequency of residues as HS using the average HS number on the interface. This feature (Equation 22) was first described by Bogan *et al.* [198]:

$$E.Factor = \frac{p_{HS/res}}{p_{res/inter}} \tag{22}$$

$E.Factor$ is defined as the ration between the proportion or percentage of a particular residue as HS ($p_{HS/res}$) and the proportion or percentage of that residue in the interface ($p_{res/inter}$). Residue-wise HS were also calculated according to Equation 23:

$$Residue - wise\ HS = \frac{n_{HS/res}}{n_{inter/res}} \tag{23}$$

$n_{HS/res}$ is the number of predicted HS for a given residue and $n_{inter/res}$ is the number of interfacial occurrences of that same residue. HS distribution was defined as the percentage of residues acting as HS out of the total number of HS.

<u>*Propensities:*</u> Protein surface residues were considered as residues with SASA larger than 1 Å^2 as calculated by the VMD software [195], core residues were considered as all non-surface residues

and interface residues were retrieved from the individual PDB files. Protein core, non-interface surface and interface residue propensities and normalized residue propensities were calculated as described in Yan *et al.* [240].

B-factors: Values for individual atoms were retrieved from all PDB files using a Python 2.7 script. Values were normalized for each complex following the protocol defined by Liu *et al*. [241], in which normalization is performed by considering individual PDB files instead of the entire dataset, as demonstrated in Equation 3:

$$B_{norm}^{i} = \frac{B^{i} - \bar{B}}{\delta_B * 1.645} \tag{24}$$

where B^{i} is the B-factor value of the atom, and $\bar{B}$ and δ_B are, respectively, the mean and sd B-factor values for the complex, rendering B_{norm}^{i}, the normalized B-factor value. Mean normalized B-factor values for each interfacial residue were also calculated.

HS regions: These regions were defined by Keskin *et al*. [201,242] as clusters of tightly packed HS that come in contact with each other. VMD [195] tailor-made scripts were utilized to calculate neighboring residues within 8 Å and 10 Å of all interfacial residues. A posterior distinction of neighboring residues as intramolecular (neighboring residues within the same monomer) or intermolecular (neighboring residues present in the partner protein) Hot-regions was also made.

Evolutionary conservation: Residue sequence-conservation was retrieved for all complexes using Consurf [124], based on the Rate4Site algorithm [125] for calculating position-specific conservation scores for each amino acid residue. Only interfacial residues were considered. The selected program for multiple sequence alignment was MAFFT [183,184] using the BLAST [118] on the UNIREF90 database [243]. For each protein sequence a maximum of 150 selected sequences were retrieved, having between 35% and 95% homology.

iv) Statistics, data visualization and webserver implementation

Statistics presented in this paper, namely mean and sd values, as well as data visualizations plots (ggplot2 based [244]) were calculated and displayed using the R-statistical package [233]. *Plotly* [245] was used for the dynamic visualization in the website, which was constructed using the *shiny* application package [246]. It is freely available at http://milou.science.uu.nl/services/SPOTONDB/ or http://45.32.153.74/spotondb/. The complete dataset with all calculated features and residue classification will be publically available, with a summary description of the columns composing the dataset in Annex II.

3) Interface Prediction – using deep-learning to identify interfacial patches in protein surface

Using a pipeline similar to the one utilized for SpotOn (described above), predicting the interface revealed itself as a more challenging endeavour. As such, more ambitious techniques and models were employed, namely Deep-Learning (DL) through the use of *h2o* package, an implementation of the h2o deep-learning framework [247], and the *bio3d* package [248] for the R programming language. The main advantage of using *h2o* is that it uses it performs computations using its own cluster, rather than in R memory, leading to much faster model training. *bio3d* was used to calculate distances between residues in PDB files to identify interfacial patches.

i) Dataset construction and features

The same dataset used for SpotOn was utilized for this part of my thesis work, with some changes: i) only H-bonds and salt-bridges were calculated concerning structural interactions and the total number of structural interactions in the interface and per residue were used as features in the construction of the final dataset, ii) only monomer-related SASA features were utilized (${}_{mon}SASA_i$, ${}_{mon/res}SASA_i$ and ${}_{mon/ave}SASA_i$), iii) evolutionary conservation was used, as calculated by Consurf [124], and iv) all interfacial residues were considered. This resulted in a dataset with 15.508 residues and 907 features, each classified as interfacial or non-interfacial. Features were filtered for near zero variability resulting in no removed features, and feature covariance was calculated across all features and a threshold of 75% was chosen to remove highly correlated features, which led to a final count of 655 features.

ii) Model hyperparametrization, parametrization and training

Concerning that *h2o* provides a highly rich environment for DL model concerning hyperparameters, a random discrete grid search was performed to determine the best possible hyperparameters.

Table 1 describes the tested hyperparameters in the random discrete grid search, which was done with a training set – 50% of the full dataset (Train) – further divided into a training subset – 60% of the training set – and a validation set – 20% of the training set – trained for 5 epochs. AUROC was used used to select the best hyperparameters. The tested hyperparameters are different activation functions (how each neuron determines the output considering its inputs), different hidden layer compositions (the number of layers (the number of elements in each array) and the number of hidden neuros in each layer), different input dropout ratios (how often should an input neuron be dropped out), and different values for L1 and L2 regularization. Activation functions are an important discussion topic and each should be described:

- Rectifier functions are equal to 0 for any x < 0 and linear for any x > 0 [249];
- Tanh functions are hyperbolic tangent functions and are composed of two plateaus at -1 and 1 when x approximates positive and negative infinity, respectively, and a smooth gradient from -1 to 1 [250];
- Maxout functions are composed of several linear models and the output is the maximal value out of all linear models for any given input [251];
- RectifierWithDropout, TanhWithDropout and MaxoutWithDropout functions are the same as Rectifier, Tanh and Maxout functions, respectively, but neurons in the hidden layer are randomly eliminated. This prevents overfitting by stochastically removing neurons which might be fitting to closely with the data [249].

L1 and L2 regularization are used to reduce overfitting and are described according to Equations 24 and 25:

$$L1 = \left\|w\right\|_1 = \sum_{i=1}^{n} |w_i| \tag{24}$$

$$L2 = \left\|w\right\|_2^2 = \sum_{i=1}^{n} w_i^2 \tag{25}$$

where L1 and L2 are the L1 and L2 regularizations and w_i is the weight of a single neuron. This restraint ensures that the weights will not grow to excessively large values, which can be a cause of overfitting [249]. Furthermore, a stopping tolerance of $5*10^{-2}$ was used considering 2 stopping rounds (if the method did not show an improvement superior to $5*10^{-2}$ in AUROC for 2 rounds the model being trained as a part of the grid search stops and the next one is trained) and the maximum squared sum of incoming weights per unit was set to 10. Concerning the search criteria, a maximum of 100 models were generated, with a maximum runtime of 360 seconds and a stopping tolerance of 10^{-2} for 5 stopping rounds.

Table 1 - Tested hyperparameters for the random discrete grid search for the deep-learning model.

Hyperparameters	Tested conditions
Activation functions	Rectifier, Tanh, Maxout, RectifierWithDropout, TanhWithDropout, MaxoutWithDropout
Hidden layer composition	[100,100], [200,200], [100,100,100], [200,200,100], [500,200,100], [200,200,200], [500,200,100], [500,500,100], [500,500,200,100]
Input dropout ratio	Values ranging from 0.01 to 0.1 with a step of 0.01
L1 regularization	Values ranging from 0 to 10^{-3} with a step of 10^{-6}
L2 regularization	Values ranging from 0 to 10^{-3} with a step of 10^{-6}

Feature pre-processing consisted only down- and up-sampling the negative and positive class, respectively, and in using Synthetic Minority Over-sampling TEchnique (SMOTE) resulting in four datasets – Full, DownSample, UpSample and SMOTEd. SMOTE is a technique to generate synthetic entries by taking each underrepresented class sample and introducing examples along the line joining the k underrepresented class nearest neighbours [252]. It is necessary to use SMOTE with some caution in this case as high-dimensionality might be a problem for SMOTE [253]. The data was scaled as a part of the *h2o* DL function, which automatically scales the data upon user request. Deep-learning models ware trained with the same data subset used for the random discrete grid search for all four datasets for 20 epochs using 10-fold cross-validation with the optimal parameters obtained from the random discrete grid search, considering the AUROC as the metric to evaluate performance across each epoch. Testing was done using 25% of the full dataset (Test). The same metrics used for SpotOn were utilized to assess the models' performance.

iii) Interfacial patch prediction

The number of α-carbons belonging to neighbouring predicted interfacial residues at different cut-off distances (5 Å, 8 Å, 10 Å, 15 Å, 20 Å and 25 Å) to test if this feature is relevant in interface prediction. The rationale behind this process is that interfaces are usually composed by more than 13 residues, as results from the SpotOnDB assessment showed. As such, the number of neighbouring predicted interfacial residues was used in an attempt to enhance interface prediction. To do so, distance matrices were calculated for all residues' α-carbons in all PDB files using the *bio3d* package [248] for the R programming language and results were filtered for surface/interfacial residues only. Upon doing so, total predicted interfacial residues at 5 Å, 8 Å, 10 Å, 15 Å, 20 Å and 25 Å cut-offs and the interface prediction were used as input to a logistic regression. The performance of the final pipeline (interfacial residue prediction and interfacial patch prediction) was assessed using the same metrics as those for the SpotOn method and the remaining 25% of the full dataset (Test2).

4) Beyond the interface – how complex structure dynamics and conformation affects the binding interface

Understanding GPCR-partner interaction is of major relevance for molecular biology. As such, dopamine receptors were used as a case study to better understand how differential binding happens in GPCRs. To do so, homology models were made of several dopamine receptors, G-proteins and arrestins, and several bioinformatics tools were used to characterize GPCR-partner complexes.

i) Homology modelling

Sequences for all proteins to be modelled were retrieved from the Universal Protein Resource (UniProt) database [254] and saved in the Fast Adaptive Shrinkage/Thresholding Algorithm (FASTA) format. The modelled proteins were dopamine receptor 1 (D1R – P21728), dopamine receptor 2 (D2R – P14416), dopamine receptor 3 (D3R – P35462), dopamine receptor 4 (D4R – P21917) and dopamine receptor 5 (D5R – P21918) from the β2-adrenergic receptor in PDBID: 3SN6 [217], the G-proteins G_q (P50148), G_z (P19086), G_{t2} (P50149), G_{i1} (P63096), G_{i2} (P04899), G_{i3} (P08754), $G_{s(sh)}$ (P63092), G_o (P04971), $G_{s(lo)}$ (GI:20147687) and G_{oB} (GI20147683) from the G_s in PDBID: 3SN6 [217], and arrestins 2 (P49407) and 3 (P32121) from the visual arrestin in PDBID: 4ZWJ [255]. All modelling was performed using MODELLER [158]. For every protein, 100 homology models and the best one was selected considering MODELLER's molpdf and DOPE scores, and, for the dopamine receptors, the distance between ICL3 and the intramembrane domain of the dopamine receptor.

ii) Structure refinement with HADDOCK

Dopamine receptors and G-proteins and arrestins were oriented using 3SN6 [217] and 4ZWJ [255], respectively using the alignment function in PyMol [196]. Upon doing so, structures were submitted to the Refinement Interface of the HADDOCK webserver [206]. This interface refines the structure of a complex submitted by the user, with focus on the interface. It simulates a thin water layer surrounding the protein with a cut-off distance preventing water molecule infiltration into the complex. To do so, it follows a three step process: i) a heating phase, during which the structure has no positional restraints (the atoms are allowed to move freely), ii) a high temperature phase, during which bond and improper torsion angle energy constants are reduced by a constant value to ensure structural flexibility, and iii) a cooling phase, during which the constants are progressively returned to their original values. During these steps, MD simulations are responsible for structural alterations in non-bonded interactions, such as Van der Waals, electrostatic and Lennard-Jones interactions.

iii) Interhelical distance

TM3-TM5 and TM5-TM6 interhelical distances were calculated using VMD [195] using the α-carbons belonging to the residues in Table 2. These residues were chosen based on Kruse *et al.* [256], corresponding to 3.54, 5.62 and 6.37 in the Ballesteros-Weinstein nomenclature [215].

Table 2 - Residues utilized for the interhelical distance calculation in TM3, TM5 and TM6 for D1R, D2R, D3R, D4R and D5R.

	TM3	**TM5**	**TM6**
D1R	108ILE	201TYR	241LEU
D2R	108VAL	185TYR	225LEU
D3R	108VAL	188TYR	228VAL
D4R	108PRO	186PHE	226LEU
D5R	108ILE	215TYR	255LEU

iv) Comparative normal mode analysis

Normal mode analysis (NMA) is a technique which aims at tracing the protein’s most relevant movements by considering the protein as a system of harmonic oscillators. This enables the calculation of the normal (orthogonal) modes of vibration – all the harmonic oscillators vibrate at the same frequency. Considering that the most relevant modes are usually the ones comprehending the largest movement, the modes with the lowest vibrational frequency are usually considered to be the most relevant [257].

Comparative NMA was performed with the *bio3d* [248] for the R programming language. To do so, protein structure and sequence are both aligned using Multiple Sequence Comparison by Log-Expectation (MUSCLE), a program used to align protein sequences. After doing so, normal mode analysis is carried out for all proteins and fluctuation values are registered for all residues. Fluctuation values correspond to the vibrational amplitude of each residue. Thanks to the protein sequence alignment, the output is highly informative since it allows the user to compare how residue movement changes for G-proteins, dopamine receptors and arrestins upon binding to different partners.

v) Data visualization and webserver implementation

Data visualization plots using ggplot2 [244] were calculated and displayed using the R programing language [233] and are presented in http://45.32.153.74/gpcr/, which was constructed using the *shiny* application package [246].

C. Results and Discussion

This section is organized according to the methodology section. As such, it comprises results regarding the SpotOn prediction method, the global assessment of PPIs SpotOnDB, the method for interface prediction using DL and the computational case study on protein structure dynamics using dopamine receptors.

1) SpotOn – from high dimensionality to a successful method

i) ML Algorithms Clustering

The various trained machine-learning algorithms were subjected to hierarchical clustering using the Jaccard similarity coefficient as a metric and presented at Moreira *et al.* [199]. The dendrogram depicted in Figure 4, allows us to distinguish 5 main algorithm clusters:

I) Cluster I (mostly RF-based models): avNNet, Boruta, ranger, rf, RRF, RRFglobal and wrsf;

II) Cluster II (mostly adaptive algorithms, bagging algorithms and decision trees/RFs): ada, adaboost, bagEarth, bagEarthGCV, bagFDA, bagFDAGCV, C5.0, C5.0Rules, C5.0Tree, ctree, ctree2, evtree, fda, gamboost, LogitBoost, ORFlog, ORFpls, ORFridge and ORFsvm;

III) Cluster III (mostly regression models): glmboost, multinom, glm and plr.

IV) Cluster IV (mostly SVMs and distance weighted algorithms): dwdPoly, dwdRadial, svmLinear, svmLinear2, svmPoly, svmRadial, svmRadialCost, svmRadialSigma and svmRadialWeights;

V) Cluster V (mostly discriminant analysis algorithms): amdai, hdda, knn, lda, lda2, loclda, nb, pda, qda, rda, stepLDA and stepQDA.

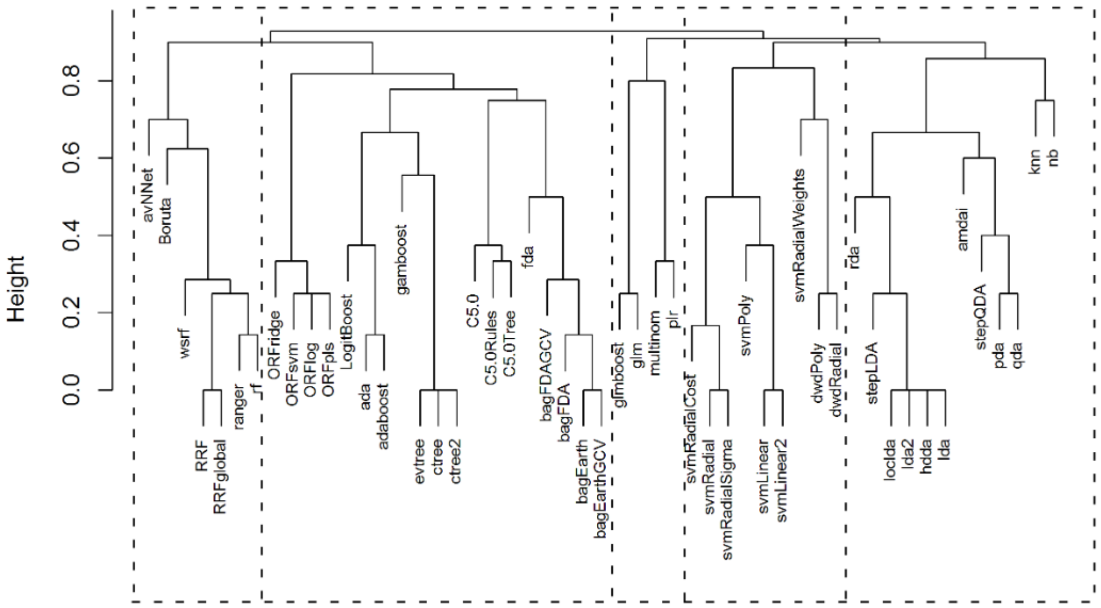


Figure 4 - Hierarchical clustering of the test machine-learning algorithms.

ii) ML algorithms Cluster Performance

The performance of the machine-learning algorithm clustering was assessed using Multivariate Analysis of Variance (MANOVA) described at Moreira *et al.* [199]. Table 3 summarizes the performance on the independent test set by presenting the mean values for each metric for the best classifier of each cluster for the different pre-processing conditions. From the various pre-processed datasets described above, the ScaledUp (dataset generated upon centering and scaling of variables and up-sampling of the minor class) was subsequently used since it yielded the best performance metrics, specifically the best mean value for AUROC and TPR (Sensitivity) in the training set.

Table 3 - Statistical metrics mean values attained for each cluster for all pre-processing conditions for both training set (Train) and testing set (Test).

	Train	Test	Train	Test
	PCA		*Scaled*	
AUROC	0.79	0.67	0.80	0.77
Accuracy	0.89	0.78	0.90	0.81
Sensitivity	0.60	0.31	0.67	0.40
Specificity	0.98	0.92	0.97	0.94
PPV	0.87	0.53	0.88	0.67
NPV	0.89	0.81	0.91	0.83
F1-score	0.67	0.38	0.75	0.49
MCC	0.68	0.29	0.71	0.42
	PCAUp		*ScaledUp*	
AUROC	0.93	0.80	0.94	0.83
Accuracy	0.93	0.79	0.97	0.79
Sensitivity	0.95	0.55	0.98	0.48
Specificity	0.93	0.86	0.96	0.88
PPV	0.93	0.57	0.96	0.57
NPV	0.94	0.87	0.98	0.85
F1-score	0.94	0.55	0.97	0.52
MCC	0.83	0.41	0.91	0.38
	PCADown		*ScaledDown*	
AUROC	0.79	0.70	0.81	0.74
Accuracy	0.91	0.75	0.90	0.76
Sensitivity	0.90	0.78	0.87	0.66
Specificity	0.92	0.74	0.93	0.80
PPV	0.92	0.48	0.92	0.51
NPV	0.91	0.92	0.89	0.88
F1-score	0.91	0.59	0.89	0.57
MCC	0.78	0.46	0.78	0.42

Ensembles of ML algorithms have shown to be quite valuable in improving classification when constructing ML models [236]. The best algorithms of each cluster for the ScaledUp pre-processing condition (ORFsvm, pda, rf, svmPoly and plr) were used as input for a logistic regression model. ORFsvm is an oblique rf – a rf composed of oblique decision trees, which differ from regular trees by taking as input linear combinations of features instead of a single feature – pda is a penalized discriminant analysis – a form of discriminant analysis adapted to high-dimensionality datasets – rf is a standard rf, svmPoly is a polynomial kernel SVM – a form of SVM that represents the space with polynomials of the original variables instead of using the variables and plr is a penalized logistic regression – a regular logistic regression with L1 and L2 regularization. A stepwise selection of relevant variables (algorithms) was performed, leading to the selection of rf, svmPoly and pda as the most relevant classifications for the logistic regression model. Training and testing metrics are provided in Table 4. Logistic regression leads to improved results as reported by all metrics, for both the full (5 variable) and rf + svmPoly + pda regression models. Even though both share practically identical metrics, the latter was chosen as the final model, since it offers the best possible predictions in the least time and simplest way when compared with the Full Regression model. The logistic regression was trained using the full dataset as it comprehends the largest amount of nonredundant protein structures and interfaces with HS information.

Table 4 - Statistical metrics for the best algorithm of each cluster of method and their combined regression model, both the "Full Regression" and the stepwise-optimized regression model (rf + svmPoly + pda) for both training and testing set.

	C5.0		**pda**		**plr**		**rf**		**svmPoly**		**Full Regression**		**rf + svmPoly + pda**	
	Train	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
AUROC	0.88	0.83	0.85	0.84	0.83	0.85	0.93	0.83	0.89	0.83	0.91	0.91	0.91	0.91
Accuracy	0.88	0.91	0.85	0.88	0.83	0.85	0.93	0.90	0.89	0.90	0.94	0.95	0.94	0.95
Sensitivity	0.78	0.68	0.86	0.76	0.82	0.84	0.87	0.71	0.80	0.68	0.98	0.98	0.98	0.98
Specificity	0.98	0.98	0.84	0.91	0.85	0.85	0.98	0.96	0.98	0.97	0.84	0.85	0.84	0.85
PPV	0.98	0.90	0.84	0.73	0.84	0.64	0.98	0.84	0.97	0.87	0.95	0.95	0.95	0.95
NPV	0.81	0.91	0.85	0.93	0.82	0.95	0.89	0.91	0.83	0.91	0.91	0.94	0.91	0.94
FPR	0.22	0.32	0.14	0.24	0.18	0.16	0.13	0.29	0.20	0.32	0.02	0.02	0.02	0.02
FNR	0.02	0.02	0.16	0.09	0.15	0.15	0.02	0.04	0.02	0.03	0.16	0.15	0.16	0.15
F1	0.86	0.78	0.85	0.74	0.83	0.73	0.92	0.77	0.88	0.76	0.96	0.97	0.96	0.97

In order to further assess the quality of the method, SpotOn was compared with other methods commonly used to perform HS prediction, namely SBHD2 (SASA-Based Hot-spot Detection)[258] (a previous version of the algorithm considering only SASA-related features), Robetta[259], K-FADE and K-CON models (KFC2-A and KFC2-B)[260], and CPORT (Consensus Prediction Of interface Residues in Transient complexes)[192], even though the latter is not a proper HS predictor but rather an interface predictor. All predictions were collected by using the respective web-servers. The performance of all tested methods is summarized in Table 5. The full dataset was used for the comparison since it is the richest nonredundant database of proteins with resolved structure and information on HS. SpotOn clearly outperforms all other methods, with a strong performance in identifying both HS and NS.

Table 5 - Comparison of the performance of SpotOn with other common methods used for HS prediction for the full dataset.

	SpotOn	**SBHD2[261]**	**Robetta[259]**	**KFC2-A[260]**	**KFC2-B**	**CPORT[192]**
AUROC	0.91	0.69	0.62	0.66	0.67	0.54
Sensitivity	0.98	0.70	0.29	0.53	0.28	0.54
Specificity	0.84	0.71	0.88	0.81	0.96	0.47
F1-score	0.96	0.62	0.39	0.56	0.42	0.42

2) SpotOnDB – a high-throughput method unravels new details on protein-protein interactions

SpotOnDB proved to be quite successful in understanding PPIs by both confirming previous data high-throughput data and redefining some information on HS that was gathered in studies with low amounts of data by utilizing SpotOn, the method and pipeline developed and described in the previous section. All data is available on a webserver – SpotOnDB – in 45.32.153.74/spotondb.

i) Protein-protein interface characterization

The PPIs in the dataset have between 13 and 156 residues with SASA between 458 and 7229 Å^2. On average, the number of residues in the interface is 47.5 (sd = 24). The amino-acid residues propensities at PPIs were calculated from the relative frequency of the different types of residues in PPIs. HS and interfacial composition on the studied complexes were presented in the webserver under the tab "HS and Interface" as well as in Figure 7A. The PPI frequencies are in close agreement with previously published studies on heterodimeric [262,263] and antibody-antigen interfaces (138 at this dataset), as well as with larger studies with over 6.000 complexes [240], which are usually less hydrophobic than homodimers. Normalized propensities of individual residues in Figure 5 show that the residues which are most likely to be represented in the protein interface when compared to the remainder of the protein are valine, lysine, leucine and tyrosine. Concerning the actual presence residues in the interface, Figure 7A shows that interfaces are enriched in glycine, leucine, serine, threonine, tyrosine and charged residues (particularly arginine), and depleted in tryptophan, cysteine, histidine.

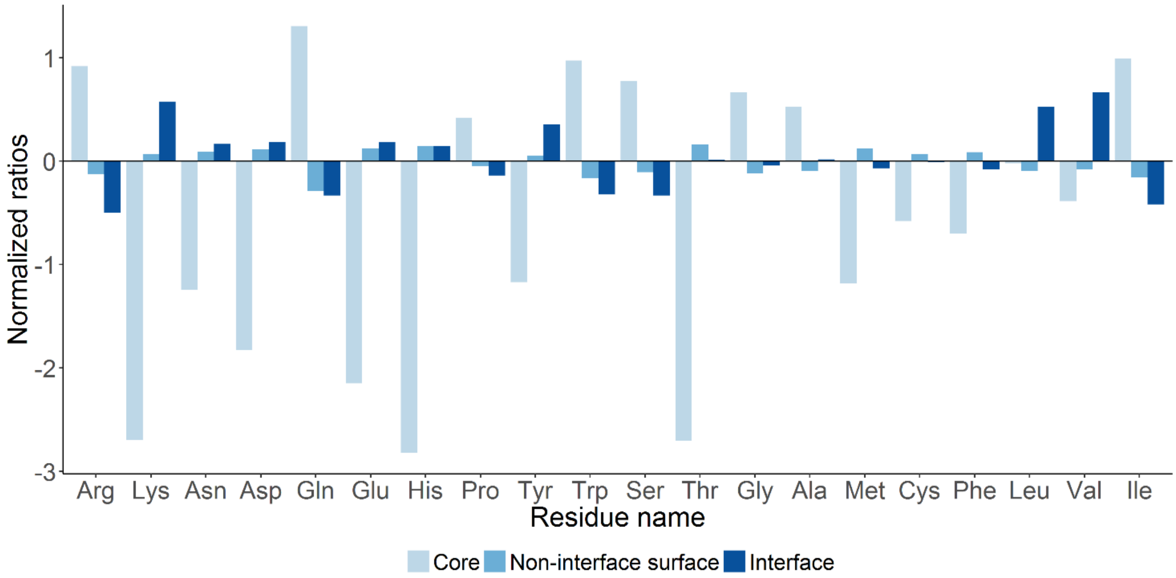


Figure 5 - Residue normalized propensity for the core, non-interface surface and interface protein regions.

ii) Hot-spots Composition

To better understand if the ML approach implemented in SpotOn would allow us to correctly classify interfacial residues as HS and NS, the algorithm specificity, sensibility and accuracy by amino-acid residue type was calculated and plotted in Figure 6. The performance of the method across all residue types is very high with accuracy values ranging from 88% to 100%.

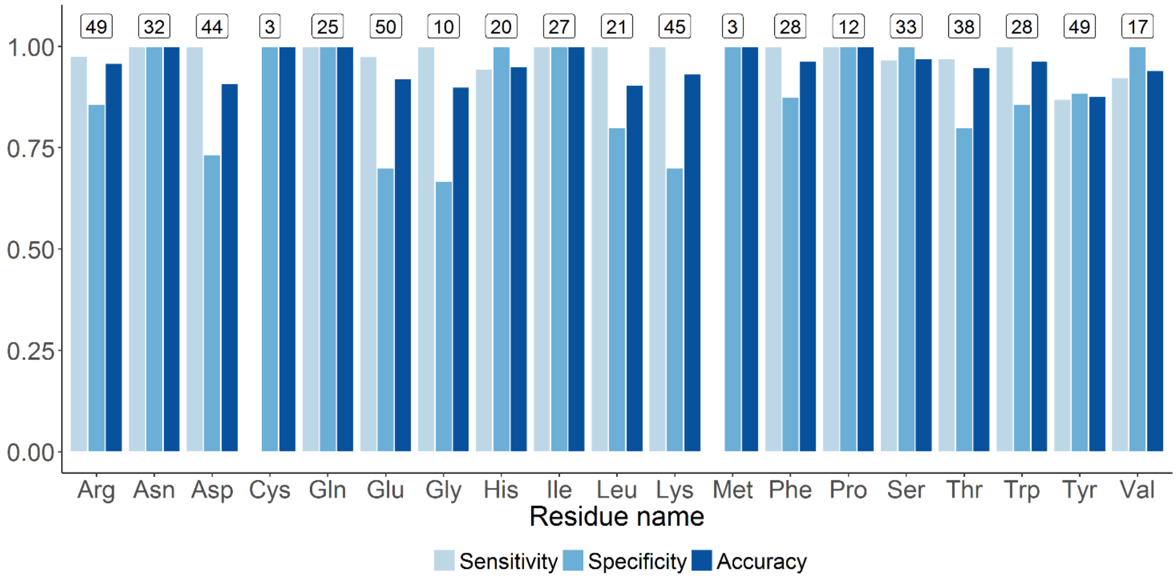


Figure 6 - Performance of SpotOn algorithm by amino-acid type. Sensitivity, specificity and accuracy are reported as well as the number of each amino-acid residue type in small boxes at the top. All values were calculated using the prediction outcomes from SpotOnDB – true positive (TP, correctly predicted HS), true negative (TN, correctly predicted NS), false positive (FP, NS predicted as HS) and false negative (FN, HS predicted as NS). The sensitivity was calculated

by dividing TP cases by the total number of predicted positives (Sensitivity = TP/(TP + FN)), the specificity by dividing TN cases by the total number of predicted negatives (TN/(TN + FP)) and the accuracy was calculated by dividing the total number of correct predictions by the total number of predictions ((TP + TN)/(TP + TN + FP + FN)).

The average HS number detected by complex was 2.6 (5.4% of all interfacial residues) and 45.0 for NS (93.6 % of interfacial residues). Bogan *et al.,* also reported a 5.4% of HS within 2325 interfacial residues analyses [198]. However, their average number of HS was 5.7 by complex (125 HS in 22 complexes), which is almost the double compared to the one attained in this study. Supported by large scale data (1403 complexes), HS are less abundant in protein-protein interface than what was initially thought. By analysing $E.Factor$ values (Figure 7B and Table 6), tryptophan comes across as the most common residue as HS ($E.Factor$ = 2.07), followed by tyrosine ($E.Factor$ = 2.00). Previous studies on HS have reported both these amino-acids as some of the most represented residues (tryptophan $E.Factor$ calculated as 3.91 and tyrosine as 2.29 in a 22 complexes study), which further validates these conclusions [198,264]. However, those studies reported tryptophan as having a much larger $E.Factor$ when compared to what was found ($E.Factor$ calculated as 3.91 vs. 2.01, respectively). This can be due to the low amount of data on tryptophan presented by Bogan *et al.*, with only 19 instances of alanine scanning mutation results for tryptophan being studied, 4 of which were HS (21.05%), as compared with, for example, 218 instances for arginine, of which 29 were HS (13.30%), and 122 for tyrosine, of which 15 were HS (12.30%) [198]. Concerning the high discrepancy of instances for each amino-acid residue, it becomes clear that tryptophan might be overrepresented as a HS due to the low amount of experimental information on its role as HS. Underrepresented residues include cysteine and methionine, widely regarded as low occurring as HS [198,264]. Authors tend to group residues with similar chemical proprieties to improve statistics due to the reduced number of experimental HS/NS residues known but this could lead to error as, for example, HS are depleted in lysine but not in arginine residues ($E.Factor$ = 1.00 vs 1.38, respectively). These results are also listed under the “HS and interface composition” tab.

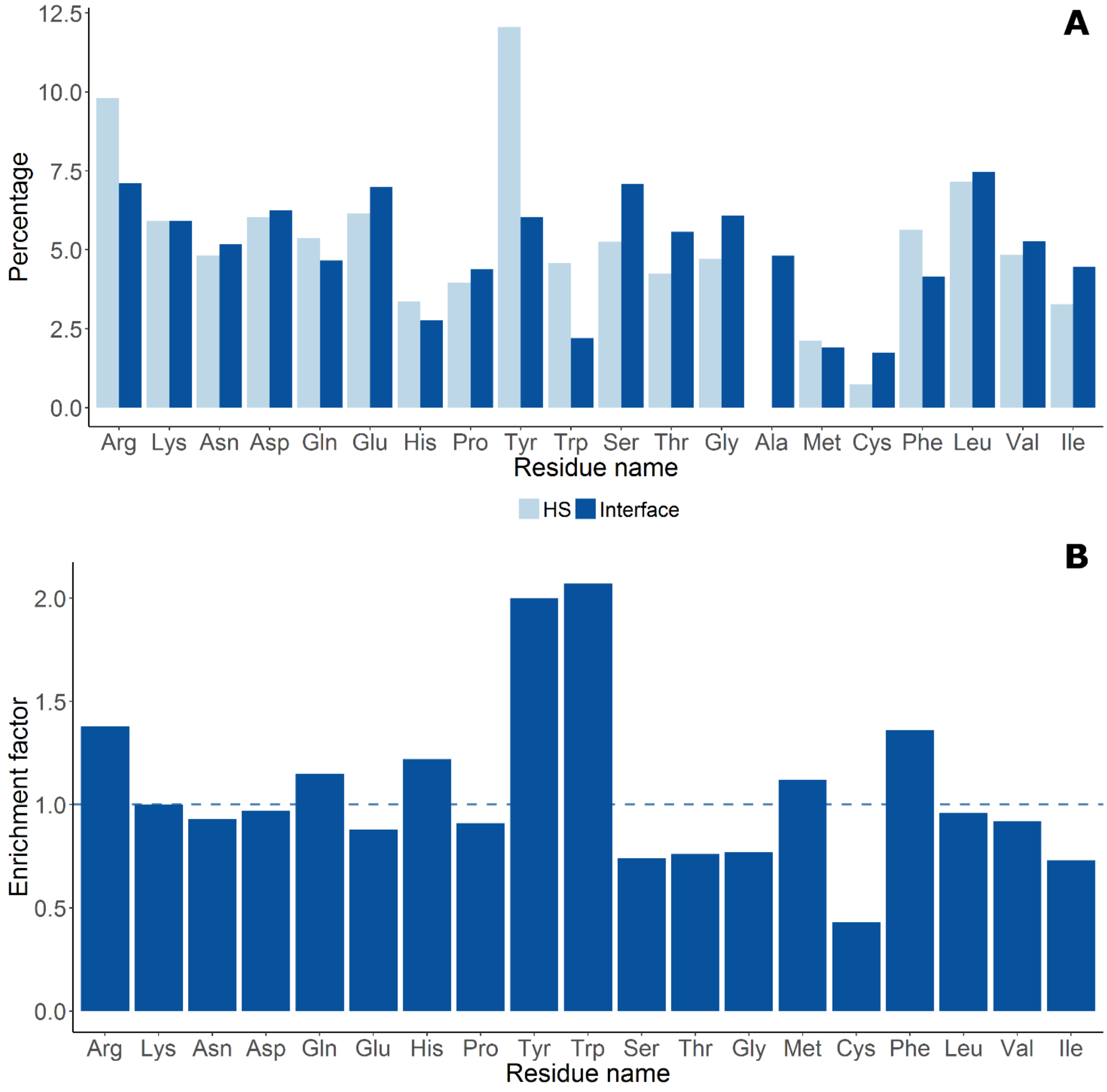


Figure 7 - HS and residues in the protein interface. A. Interface and HS composition by amino-acid residue type. B. HS enrichment by amino-acid residue type.

Table 6 - Interface and HS composition as well as HS enrichment for each amino-acid residue type. Interface (%) represents the percentage of each residue in the interface by dividing the number of each interfacial residue by the total number of interfacial residues; HS (%) represents the percentage of each residue acting as HS when compared to the total interface by dividing the number of each HS residue by the total number of interfacial residues; HS distribution (%) represents the percentage of each residue acting as HS when compared to the total number of HS in the interface by dividing the number of each HS residue by the total number of residues classified as HS; HS Enrichment represents how each residue is represented as a HS when compared to the average HS distribution as described in the Methods section; Residue-wise HS (%) represents the amount of each interfacial residue classified as a HS when compared to the presence of that residue in the interface.

Residue	Interface (%)	HS (%)	HS distribution (%)	HS Enrichment	Residue-wise HS (%)
Ala	4.81	-	-	-	-
Arg	7.10	0.53	9.80	1.38	10.67
Asn	5.18	0.26	4.82	0.93	5.25
Asp	6.25	0.33	6.03	0.97	6.57
Cys	1.74	0.04	0.74	0.43	0.81
Gln	6.98	0.29	5.37	1.15	5.85
Glu	4.66	0.33	6.14	0.88	6.69
Gly	6.08	0.26	4.71	0.77	5.13
His	2.76	0.18	3.36	1.22	3.66
Ile	4.46	0.18	3.28	0.73	3.57
Leu	7.47	0.39	6.94	0.96	7.79
Lys	5.91	0.32	7.16	1.00	6.45
Met	1.90	0.12	2.12	1.12	2.31
Phe	4.15	0.31	5.64	1.36	6.15
Pro	4.38	0.22	3.96	0.91	4.32
Ser	7.09	0.29	5.26	0.74	5.73
Thr	5.57	0.23	4.24	0.76	4.62
Trp	2.20	0.25	4.57	2.07	4.98
Tyr	6.04	0.66	12.06	2.00	13.13
Val	5.28	0.26	4.84	0.92	5.28

iii) Hot-regions

Li *et al.* [265] evaluated for 18 complexes the composition of complemented pockets, not necessary Hot-regions, and concluded that tryptophan, glycine, proline, cysteine, tyrosine and glutamate were more likely to be conserved at these regions. Intramolecular and intermolecular packing around HS and NS was analysed using 2 different Cα-Cα distance cut-offs: 8 and 10 Å. With more than 60.000 interfacial residues classified as HS/NS in the database, there is not a meaningful difference of intramolecular packing between the two classification groups as on average 8.10 neighbour residues are found for HS (sd = 0.53) versus 8.83 for NS (sd = 0.59), using an 8 Å cut-off distance (Table 7). However, the situation completely changes when analysing intermolecular packing. On average, 2.34 (sd = 0.58) and 1.46 (sd = 0.36) residues surround HS and NS, respectively. Table 7 also shows the rather high sd values (around 25%) for the number of both intra- and intermolecular neighbouring residues, illustrating the high variability that can be found in the size of the hot-region, which was not previously found by other authors on previous studies [201,266].

Further observation of these results is available in the SpotOnDB under the “HS regions” tab. Cysteine, glycine, serine, threonine and proline were the residues with a higher number of neighbours when acting as HS. Excluding threonine, these are mostly small residues and present shorter average rays – cysteine and glycine have two of the smallest average rays [267] – which should explain their high above average number of neighbouring residues both as HS and NS.

Table 7 - Average number of neighbouring residues (Mean) and standard-deviation (sd) for intramonomer and intermonomer HS and HS.

	Neighbouring residues	
	Mean	*sd*
Intramonomer HS	8.10	0.53
Intramonomer NS	8.79	0.59
Intermonomer HS	2.34	0.58
Intermonomer NS	1.46	0.36

On average, when compared to NS, HS have nearly three times as much HS as neighbours (0.97 neighbouring HS for HS, when compared to 0.33 for NS (sd = 1.19 and sd = 0.69, respectively) as observable in Figure 8). When combined with the low prevalence of HS in the protein-protein interface, these findings are key in confirming the existence of the already reported cooperative hot-regions [201].

-

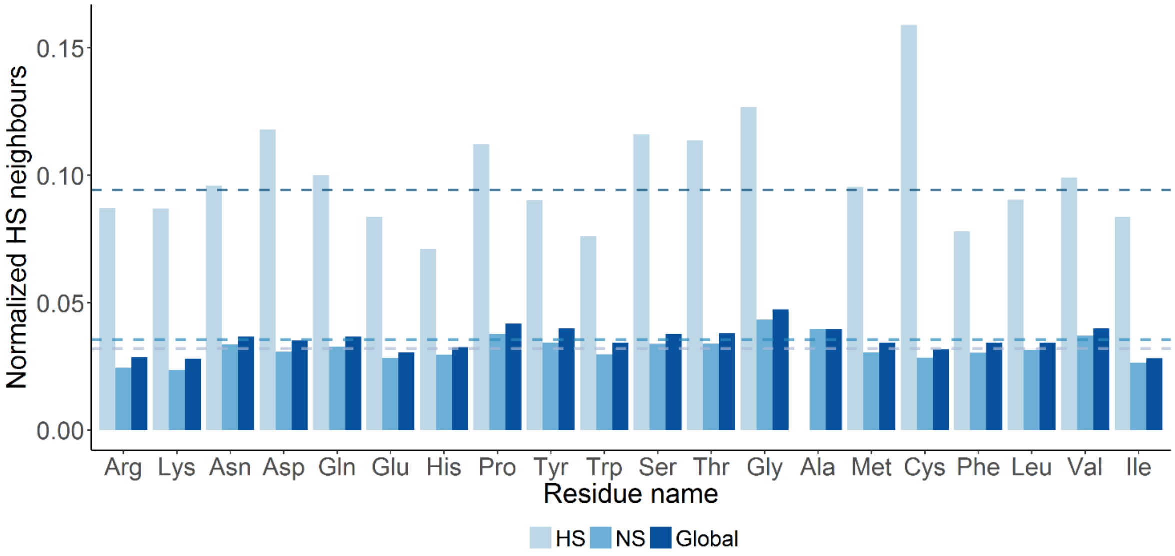


Figure 8 - Hot-regions by residue type. The number of HS is plotted against the residue type with average HS, NS and global (all interfacial residues) values represented by dashed lines Normalizing was done by dividing the average number of HS neighbours by the number of neighbours.

iv) B-factors

The reported data on mobility of HS vs NS comes from a study on conserved vs non-conserved residues by Erdemli *et al.* [268] upon MD simulation of 17 protein-protein complexes. These authors concluded at the time, by analysis of Root-Mean-Square-Deviation (RMSD) of the two regions, that conserved ones' show lower mobility. Whereas this result cannot be directly compared to our, as HS classification is not directly linked to conservation, it can be used as a guideline of the current knowledge in the field. Here, normalized B-factors for all atoms from their crystal structures were gathered and their averages over all atoms, solely for backbone or for side-chains of the individual residues were calculated. The illustrative plots can be found at the SpotOnDB webserver, under the "B-factors" tab. No significant difference between HS and NS backbone B-factor values were found. There is, however, a slight decrease in the HS side-chain B-factors values within the SD. These results point to a similar mobility between HS and any other interfacial residue.

v) Evolutionary conservation

Evolutionary conservation has been considered an important feature when studying protein structure as it explains the prevalence of some residues in key functional sites [269], namely in PPIs [268,270]. Normalized conservation scores (Cons) were calculated for all interfacial residues. A positive value means that the residue is conserved at PPI whereas a negatives value means it is not. A total of 135 monomers (6.5%) did not yield any conservation scores due to the low availability of sequence homologues. Figure 9 shows that HS are more conserved than other residues (Average $Cons_{HS}$ = 0.17, sd = 1.29; Average $Cons_{NS}$ = -0.01, sd = 1.09). This suggests that HS occurrence in PPIs is evolution-driven, as previously suggested [264,265,271].

The most conserved residues in the interface are tyrosine, tryptophan and asparagine, while cysteine appears to be, by far, the least conserved residue as HS ($Cons_{Tyr}$ = 0.33, sd = 1.42; $Cons_{Trp}$ = 0.12, sd = 1.54; $Cons_{Asn}$ = 0.10, sd = 1.20; $Cons_{Cys}$ = -0.76, sd = 0.66). Tryptophan and tyrosine roles was already reported by other authors [201,271]. Phenylalanine and arginine, on the other hand, have been reported as conserved by the same studies [201,271] whereas the data here presented shows low conservation score for both. Arginine conservation was already challenged as it was shown to be the second least common residue in key positions of ligand binding [272]. Leucine seems to be particularly conserved within HS residues. These results are present in the SpotOnDB server under the "Evolutionary Conservation" tab.

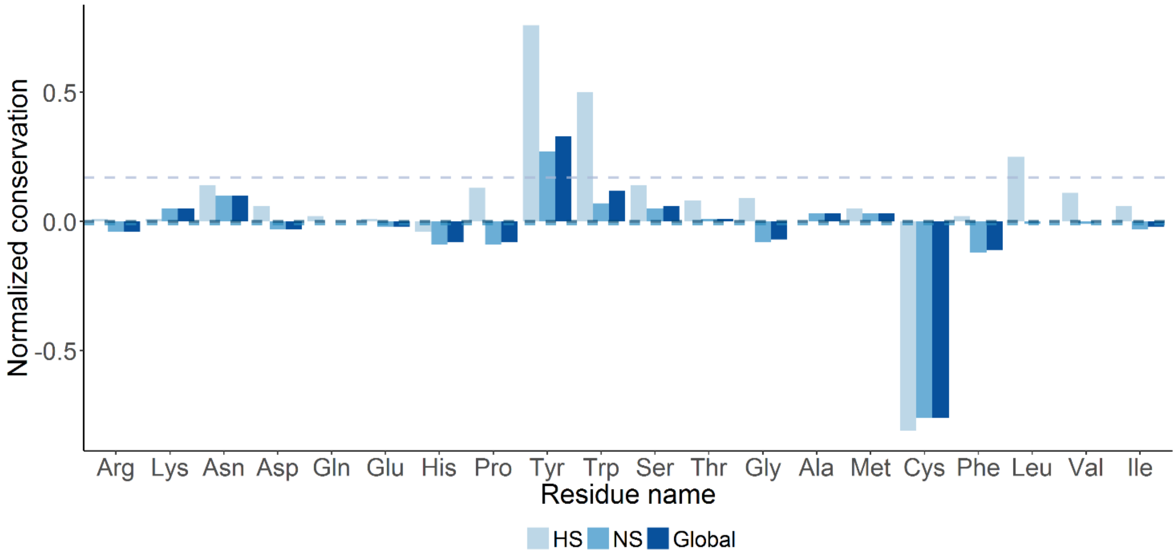


Figure 9 - Average conservation by residue type. The conservation score from Consurf is plotted against the residue type with average HS, NS and global (all interfacial residues) values represented by dashed lines.

vi) Solvent-accessible surface area of interfacial residues

Solvent occlusion had already been demonstrated as a key aspect of HS by the O-ring theory established by Bogan *et al.* [268]. Work done by Moreira *et al.* [261,273-275] in a large number of complexes and taking also into account dynamics supports this theory that SASA is considerably diminished upon complex formation in HS when compared with other interfacial residues. According to these results, average $_{com}SASA_i$ for HS are nearly 7 times lower than those for NS. Furthermore, $_{rel}SASA_i$, which provides a way of directly comparing bound and unbound SASA values and differentiate situations that lead to same $\Delta SASA_i$ but in which the residue is full occluded upon complexation (check Martins *et al.* [261]), is nearly 1.6 times larger in HS and NS (Figure 10). Further statistics on SASA values can be visualized in the SpotOnDB webserver, under the “SASA” tab.

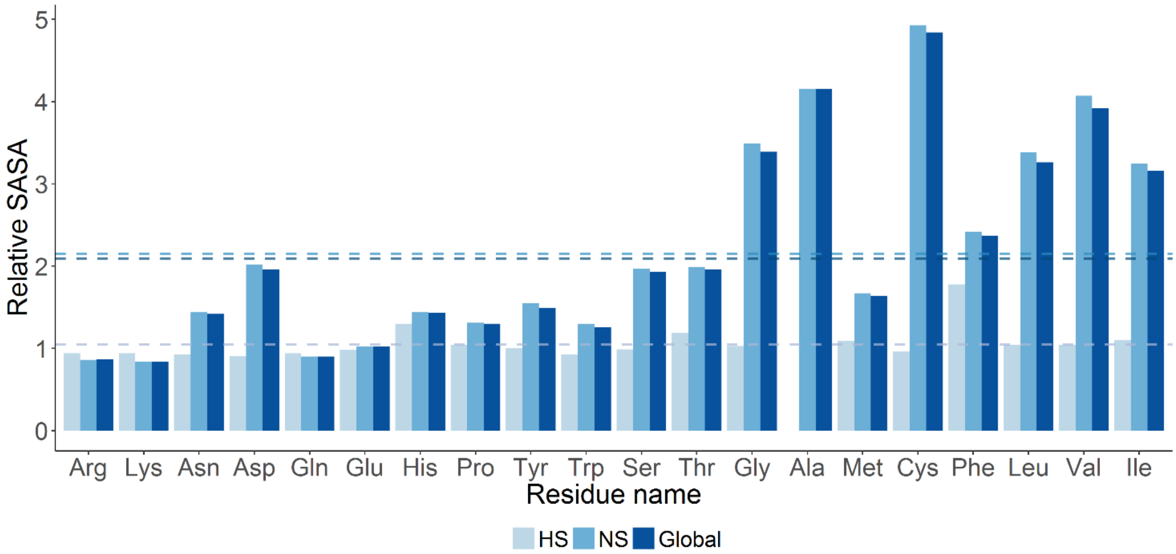


Figure 10 - Relative solvent-accessible surface area values. A. relSASAi average values by interfacial residue. Dashed lines represent the average values for HS, NS and global (all interfacial residues).

vii) Structural interactions at protein-protein interfaces

The number of atoms in short distances (2.5 or 4 Å cut-off) of HS seems to be higher than for NS, especially for positively charged and aromatic residues. Further information on atom abundance within both cut-offs can be found on the SpotOnDB web platform, under the “Structural Information” tab. Hydrophobic interactions appear to be key elements as well in HS, which present on average 2.19 times more hydrophobic interactions than NS, confirming findings by Liu *et al.* [241]. While a study reports H-bonds and salt bridges as irrelevant in defining HS [201], the majority of previous studies reported polar residues as more important either for HS or as highly conserved and important interfacial residues [271,276-279]. Here, both salt-bridges in charged residues and HB across all residues are 1.29 and 1.84 times more common, respectively.

These results contrast with those by Keskin *et al*. using a less accurate classification algorithm, that electrostatics, although relevant for PPIs, is not as relevant for HS formation as H-bonds and hydrophobic interactions [201]. To facilitate visualization and comparison of the presented data on H-bonds, salt-bridges and hydrophobic interactions, Figure 11 shows the ratio of the average number of interactions between HS and NS.

Concerning the distribution of the number of structural interactions per interface area, these results fall closely to those already determined, with 0.63 H-bonds for every 100 Å, which roughly translates to approximately 1 H-bond by every 100-200 Å, as reported by Jones & Thornton [277,278] and Janin [270]. Additionally, high correlation between interface area, and H-bonds and hydrophobic interactions (*r = 0.65* and *r = 0.66*, respectively) was also observed as backed by previous studies [237] and contrarily to the low correlation observed between interface area and salt-bridges and HS (*r = 0.45* and *r = 0.20*, respectively). Furthermore, even though H-bonds and salt-bridges are more common around HS, hydrophobic interactions appear to be 4.52 times more common when considering HS. For more information on this topic, please refer to the "Area distribution" tab in the SpotOnDB webserver.

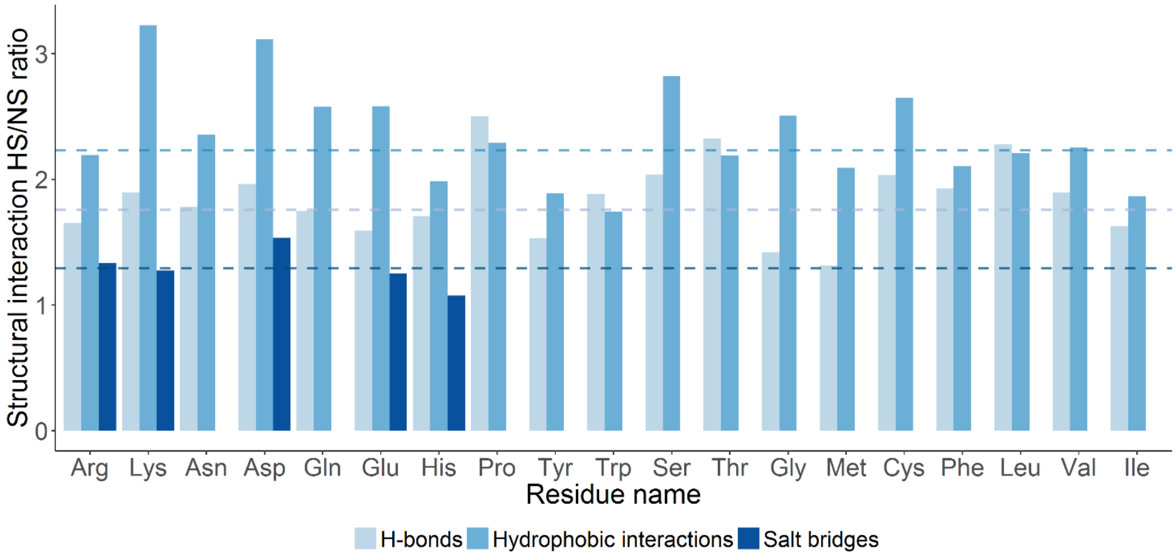


Figure 11 - HS/NS ratio for H-bonds, hydrophobic interactions and salt bridges. Dashed lines represent the average values for H-bonds, hydrophobic interactions and salt bridges.

To further understand contributions by the various structural interactions, Van der Waals and electrostatic energies were retrieved from all HADDOCK [206] experiments run until the moment (roughly 27.000). Interestingly, it was found that minimizing the electrostatic energy is far more important than minimizing other types of energy in docking processes using HADDOCK (in average, electrostatic energy was 3.91 times lower than Van der Waals energy and 16.79 times lower than desolvation energy). Nonetheless, some complementarity might be useful in docking tools – using electrostatic interactions over other types of interactions to determine the best binding pose in rigid docking and then, using sequence-based methods for the detection of important residues and hot-spots to drive flexible docking and interface refinement, obtain a more realistic structure from computational methods.

3) Interface prediction – a tentative application of DL

i) Hyperparametrization and model training

Grid search for the optimal parameters rendered are presented in

Table 8. It is hard to determine the activation function responsible for the best results. A 2013 study using back-propagation neural networks concluded that, while different activation functions rendered different results, variations in error were small when comparing all of them [280]. As such, it becomes different to actually state the reasons behind a better performance by an activation function. The number of hidden layers and the number of neurons in each layer is a complicated matter as well – while some guidelines have been traced (the number of neurons on each hidden layer should be no more than those in the input layer (number of features) and no less than the output layer (number of classes) and the number of hidden layers should be determined regarding the problems complexity [281]) and there have been some efforts to determine the best number of hidden neurons for tasks such as image classification [282], no actual rule of thumb can be found. Even when using web-forums such as Quora (https://www.quora.com), ResearchGate (https://www.researchgate.net/home) and stackoverflow (https://stackoverflow.com), which have a strong community of ML researchers, no absolute answer can be found. This lack of coherence and guidelines was one of the motivations to perform a grid search to determine the best possible hyperparameters. Input dropout is a fairly delicate process – while it prevents the network from overfitting, excessive dropout might cause underfitting by removing too much features [283]. As such, it is understandable why no low values were assumed for Input Dropout Ratio. Similar to this, L1 and L2 regularization are also key in preventing overfitting but overestimation might lead to underfitting. Hence, it is understandable that fairly stable values are observable across all pre-processing conditions for these hyperparameters, with few approaching the stipulated *minima* or *maxima* (excluding Input Dropout Ratio for DownSample and L2 regularization for UpSample) for random discrete grid search.

Table 8 - Optimal hyperparameters according to the random discrete grid search for the deep-learning network.

	Pre-processing conditions			
	Standard	*DownSample*	*UpSample*	*SMOTEd*
Activation functions	Maxout	TanhWithDropout	RectifierWithDropout	Rectifier
Hidden layer composition	[200,200,100]	[500,500,200,100]	[500,500,200,100]	[100,100]
Input dropout ratio	0.05	0.09	0.06	0.08
L1 regularization	$7.40*10^{-4}$	$7.23*10^{-4}$	$3.40*10^{-5}$	$1.71*10^{-4}$
L2 regularization	$1.44*10^{-4}$	$8.20*10^{-4}$	$9.40*10^{-4}$	$1.01*10^{-4}$

The network did not achieve the expected results, as observable in Table 9, while in some cases accuracy achieves good results as the network is able to correctly predict a high number of interfacial and non-interfacial residues correctly, but its PPV is far too high as it predicts many non-interfacial residues as interfacial residues. As such, the network is apparently overfitting in lower PPV cases (DownSample, UpSample and SMOTEd) and underfitting (not able to predict positive cases – interfacial residues in this case) in low-sensitivity cases (Standard). It should also be considered that interfacial residues were determined considering the known interface – the database has no knowledge on other possible interfaces in the protein. As such, FP (residues being predicted as interfacial but are represented as non-interfacial in the database) might in fact be TP because the database does not contain all the information on interfacial residues. Nonetheless, even considering this aspect, there is no practical way of assessing how far this might correct and, as such, while it can be considered as a valid explanation for a few incorrectly predicted interfacial residues (considering that the method is reliable), this factor will remain unaddressed for this case.

Table 9 - Statistical metrics values attained for all pre-processing conditions for both training set (Train) and testing set (Test).

	Train	Test	Train	Test
	Standard		*DownSample*	
AUROC	0.60	0.56	0.55	0.54
Accuracy	0.93	0.92	0.74	0.73
Sensitivity	0.53	0.30	0.91	0.78
Specificity	0.94	0.94	0.73	0.73
PPV	0.22	0.14	0.10	0.09
NPV	0.98	0.98	1.00	0.99
F1-score	0.31	0.19	0.19	0.16
MCC	0.31	0.17	0.25	0.20
	UpSample		*SMOTEd*	
AUROC	0.56	0.55	0.58	0.55
Accuracy	0.78	0.78	0.83	0.83
Sensitivity	0.98	0.84	0.94	0.68
Specificity	0.78	0.78	0.83	0.83
PPV	0.13	0.12	0.16	0.12
NPV	1.00	0.99	1.00	0.99
F1-score	0.23	0.20	0.27	0.21
MCC	0.31	0.26	0.35	0.24

ii) Interfacial patch prediction

The rationale behind this methodology was to: i) predict interfacial residues and ii) confirm the predicted residues by assessing their neighbouring interfacial residues with a logistic regression. However, given that the first step failed to provide good results, using the predictions from this step proved to be ineffective in the second and final step. In fact, predictions from the logistic regression were in fact worse than those coming from the DL network. As such, this approach, while supported by theory, has no practicality with this particular setting.

4) Structural dynamics as influential in understanding GPCR-partner differential binding

The method for interfacial analysis was developed by O. Sensoy, J. Shabbir, José G. Almeida (myself), Irina S. Moreira and G. Morra and has been submitted for publication [284]. The calculations on structural and evolutionary characterization of protein GPCR-partner interface and resulting data treatment were performed by my colleague António José Preto, while I focused mostly on complex structure dynamics using NMA and considering interhelical distances.

i) Normal mode analysis of complex structure

Considering GPCR-arrestin interaction, interface residues appear consistently at the sites belonging to ICL1-3 (information retrieved by my colleague António José Preto). Information on this topic can be visualized on the server developed for this project (http://45.32.153.74/gpcr/), in the “Comparative NMA” section. Concerning the first two ICLs (ICL1 and ICL2), no major differences are observable in residue fluctuation. However, for ICL3, the binding of arrestin affects differently all dopamine receptors. Concerning differences in arrestin residue fluctuation while binding different dopamine receptors, no major fluctuations can be observed. Regarding most G-proteins, GPCRs do not present a high number of interface residues at ICL1, however doing so at ICL2. However, when considering Gs, ICL1 features a higher number of interacting residues. Across all G-protein-GPCR interactions, ICL3 appears to be the most affected region concerning protein motion Further discussing G-proteins, no major interfacial residue fluctuation difference is observed across different GPCR binding interactions. Considering dopamine receptors, only Gs displays relevant fluctuations at residues around 50, which roughly corresponds to ICL1.

NMA results for DR-Arr complexes have shown similar motion for ICL1 and ICL2, as suggested by structural images, while demonstrating higher fluctuation levels for ICL3, showing an important role for this protein region, as demonstrated for other GPCRs [285,286]. When discussing DR-G-protein interaction, ICL1 does not play a major role in the interface for most DR-G-protein complexes, excluding DR-Gs interactions, which concurs with the images in the “Comparative NMA” section of the webserver, where ICL1 is the furthest loop from G-proteins. This might point towards an important role for this region when considering DR-Gs interactions. Furthermore, no major fluctuation differences are observed in G-proteins, while large ones are observed in the ICL3 region of all GPCRs. This might point towards the predominant role of GPCRs – namely the ICL3 region – over G-proteins when determining binding affinity. D4R, which was shown to interact more strongly with arrestins, shows lower fluctuations in the ICL3 when compared to other dopamine receptors. Since lower fluctuations point towards reduced motion of a given protein region, these interactions might be stabilizing the DR-arrestin complexes. Following a similar logic, the opposite for the ICL3 can be concluded, whose interaction, due to its high motility, is a protein region with lower stability when complexed with Arrs.

ii) Interhelical Distance

Interhelical distance is a great tool to distinguish GPCR-G-protein binding from GPCR-arrestin binding – at least one of the measured interhelical distances is much larger in arrestin-bound GPCRs than in G-protein-bound ones. Particularly, all dopamine receptors excluding D2R show larger TM3-TM5 distances when bound to arrestin, while all dopamine receptors excluding D5R show larger TM5-TM6 distances when bound to arrestin. Furthermore, arrestins and G-proteins tend to cluster separately for each dopamine receptor, pointing towards common conformational changes when bound to different partners, implying that these distances and the associated conformational alterations might be evolutionarily conserved.

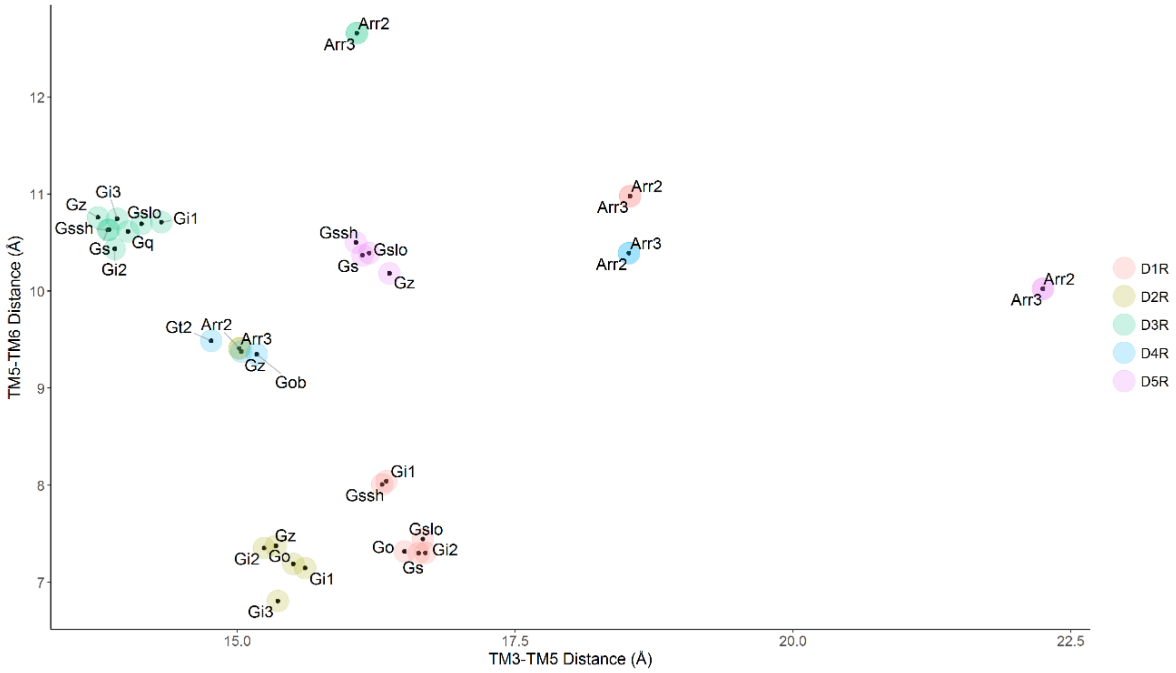


Figure 12 - Interhelical distances plotted according to TM3-TM5 distance and TM5-TM6 distance, both measured in Å.

Even though results from interhelical distance measurements signal that clear and distinct structural changes happen across all dopamine receptors concerning TM movement, these are not as apparently distinguishable when considering fluctuations values from NMA. Nonetheless, it can be considered that fairly different TM movement might be giving rise to distinct structural patterns in important dopamine receptor regions that are highly relevant for GPCR-partner interaction – ICL3 has been described has having several important physiological roles, ranging from receptor activation to endocytosis regulation [285-288]. As such, is no surprise that it may be involved in both arrestin and G-protein binding and that key conformational changes in TMs are leading to this sort of selectivity.

D. Conclusions

During my thesis work I participated in the development of a method to predict HS in PPIs (SpotOn) [199], I performed a high-throughput assessment of protein-protein interfaces and their HS, attempted to implement interfacial residue prediction using DL and performed studies to understand how global protein dynamics affect PPIs. As a result of this work we have successfully published a review paper on membrane protein computational study [160] and a research paper on the SpotOn method [199]. We have also a book chapter describing a computational pipeline for the study of GPCR-partner interactions [284] and a review article on *in silico* methods for drug development in PD-relevant GPCRs[96]. Another paper describing the SOPTONDB is in finalization [289], as well as a paper concerning the differential binding of dopamine receptors to their different partners [290] . Annex III contains a table with all papers and their abstract, as well as the first page of each of the papers.

Results for the SpotOn webserver were relevant as this method outperforms all other computational knowledge-based methods for HS detection. A key aspect of this method is using a non-redundant database (all entries can be considered as valuable when training the algorithm) and all information is derived from experimental studies with structural information – as such, when testing the algorithm, it is possible to say with a high degree of confidence that the performance metrics are valid. By combining a high number of features, an ensemble of ML algorithms and methods to handle overfitting, such as cross-validation and intrinsic algorithm feature selection. ML upon PCA performed poorly when compared to a simple scaling procedure. This is likely to have occurred since PCA uses feature covariance as a way to reduce dimensions and does not consider feature importance (a feature might have lower variance but higher importance, which is detrimental in a PCA). The created pipeline will allow this method to be easily improved when additional experimental information is discovered. Furthermore, cold-spots can also be eventually implemented if their existence is proven and upon existence of sufficient data [291].

SpotOnDB was also successful, resulting in new information of PPIs by combining a rather large database of non-redundant protein complexes [218] with the SpotOn pipeline. The most relevant results include the inter-monomer clustering of residues around HS and the high number of HS neighbours in residues classified as HS, which contributes to the existence of hot-regions [201], and the establishment of an importance order for structural interactions – hydrophobic interactions were more frequent in HS than H-bonds, which were more frequent in HS than salt bridges. All these interactions have been considered as important in interface residues [162,237,238,279,292], but their relative importance had never been assessed in HS, which makes this novel finding highly relevant for the understanding of PPIs and HS. Generally, residues which are more common as HS have typically a higher number of structural features.

Interface prediction was not as successful as it did not performed as expected. The features retrieved are likely to be not as important as initially considered to identify protein interfaces when compared with non-interfacial surface residues. To conclude this, the results from the different pre-processing conditions are highly relevant. DownSample has a lower number of negative class (non-interfacial surface) residues, which addresses the unbalanced dataset at the expense of losing the majority of the non-interfacial surface residues in the dataset. Results ailing from this pre-processing condition could be initially attributed to the loss of negative class entries in the dataset. However, upon considering the results for UpSample – which increases the number of positive class entries by random sampling with replacement – and SMOTEd – which generates synthetic entries through a k-nearest neighbours-based approach – the results do not improve as would be expected. For both cases, if the features are not the most appropriate for the problem, up-sampling is not expected to solve the problem – up-sampling simply repeats underrepresented class entries and SMOTE generates new entries in the dataset based on the feature space (if the features are not adequate for the problem, generating new entries based on those features is likely to have no relevant effect) and is inappropriate for high-dimension datasets [253]. As such, a better search for features should be performed. Future works using DL for interfacial prediction should include the utilization of convolutional networks for sequence feature extraction – which has been done for intraprotein contact prediction using only the protein sequence [293] and using both the protein sequence and evolutionary coupling scores [194] – and also for structural feature extraction – which has been recently used to assess the affinity of protein-ligand interactions, a construction which would be very useful in assessing the reliability of the prediction or in docking problems if adapted to protein-protein interactions [294]. Furthermore, implementing residue propensities such as those in Figure 5 can also be useful as a feature in interface prediction or as a way to understand if the prediction is reliable.

Using protein dynamics to understand how interfaces change was also successful, as results from this study correspond to those obtained by our group considering structural interactions. However, to get a better understanding of protein motion and dynamics it is essential to perform actual MD, as NMA has its limitations [295]. Nonetheless, NMA represents a good exploratory analysis and aids us in understanding protein dynamics when no computational resources are available to perform more demanding techniques such as MD for a high number of structures as the ones tested in this work.

I believe that throughout this thesis work I was able to not only learn many important aspects of bioinformatics – both at a theoretical and practical level – and structural biology, but to contribute to the scientific community with relevant knowledge and methodology for PPIs. Additional effort in the future to complement computational work with experimental work is highly important as well, as it will be crucial in understanding how molecular and structural changes affect cells and organisms.

E. References

1 Nelson, D. L. & Cox, M. M. *Lehninger Principles of Biochemistry*. (W. H. Freeman, 2012).

2 Lamond, A. I. Molecular biology of the cell, 4th edition. *Nature* **417**, 383-383, doi:Doi 10.1038/417383a (2002).

3 Williamson, M. P. & Sutcliffe, M. J. Protein-protein interactions. *Biochemical Society transactions* **38**, 875-878, doi:10.1042/BST0380875 (2010).

4 Barrett, P. J. *et al.* The Quiet Renaissance of Protein NMR. *Biochemistry* **52**, 1303-1320, doi:10.1021/bi4000436 (2013).

5 Wider, G. & Wuthrich, K. NMR spectroscopy of large molecules and multimolecular assemblies in solution. *Curr Opin Struct Biol* **9**, 594-601 (1999).

6 Foster, M. P., McElroy, C. A. & Amero, C. D. Solution NMR of large molecules and assemblies. *Biochemistry* **46**, 331-340, doi:10.1021/bi0621314 (2007).

7 Frueh, D. P. Practical aspects of NMR signal assignment in larger and challenging proteins. *Progress in nuclear magnetic resonance spectroscopy* **78**, 47-75, doi:10.1016/j.pnmrs.2013.12.001 (2014).

8 Ding, X., Zhao, X. & Watts, A. G-protein-coupled receptor structure, ligand binding and activation as studied by solid-state NMR spectroscopy. *The Biochemical journal* **450**, 443-457, doi:10.1042/BJ20121644 (2013).

9 Blakeley, M. P., Hasnain, S. S. & Antonyuk, S. V. Sub-atomic resolution X-ray crystallography and neutron crystallography: promise, challenges and potential. *IUCrJ* **2**, 464-474, doi:10.1107/S2052252515011239 (2015).

10 Li, X. *et al.* Structure of a presenilin family intramembrane aspartate protease. *Nature* **493**, 56-61 (2013).

11 Cherezov, V. *et al.* High Resolution Crystal Structure of an Engineered Human β(2)-Adrenergic G protein-Coupled Receptor. *Science (New York, N.Y.)* **318**, 1258-1265, doi:10.1126/science.1150577 (2007).

12 Van Horn, W. D. *et al.* Solution NMR Structure of Membrane-Integral Diacylglycerol Kinase. *Science (New York, N.Y.)* **324**, 1726-1729, doi:10.1126/science.1171716 (2009).

13 Koehler Leman, J., Ulmschneider, M. B. & Gray, J. J. Computational modeling of membrane proteins. *Proteins* **83**, 1-24, doi:10.1002/prot.24703 (2015).

14 Milne, J. L. *et al.* Cryo-electron microscopy--a primer for the non-microscopist. *The FEBS journal* **280**, 28-45, doi:10.1111/febs.12078 (2013).

15 Liao, M., Cao, E., Julius, D. & Cheng, Y. Structure of the TRPV1 ion channel determined by electron cryo-microscopy. *Nature* **504**, 107-112, doi:10.1038/nature12822 (2013).

16 Park, E., Campbell, E. B. & MacKinnon, R. Structure of a CLC chloride ion channel by cryo-electron microscopy. *Nature* **541**, 500-505, doi:10.1038/nature20812 (2017).

17 Lawson, C. L. *et al.* EMDataBank unified data resource for 3DEM. *Nucleic Acids Res* **44**, D396-403, doi:10.1093/nar/gkv1126 (2016).

18 Doerr, A. Membrane protein structures. *Nat Meth* **6**, 35-35, doi:10.1038/nmeth.f.240 (2009).

19 Moraes, I., Evans, G., Sanchez-Weatherby, J., Newstead, S. & Stewart, P. D. S. Membrane protein structure determination — The next generation. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **1838**, 78-87, doi:http://dx.doi.org/10.1016/j.bbamem.2013.07.010 (2014).

20 Lee, A. G. How lipids affect the activities of integral membrane proteins. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **1666**, 62-87, doi:http://dx.doi.org/10.1016/j.bbamem.2004.05.012 (2004).

21 Grouleff, J., Irudayam, S. J., Skeby, K. K. & Schiøtt, B. The influence of cholesterol on membrane protein structure, function, and dynamics studied by molecular dynamics

simulations. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **1848**, 1783-1795, doi:http://dx.doi.org/10.1016/j.bbamem.2015.03.029 (2015).

22 Grouleff, J., Irudayam, S. J., Skeby, K. K. & Schiott, B. The influence of cholesterol on membrane protein structure, function, and dynamics studied by molecular dynamics simulations. *Biochimica et biophysica acta* **1848**, 1783-1795, doi:10.1016/j.bbamem.2015.03.029 (2015).

23 Alonso, M. A. & Millán, J. The role of lipid rafts in signalling and membrane trafficking in T lymphocytes. *Journal of Cell Science* **114**, 3957-3965 (2001).

24 Brown, D. A. & London, E. Functions of lipid rafts in biological membranes. *Annual review of cell and developmental biology* **14**, 111-136, doi:10.1146/annurev.cellbio.14.1.111 (1998).

25 Escribá, P. V. *et al.* Role of lipid polymorphism in G protein-membrane interactions: nonlamellar-prone phospholipids and peripheral protein binding to membranes. *Proceedings of the National Academy of Sciences* **94**, 11375-11380 (1997).

26 UniProt: a hub for protein information. *Nucleic acids research* **43**, D204-212, doi:10.1093/nar/gku989 (2015).

27 Whisstock, J. C. & Lesk, A. M. Prediction of protein function from protein sequence and structure. *Q Rev Biophys* **36**, 307-340 (2003).

28 Petersen, T. N. *et al.* Prediction of protein secondary structure at 80% accuracy. *Proteins* **41**, 17-20 (2000).

29 Rost, B. & Sander, C. Prediction of protein secondary structure at better than 70% accuracy. *J Mol Biol* **232**, 584-599, doi:10.1006/jmbi.1993.1413 (1993).

30 Kretsinger, R. H., Ison, R. E. & Hovmoller, S. Prediction of protein structure. *Methods Enzymol* **383**, 1-27, doi:10.1016/S0076-6879(04)83001-5 (2004).

31 Shortle, D. Prediction of protein structure. *Current biology : CB* **10**, R49-51 (2000).

32 Argos, P. & Rao, J. K. Prediction of protein structure. *Methods Enzymol* **130**, 185-207 (1986).

33 Edwards, Y. J. & Cottage, A. Prediction of protein structure and function by using bioinformatics. *Methods Mol Biol* **175**, 341-375, doi:10.1385/1-59259-235-X:341 (2001).

34 Nanni, L., Brahnam, S. & Lumini, A. Prediction of protein structure classes by incorporating different protein descriptors into general Chou's pseudo amino acid composition. *Journal of theoretical biology* **360**, 109-116, doi:10.1016/j.jtbi.2014.07.003 (2014).

35 Hartlmuller, C., Gobl, C. & Madl, T. Prediction of Protein Structure Using Surface Accessibility Data. *Angewandte Chemie*, doi:10.1002/anie.201604788 (2016).

36 Al-Lazikani, B., Hill, E. E. & Morea, V. Protein structure prediction. *Methods Mol Biol* **453**, 33-85, doi:10.1007/978-1-60327-429-6_2 (2008).

37 Westhead, D. R. & Thornton, J. M. Protein structure prediction. *Current opinion in biotechnology* **9**, 383-389 (1998).

38 Benner, S. A., Geroff, D. L. & Rozzell, J. D. Protein structure prediction. *Science* **274**, 1448b-1449b, doi:10.1126/science.274.5292.1448b (1996).

39 Barton, G. J. & Russell, R. B. Protein structure prediction. *Nature* **361**, 505-506, doi:10.1038/361505b0 (1993).

40 Robson, B. & Garnier, J. Protein structure prediction. *Nature* **361**, 506, doi:10.1038/361506a0 (1993).

41 Garnier, J. Protein structure prediction. *Biochimie* **72**, 513-524 (1990).

42 Hui, W. Q., Cheng, Q., Liu, T. Y. & Ouyang, Q. Homology modeling, docking, and molecular dynamics simulation of the receptor GALR2 and its interactions with galanin and a positive allosteric modulator. *Journal of molecular modeling* **22**, 90, doi:10.1007/s00894-016-2944-x (2016).

43 Nugent, T. De novo membrane protein structure prediction. *Methods Mol Biol* **1215**, 331-350, doi:10.1007/978-1-4939-1465-4_15 (2015).

44 Nugent, T. & Jones, D. T. Accurate de novo structure prediction of large transmembrane protein domains using fragment-assembly and correlated mutation analysis. *Proc Natl Acad Sci U S A* **109**, E1540-1547, doi:10.1073/pnas.1120036109 (2012).

45 Seaton, B. A. & Roberts, M. F. in *Biological Membranes: A Molecular Perspective from Computation and Experiment* (eds Kenneth M. Merz & Benoît Roux) 355-403 (Birkhäuser Boston, 1996).

46 Whited, A. M. & Johs, A. The interactions of peripheral membrane proteins with biological membranes. *Chemistry and physics of lipids* **192**, 51-59, doi:10.1016/j.chemphyslip.2015.07.015 (2015).

47 Monje-Galvan, V. & Klauda, J. B. Peripheral membrane proteins: Tying the knot between experiment and computation. *Biochim Biophys Acta* **1858**, 1584-1593, doi:10.1016/j.bbamem.2016.02.018 (2016).

48 London, S., Gurdal, O. & Gall, C. Automatic Export of PubMed Citations to EndNote. *Medical reference services quarterly* **29**, 146-153, doi:10.1080/02763861003723317 (2010).

49 Arinaminpathy, Y., Khurana, E., Engelman, D. M. & Gerstein, M. B. Computational analysis of membrane proteins: the largest class of drug targets. *Drug discovery today* **14**, 1130-1135, doi:10.1016/j.drudis.2009.08.006 (2009).

50 Alford, R. F. *et al.* An Integrated Framework Advancing Membrane Protein Modeling and Design. *PLoS Computational Biology* **11**, e1004398, doi:10.1371/journal.pcbi.1004398 (2015).

51 Gromiha, M. M. & Ou, Y. Y. Bioinformatics approaches for functional annotation of membrane proteins. *Brief Bioinform* **15**, 155-168, doi:10.1093/bib/bbt015 (2014).

52 Peyronnet, R., Tran, D., Girault, T. & Frachisse, J. M. Mechanosensitive channels: feeling tension in a world under pressure. *Frontiers in Plant Science* **5**, doi:10.3389/fpls.2014.00558 (2014).

53 Tice, C. M. & Zheng, Y. J. Non-canonical modulators of nuclear receptors. *Bioorganic & medicinal chemistry letters* **26**, 4157-4164, doi:10.1016/j.bmcl.2016.07.067 (2016).

54 Lodish, H. *et al.* Section 3.4-Section 3.4 (W. H. Freeman, 2000).

55 White, S. H. & Wimley, W. C. Membrane protein folding and stability: physical principles. *Annual review of biophysics and biomolecular structure* **28**, 319-365, doi:10.1146/annurev.biophys.28.1.319 (1999).

56 Schulz, G. E. Transmembrane beta-barrel proteins. *Advances in protein chemistry* **63**, 47-70 (2003).

57 Bernaudat, F. *et al.* (2011).

58 Zoonens, M. & Miroux, B. Expression of membrane proteins at the Escherichia coli membrane for structural studies. *Methods in Molecular Biology* **601**, 49-66 (2010).

59 Wagner, S. *et al.* Tuning Escherichia coli for membrane protein overexpression. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 14371-14376, doi:10.1073/pnas.0804090105 (2008).

60 Rosenbusch, J. P., Lustig, A., Grabo, M., Zulauf, M. & Regenass, M. Approaches to determining membrane protein structures to high resolution: do selections of subpopulations occur? *Micron* **32**, 75-90, doi:10.1016/S0968-4328(00)00021-4 (2001).

61 Privé, G. G. Detergents for the stabilization and crystallization of membrane proteins. *Methods* **41**, 388-397, doi:10.1016/j.ymeth.2007.01.007 (2007).

62 Carpenter, E. P., Beis, K., Cameron, A. D. & Iwata, S. Overcoming the challenges of membrane protein crystallography. *Current Opinion in Structural Biology* **18**, 581-586, doi:10.1016/j.sbi.2008.07.001 (2008).

63 Cherezov, V., Peddi, A., Muthusubramaniam, L., Zheng, Y. F. & Caffrey, M. A robotic system for crystallizing membrane and soluble proteins in lipidic mesophases. *Acta Crystallographica Section D* **60**, 1795-1807 (2004).

64 Hunte, C., Koepke, J., Lange, C., Roßmanith, T. & Michel, H. Structure at 2.3 Å resolution of the cytochrome bc1 complex from the yeast Saccharomyces cerevisiae co-crystallized with an antibody Fv fragment. *Structure* **8**, 669-684, doi:10.1016/S0969-2126(00)00152-0 (2000).

65 Liang, B. & Tamm, L. NMR as a tool to investigate the structure, dynamics and function of membrane proteins. *Nat Struct Mol Biol* **23**, 468-474, doi:doi: 10.1038/nsmb.3226 (2016).

66 Oxenoid, K. & Chou, J. J. A functional NMR for membrane proteins: dynamics, ligand binding, and allosteric modulation. *Protein science : a publication of the Protein Society* **25**, 959-973, doi:10.1002/pro.2910 (2016).

67 Murray, D. T., Das, N. & Cross, T. A. Solid State NMR Strategy for Characterizing Native Membrane Protein Structures. *Accounts of Chemical Research* **46**, 2172-2181, doi:10.1021/ar3003442 (2013).

68 Shahid, S. A. *et al.* Membrane-protein structure determination by solid-state NMR spectroscopy of microcrystals. *Nat Meth* **9**, 1212-1217 (2012).

69 Watts, A. *et al.* (ed A. Kristina Downing) 403-473 (Humana Press, 2004).

70 Kaplan, M., Pinto, C., Houben, K. & Baldus, M. Nuclear magnetic resonance (NMR) applied to membrane-protein complexes. *Q Rev Biophys* **49**, e15, doi:10.1017/S003358351600010X (2016).

71 Fredriksson, R., Lagerstrom, M. C., Lundin, L. G. & Schioth, H. B. The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints. *Molecular pharmacology* **63**, 1256-1272, doi:10.1124/mol.63.6.1256 (2003).

72 Heng, B. C., Aubel, D. & Fussenegger, M. An overview of the diverse roles of G-protein coupled receptors (GPCRs) in the pathophysiology of various human diseases. *Biotechnology advances* **31**, 1676-1694, doi:10.1016/j.biotechadv.2013.08.017 (2013).

73 Kobilka, B. K. G protein coupled receptor structure and activation. *Biochim Biophys Acta* **1768**, 794-807, doi:10.1016/j.bbamem.2006.10.021 (2007).

74 Maurice, P. *et al.* GPCR-interacting proteins, major players of GPCR function. *Advances in pharmacology (San Diego, Calif.)* **62**, 349-380, doi:10.1016/b978-0-12-385952-5.00001-4 (2011).

75 Weinstein, L. S., Chen, M. & Liu, J. Gs(alpha) mutations and imprinting defects in human disease. *Annals of the New York Academy of Sciences* **968**, 173-197 (2002).

76 Taussig, R., Iniguez-Lluhi, J. A. & Gilman, A. G. Inhibition of adenylyl cyclase by Gi alpha. *Science* **261**, 218-221 (1993).

77 Berstein, G. *et al.* Phospholipase C-beta 1 is a GTPase-activating protein for Gq/11, its physiologic regulator. *Cell* **70**, 411-418 (1992).

78 Suzuki, N., Hajicek, N. & Kozasa, T. Regulation and physiological functions of G12/13-mediated signaling pathways. *Neuro-Signals* **17**, 55-70, doi:10.1159/000186690 (2009).

79 Akgoz, M., Azpiazu, I., Kalyanaraman, V. & Gautam, N. Role of the G protein gamma subunit in beta gamma complex modulation of phospholipase Cbeta function. *J Biol Chem* **277**, 19573-19578, doi:10.1074/jbc.M201546200 (2002).

80 Reuveny, E. Structural biology: Ion channel twists to open. *Nature* **498**, 182-183, doi:10.1038/nature12255 (2013).

81 Oakley, R. H., Laporte, S. A., Holt, J. A., Barak, L. S. & Caron, M. G. Association of beta-arrestin with G protein-coupled receptors during clathrin-mediated endocytosis dictates the profile of receptor resensitization. *J Biol Chem* **274**, 32248-32257 (1999).

82 Hara, M. R. *et al.* A stress response pathway regulates DNA damage through beta2-adrenoreceptors and beta-arrestin-1. *Nature* **477**, 349-353, doi:10.1038/nature10368 (2011).

83 Boerrigter, G. *et al.* Cardiorenal actions of TRV120027, a novel ss-arrestin-biased ligand at the angiotensin II type I receptor, in healthy and heart failure canines: a novel

therapeutic strategy for acute heart failure. *Circulation. Heart failure* **4**, 770-778, doi:10.1161/CIRCHEARTFAILURE.111.962571 (2011).

84 Whalen, E. J., Rajagopal, S. & Lefkowitz, R. J. Therapeutic potential of beta-arrestin- and G protein-biased agonists. *Trends in molecular medicine* **17**, 126-139, doi:10.1016/j.molmed.2010.11.004 (2011).

85 Sengupta, D. & Chattopadhyay, A. Molecular dynamics simulations of GPCR–cholesterol interaction: An emerging paradigm. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **1848**, 1775-1782, doi:http://dx.doi.org/10.1016/j.bbamem.2015.03.018 (2015).

86 Park, J. H. *et al.* Opsin, a structural model for olfactory receptors? *Angewandte Chemie* **52**, 11021-11024, doi:10.1002/anie.201302374 (2013).

87 Wu, G., Davis, J. E. & Zhang, M. Regulation of alpha2B-Adrenerigc Receptor Export Trafficking by Specific Motifs. *Progress in molecular biology and translational science* **132**, 227-244, doi:10.1016/bs.pmbts.2015.03.004 (2015).

88 Duvernay, M. T. *et al.* A single conserved leucine residue on the first intracellular loop regulates ER export of G protein-coupled receptors. *Traffic* **10**, 552-566, doi:10.1111/j.1600-0854.2009.00890.x (2009).

89 Perez-Aguilar, J. M., Shan, J., LeVine, M. V., Khelashvili, G. & Weinstein, H. A functional selectivity mechanism at the serotonin-2A GPCR involves ligand-dependent conformations of intracellular loop 2. *J Am Chem Soc* **136**, 16044-16054, doi:10.1021/ja508394x (2014).

90 Katritch, V., Cherezov, V. & Stevens, R. C. Diversity and modularity of G protein-coupled receptor structures. *Trends in pharmacological sciences* **33**, 17-27, doi:10.1016/j.tips.2011.09.003 (2012).

91 Verzijl, D. *et al.* Helix 8 of the viral chemokine receptor ORF74 directs chemokine binding. *J Biol Chem* **281**, 35327-35335, doi:10.1074/jbc.M606877200 (2006).

92 Sensoy, O. & Weinstein, H. A mechanistic role of Helix 8 in GPCRs: Computational modeling of the dopamine D2 receptor interaction with the GIPC1-PDZ-domain. *Biochim Biophys Acta* **1848**, 976-983, doi:10.1016/j.bbamem.2014.12.002 (2015).

93 Moreira, I. S. Structural features of the G-protein/GPCR interactions. *Biochimica et Biophysica Acta (BBA) - General Subjects* **1840**, 16-33, doi:http://dx.doi.org/10.1016/j.bbagen.2013.08.027 (2014).

94 Christopoulos, A. Allosteric binding sites on cell-surface receptors: novel targets for drug discovery. *Nat Rev Drug Discov* **1**, 198-210 (2002).

95 Conn, P. J., Christopoulos, A. & Lindsley, C. W. Allosteric modulators of GPCRs: a novel approach for the treatment of CNS disorders. *Nature reviews. Drug discovery* **8**, 41-54, doi:10.1038/nrd2760 (2009).

96 Lemos, A. *et al.* In silico studies targeting G-protein coupled receptors for drug research against Parkinson’s disease. *Current Neuropharmacology* (Submitted).

97 Snyder, S. H., Taylor, K. M., Coyle, J. T. & Meyerhoff, J. L. The role of brain dopamine in behavioral regulation and the actions of psychotropic drugs. *The American journal of psychiatry* **127**, 199-207, doi:10.1176/ajp.127.2.199 (1970).

98 Missale, C., Nash, S. R., Robinson, S. W., Jaber, M. & Caron, M. G. Dopamine receptors: from structure to function. *Physiological reviews* **78**, 189-225 (1998).

99 Iversen, S. D. & Iversen, L. L. Dopamine: 50 years in perspective. *Trends in neurosciences* **30**, 188-193, doi:10.1016/j.tins.2007.03.002 (2007).

100 Scatton, B., Javoy-Agid, F., Rouquier, L., Dubois, B. & Agid, Y. Reduction of cortical dopamine, noradrenaline, serotonin and their metabolites in Parkinson's disease. *Brain research* **275**, 321-328 (1983).

101 Seeman, P. *et al.* Human brain D1 and D2 dopamine receptors in schizophrenia, Alzheimer's, Parkinson's, and Huntington's diseases. *Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology* **1**, 5-15 (1987).

102 Seeman, P., Niznik, H. B., Guan, H. C., Booth, G. & Ulpian, C. Link between D1 and D2 dopamine receptors is reduced in schizophrenia and Huntington diseased brain. *Proc Natl Acad Sci U S A* **86**, 10156-10160 (1989).

103 Arthur L, S. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development* **3**, 210-229 (1959).

104 Drew Conway, J. M. W. *Machine Learning for Hackers*. (1989).

105 Dasgupta, A., Sun, Y. V., Konig, I. R., Bailey-Wilson, J. E. & Malley, J. D. Brief review of regression-based and machine learning methods in genetic epidemiology: the Genetic Analysis Workshop 17 experience. *Genetic epidemiology* **35 Suppl 1**, S5-11, doi:10.1002/gepi.20642 (2011).

106 Lin, H. N., Chen, C. T., Sung, T. Y., Ho, S. Y. & Hsu, W. L. Protein subcellular localization prediction of eukaryotes using a knowledge-based approach. *BMC Bioinformatics* **10 Suppl 15**, S8, doi:10.1186/1471-2105-10-S15-S8 (2009).

107 Peterson, L. E. K-nearest neighbor. *Scholarpedia* **4**, 1883, doi:doi::10.4249/scholarpedia.1883 (2009).

108 Krishna, K. & Murty, M. N. Genetic K-means algorithm. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* **29**, 433-439, doi:10.1109/3477.764879 (1999).

109 Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **48**, 443-453 (1970).

110 Heringa, J. & Taylor, W. R. Three-dimensional domain duplication, swapping and stealing. *Curr Opin Struct Biol* **7**, 416-421 (1997).

111 Katti, M. V., Sami-Subbu, R., Ranjekar, P. K. & Gupta, V. S. Amino acid repeat patterns in protein sequences: their diversity and structural-functional implications. *Protein Sci* **9**, 1203-1209, doi:10.1110/ps.9.6.1203 (2000).

112 Smith, T. F. & Waterman, M. S. Identification of common molecular subsequences. *J Mol Biol* **147**, 195-197 (1981).

113 Waterman, M. S. & Eggert, M. A new algorithm for best subsequence alignments with application to tRNA-rRNA comparisons. *J Mol Biol* **197**, 723-728 (1987).

114 Bailey, T. L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology* **2**, 28-36 (1994).

115 Notredame, C., Higgins, D. G. & Heringa, J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* **302**, 205-217, doi:10.1006/jmbi.2000.4042 (2000).

116 Simossis, V., Kleinjung, J. & Heringa, J. An overview of multiple sequence alignment. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]* **Chapter 3**, Unit 3 7, doi:10.1002/0471250953.bi0307s03 (2003).

117 Apweiler, R. *et al.* UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res* **32**, D115-119, doi:10.1093/nar/gkh131 (2004).

118 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403-410, doi:10.1016/S0022-2836(05)80360-2 (1990).

119 Dayhoff, M. O., Schwartz, R. M. & Orcutt, B. C. A model of evolutionary change in proteins. *Atlas of protein sequence and structure* **5**, 345-351, doi:citeulike-article-id:4442167 (1978).

120 Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* **89**, 10915-10919 (1992).

121 Eddy, S. R. Hidden Markov models. *Curr Opin Struct Biol* **6**, 361-365 (1996).

122 Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics (Oxford, England)* **23**, 2947-2948, doi:10.1093/bioinformatics/btm404 (2007).

123 Huang, X. & Miller, W. A time-efficient, linear-space local similarity algorithm. *Advances in Applied Mathematics* **12**, 337-357, doi:http://dx.doi.org/10.1016/0196-8858(91)90017-D (1991).

124 Ashkenazy, H. *et al.* ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Res* **44**, W344-350, doi:10.1093/nar/gkw408 (2016).

125 Pupko, T., Bell, R. E., Mayrose, I., Glaser, F. & Ben-Tal, N. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics (Oxford, England)* **18 Suppl 1**, S71-77 (2002).

126 Dorn, M., E Silva, M. B., Buriol, L. S. & Lamb, L. C. Three-dimensional protein structure prediction: Methods and computational strategies. *Computational Biology and Chemistry* **53**, 251-276, doi:10.1016/j.compbiolchem.2014.10.001 (2014).

127 Ichiye, T. & Karplus, M. Collective motions in proteins: a covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations. *Proteins* **11**, 205-217, doi:10.1002/prot.340110305 (1991).

128 Amadei, A., Linssen, A. B. & Berendsen, H. J. Essential dynamics of proteins. *Proteins* **17**, 412-425, doi:10.1002/prot.340170408 (1993).

129 Fenwick, R. B., Esteban-Martin, S. & Salvatella, X. Understanding biomolecular motion, recognition, and allostery by use of conformational ensembles. *European biophysics journal : EBJ* **40**, 1339-1355, doi:10.1007/s00249-011-0754-8 (2011).

130 Kim, J. I., Na, S. & Eom, K. Domain decomposition-based structural condensation of large protein structures for understanding their conformational dynamics. *J Comput Chem* **32**, 161-169, doi:10.1002/jcc.21613 (2011).

131 de Oliveira, C. A., Grant, B. J., Zhou, M. & McCammon, J. A. Large-scale conformational changes of Trypanosoma cruzi proline racemase predicted by accelerated molecular dynamics simulation. *PLoS Comput Biol* **7**, e1002178, doi:10.1371/journal.pcbi.1002178 (2011).

132 Wen, P. C. & Tajkhorshid, E. Conformational coupling of the nucleotide-binding and the transmembrane domains in ABC transporters. *Biophys J* **101**, 680-690, doi:10.1016/j.bpj.2011.06.031 (2011).

133 Stolzenberg, S., Khelashvili, G. & Weinstein, H. Structural intermediates in a model of the substrate translocation path of the bacterial glutamate transporter homologue GltPh. *The journal of physical chemistry. B* **116**, 5372-5383, doi:10.1021/jp301726s (2012).

134 van Gunsteren, W. F. & Berendsen, H. J. C. Computer Simulation of Molecular Dynamics: Methodology, Applications, and Perspectives in Chemistry. *Angewandte Chemie International Edition in English* **29**, 992-1023, doi:10.1002/anie.199009921 (1990).

135 Brooks, B. R. *et al.* CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry* **4**, 187-217, doi:10.1002/jcc.540040211 (1983).

136 Pearlman, D. A. *et al.* AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Computer Physics Communications* **91**, 1-41, doi:10.1016/0010-4655(95)00041-D (1995).

137 Berendsen, H. J. C., van der Spoel, D. & van Drunen, R. GROMACS: A message-passing parallel molecular dynamics implementation. *Computer Physics Communications* **91**, 43-56, doi:10.1016/0010-4655(95)00042-E (1995).

138 Hess, B., Kutzner, C., van der Spoel, D. & Lindahl, E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *Journal of Chemical Theory and Computation* **4**, 435-447, doi:10.1021/ct700301q (2008).

139 Büssow, K. Anna Tramontano: Protein structure prediction. Concepts and applications. *Analytical and Bioanalytical Chemistry* **386**, 1579-1580, doi:10.1007/s00216-006-0812-8 (2006).

140 Rohl, C. A., Strauss, C. E. M., Misura, K. M. S. & Baker, D. Protein Structure Prediction Using Rosetta. *Methods in Enzymology* **383**, 66-93, doi:10.1016/S0076-6879(04)83004-0 (2004).

141 Jones, T. A. & Kleywegt, G. J. CASP3 comparative modeling evaluation. *Proteins: Structure, Function, and Bioinformatics* **37**, 30-46, doi:10.1002/(SICI)1097-0134(1999)37:3+<30::AID-PROT6>3.0.CO;2-S (1999).

142 Gray, J. J. *et al.* Vol. 331 281-299 (2003).

143 Rohl, C. A., Strauss, C. E. M., Chivian, D. & Baker, D. Modeling structurally variable regions in homologous proteins with Rosetta. *Proteins* **55**, 656-677, doi:10.1002/prot.10629 (2004).

144 Jones, D. T. Predicting novel protein folds by using FRAGFOLD. *Proteins: Structure, Function and Genetics* **45**, 127-132, doi:10.1002/prot.1171 (2001).

145 Zemla, A., Venclovas, Č., Moult, J. & Fidelis, K. Processing and evaluation of predictions in CASP4. *Proteins: Structure, Function and Genetics* **45**, 13-21, doi:10.1002/prot.10052 (2001).

146 McGuffin, L. J., Bryson, K. & Jones, D. T. The PSIPRED protein structure prediction server. *Bioinformatics (Oxford, England)* **16**, 404-405 (2000).

147 Finkelstein, A. V. & Ptitsyn, O. B. Why do globular proteins fit the limited set of folding patterns? *Progress in biophysics and molecular biology* **50**, 171-190 (1987).

148 Levitt, M. & Chothia, C. Structural patterns in globular proteins. *Nature* **261**, 552-558 (1976).

149 Wang, Z. X. A re-estimation for the total numbers of protein folds and superfamilies. *Protein engineering* **11**, 621-626 (1998).

150 Abual-Rub, M. S. & Abdullah, R. A Survey of protein fold recognition algorithms. *Journal of Computer Science* **4**, 768-776, doi:10.3844/jcssp.2008.768.776 (2008).

151 Sillitoe, I. *et al.* CATH: comprehensive structural and functional annotations for genome sequences. *Nucleic Acids Research* **43**, D376-D381, doi:10.1093/nar/gku947 (2015).

152 Jones, D. T. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *Journal of molecular biology* **287**, 797-815, doi:10.1006/jmbi.1999.2583 (1999).

153 McGuffin, L. J. & Jones, D. T. Improvement of the GenTHREADER method for genomic fold recognition. *Bioinformatics (Oxford, England)* **19**, 874-881 (2003).

154 Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25** (1997).

155 Holm, L. & Sander, C. The FSSP database of structurally aligned protein fold families. *Nucleic Acids Research* **22**, 3600-3609 (1994).

156 Lobley, A., Sadowski, M. I. & Jones, D. T. pGenTHREADER and pDomTHREADER: new methods for improved protein fold recognition and superfamily discrimination. *Bioinformatics (Oxford, England)* **25**, 1761-1767, doi:10.1093/bioinformatics/btp302 (2009).

157 Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Research* **28**, 235-242 (2000).

158 Eswar, N. *et al.* Comparative Protein Structure Modeling Using Modeller. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]* **0 5**, Unit-5.6, doi:10.1002/0471250953.bi0506s15 (2006).

159 Rackovsky, S. Global characteristics of protein sequences and their implications. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 8623-8626, doi:10.1073/pnas.1001299107 (2010).

160 Almeida, J. G., Preto, A. J., Koukos, P. I., Bonvin, A. & Moreira, I. S. MEMBRANE PROTEINS STRUCTURES: A review on computational modeling tools. *Biochim Biophys Acta*, doi:10.1016/j.bbamem.2017.07.008 (2017).

161 Nugent, T. & Jones, D. T. Transmembrane protein topology prediction using support vector machines. *BMC Bioinformatics* **10**, 159-159, doi:10.1186/1471-2105-10-159 (2009).

162 Wimley, W. C. & White, S. H. Experimentally determined hydrophobicity scale for proteins at membrane interfaces. *Nature structural biology* **3**, 842-848 (1996).

163 Koehler, J., Woetzel, N., Staritzbichler, R., Sanders, C. R. & Meiler, J. A unified hydrophobicity scale for multispan membrane proteins. *Proteins* **76**, 13-29, doi:10.1002/prot.22315 (2009).

164 Viklund, H. & Elofsson, A. OCTOPUS: improving topology prediction by two-track ANN-based preference scores and an extended topological grammar. *Bioinformatics (Oxford, England)* **24**, 1662-1668, doi:10.1093/bioinformatics/btn221 (2008).

165 Hayat, S. & Elofsson, A. BOCTOPUS: improved topology prediction of transmembrane beta barrel proteins. *Bioinformatics (Oxford, England)* **28**, 516-522, doi:10.1093/bioinformatics/btr710 (2012).

166 Viklund, H., Bernsel, A., Skwark, M. & Elofsson, A. SPOCTOPUS: a combined predictor of signal peptides and membrane protein topology. *Bioinformatics (Oxford, England)* **24**, 2928-2929, doi:10.1093/bioinformatics/btn550 (2008).

167 Tsaousis, G. N. *et al.* ExTopoDB: a database of experimentally derived topological models of transmembrane proteins. *Bioinformatics (Oxford, England)* **26**, 2490-2492, doi:10.1093/bioinformatics/btq362 (2010).

168 Angarica, V. E. & Sancho, J. Protein dynamics governed by interfaces of high polarity and low packing density. *PloS one* **7**, e48212-e48212, doi:10.1371/journal.pone.0048212 (2012).

169 Nguyen, D., Helms, V. & Lee, P.-H. PRIMSIPLR: prediction of inner-membrane situated pore-lining residues for alpha-helical transmembrane proteins. *Proteins* **82**, 1503-1511, doi:10.1002/prot.24520 (2014).

170 Shen, H. & Chou, J. J. MemBrain: Improving the Accuracy of Predicting Transmembrane Helices. *PLoS ONE* **3**, e2399-e2399 (2008).

171 Meruelo, A. D., Samish, I. & Bowie, J. U. TMKink: a method to predict transmembrane helix kinks. *Protein science : a publication of the Protein Society* **20**, 1256-1264, doi:10.1002/pro.653 (2011).

172 Ebejer, J. P., Hill, J. R., Kelm, S., Shi, J. & Deane, C. M. Memoir: template-based structure prediction for membrane proteins. *Nucleic Acids Res* **41**, W379-383, doi:10.1093/nar/gkt331 (2013).

173 Kelm, S., Shi, J. & Deane, C. M. MEDELLER: homology-based coordinate generation for membrane proteins. *Bioinformatics (Oxford, England)* **26**, 2833-2840, doi:10.1093/bioinformatics/btq554 (2010).

174 Kozma, D. & Tusnady, G. E. TMFoldWeb: a web server for predicting transmembrane protein fold class. *Biology direct* **10**, 54, doi:10.1186/s13062-015-0082-5 (2015).

175 Kozma, D. & Tusnady, G. E. TMFoldRec: a statistical potential-based transmembrane protein fold recognition tool. *BMC Bioinformatics* **16**, 201, doi:10.1186/s12859-015-0638-5 (2015).

176 Yarov-Yarovoy, V., Schonbrun, J. & Baker, D. Multipass membrane protein structure prediction using Rosetta. *Proteins* **62**, 1010-1025, doi:10.1002/prot.20817 (2006).

177 Yarov-Yarovoy, V., Baker, D. & Catterall, W. A. Voltage sensor conformations in the open and closed states in ROSETTA structural models of K(+) channels. *Proc Natl Acad Sci U S A* **103**, 7292-7297, doi:10.1073/pnas.0602350103 (2006).

178 Vargas, E. *et al.* An emerging consensus on voltage-dependent gating from computational modeling and molecular dynamics simulations. *The Journal of general physiology* **140**, 587-594, doi:10.1085/jgp.201210873 (2012).

179 Koehler Leman, J., Mueller, B. K. & Gray, J. J. Expanding the toolkit for membrane protein modeling in Rosetta. *Bioinformatics (Oxford, England)* **33**, 754-756, doi:10.1093/bioinformatics/btw716 (2017).

180 Zhang, Q. C., Petrey, D., Norel, R. & Honig, B. H. Protein interface conservation across structure space. *Proc Natl Acad Sci USA* **107** (2010).

181 Landau, M. *et al.* ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res* **33**, W299-302, doi:10.1093/nar/gki370 (2005).

182 Higgins, D. G. & Sharp, P. M. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene* **73**, 237-244 (1988).

183 Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* **30**, 3059-3066, doi:Doi 10.1093/Nar/Gkf436 (2002).

184 Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113, doi:10.1186/1471-2105-5-113 (2004).

185 Hopf, T. A. *et al.* Sequence co-evolution gives 3D contacts and structures of protein complexes. *Elife* **3**, e03430 (2014).

186 Marks, D. S. *et al.* Protein 3D structure computed from evolutionary sequence variation. *PLoS One* **6**, e28766, doi:10.1371/journal.pone.0028766 (2011).

187 Xue, L. C., Dobbs, D. & Honavar, V. HomPPI: a class of sequence homology based protein-protein interface prediction methods. *BMC Bioinformatics* **12**, 1-24, doi:10.1186/1471-2105-12-244 (2011).

188 Xue, L. C., Dobbs, D., Bonvin, A. M. J. J. & Honavar, V. Computational prediction of protein interfaces: A review of data driven methods. *FEBS Letters* **589**, 3516-3526, doi:10.1016/j.febslet.2015.10.003 (2015).

189 Murakami, Y. & Mizuguchi, K. Applying the Naive Bayes classifier with kernel density estimation to the prediction of protein-protein interaction sites. *Bioinformatics (Oxford, England)* **26** (2010).

190 Porollo, A. & Meller, J. Prediction-based fingerprints of protein-protein interactions. *Proteins* **66** (2007).

191 de Vries, S. J., van Dijk, A. D. & Bonvin, A. M. WHISCY: what information does surface conservation yield? Application to data-driven docking. *Proteins* **63**, 479-489, doi:10.1002/prot.20842 (2006).

192 de Vries, S. J. & Bonvin, A. M. J. J. CPORT: A Consensus Interface Predictor and Its Performance in Prediction-Driven Docking with HADDOCK. *PLoS ONE* **6**, e17695-e17695 (2011).

193 He, B., Mortuza, S. M., Wang, Y., Shen, H. B. & Zhang, Y. NeBcon: Protein contact map prediction using neural network training coupled with naive Bayes classifiers. *Bioinformatics (Oxford, England)*, doi:10.1093/bioinformatics/btx164 (2017).

194 Wang, S., Sun, S., Li, Z., Zhang, R. & Xu, J. Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. *PLoS Comput Biol* **13**, e1005324, doi:10.1371/journal.pcbi.1005324 (2017).

195 Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *Journal of molecular graphics* **14**, 33-38, 27-38 (1996).

196 Schrodinger, LLC. *The PyMOL Molecular Graphics System, Version 1.8* (2015).

197 Vangone, A., Spinelli, R., Scarano, V., Cavallo, L. & Oliva, R. COCOMAPS: a web application to analyze and visualize contacts at the interface of biomolecular complexes. *Bioinformatics (Oxford, England)* **27**, 2915-2916, doi:10.1093/bioinformatics/btr484 (2011).

198 Bogan, A. A. & Thorn, K. S. Anatomy of hot spots in protein interfaces. *J Mol Biol* **280** (1998).

199 Moreira, I. S. *et al.* SpotOn: a web server for protein-protein binding hot-spots. *Sci. Rep. (accepted)* (2017).

200 Wagner, S. *et al.* Consequences of membrane protein overexpression in Escherichia coli. *Molecular & cellular proteomics : MCP* **6**, 1527-1550, doi:10.1074/mcp.M600431-MCP200 (2007).

201 Keskin, O., Ma, B. & Nussinov, R. Hot regions in protein--protein interactions: the organization and contribution of structurally conserved hot spot residues. *J Mol Biol* **345**, 1281-1294, doi:10.1016/j.jmb.2004.10.077 (2005).

202 Cukuroglu, E., Gursoy, A. & Keskin, O. HotRegion: a database of predicted hot spot clusters. *Nucleic Acids Res* **40**, D829-833, doi:10.1093/nar/gkr929 (2012).

203 Tuncbag, N., Keskin, O. & Gursoy, A. HotPoint: hot spot prediction server for protein interfaces. *Nucleic Acids Res* **38**, W402-406, doi:10.1093/nar/gkq323 (2010).

204 Vakser, I. A. Protein-protein docking: from interaction to interactome. *Biophys J* **107**, 1785-1793, doi:10.1016/j.bpj.2014.08.033 (2014).

205 Pierce, B. G. *et al.* ZDOCK server: interactive docking prediction of protein-protein complexes and symmetric multimers. *Bioinformatics (Oxford, England)* **30**, 1771-1773, doi:10.1093/bioinformatics/btu097 (2014).

206 Van Zundert, G. C. P. *et al.* The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *Journal of Molecular Biology* **428**, 720-725, doi:10.1016/j.jmb.2015.09.014 (2016).

207 Asadabadi, E. B. & Abdolmaleki, P. Predictions of Protein-Protein Interfaces within Membrane Protein Complexes. *Avicenna journal of medical biotechnology* **5**, 148-157 (2013).

208 Bordner, A. J. Predicting protein-protein binding sites in membrane proteins. *BMC Bioinformatics* **10**, 312, doi:10.1186/1471-2105-10-312 (2009).

209 Li, B. *et al.* Accurate Prediction of Contact Numbers for Multi-Spanning Helical Membrane Proteins. *Journal of chemical information and modeling* **56**, 423-434, doi:10.1021/acs.jcim.5b00517 (2016).

210 Neuvirth, H., Raz, R. & Schreiber, G. ProMate: a structure based prediction program to identify the location of protein-protein binding sites. *J Mol Biol* **338** (2004).

211 Ahmad, S. & Mizuguchi, K. Partner-Aware Prediction of Interacting Residues in Protein-Protein Complexes from Sequence Data. *PLoS ONE* **6**, e29104-e29104 (2011).

212 ul Amir Afsar Minhas, F., Geiss, B. J. & Ben-Hur, A. PAIRpred: Partner-specific prediction of interacting residues from sequence and structure. *Proteins* **82**, 1142-1155, doi:10.1002/prot.24479 (2014).

213 Hurwitz, N., Schneidman-Duhovny, D. & Wolfson, H. J. Memdock: An {alpha}-helical membrane protein docking algorithm. *Bioinformatics (Oxford, England)*, btw184--btw184-, doi:10.1093/bioinformatics/btw184 (2016).

214 Isberg, V. *et al.* GPCRdb: an information system for G protein-coupled receptors. *Nucleic Acids Res* **45**, 2936, doi:10.1093/nar/gkw1218 (2017).

215 Ballesteros, J. A. & Weinstein, H. in *Methods in Neurosciences* Vol. Volume 25 (ed C. Sealfon Stuart) 366-428 (Academic Press, 1995).

216 Esguerra, M., Siretskiy, A., Bello, X., Sallander, J. & Gutierrez-de-Teran, H. GPCR-ModSim: A comprehensive web based solution for modeling G-protein coupled receptors. *Nucleic Acids Res* **44**, W455-462, doi:10.1093/nar/gkw403 (2016).

217 Rasmussen, S. G. *et al.* Crystal structure of the beta2 adrenergic receptor-Gs protein complex. *Nature* **477**, 549-555, doi:10.1038/nature10361 (2011).

218 Yu, J. & Guerois, R. PPI4DOCK: large scale assessment of the use of homology models in free docking over more than 1000 realistic targets. *Bioinformatics (Oxford, England)* **32**, 3760-3767, doi:10.1093/bioinformatics/btw533 (2016).

219 Thorn, K. S. & Bogan, A. A. ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. *Bioinformatics (Oxford, England)* **17**, 284-285 (2001).

220 Fischer, T. B. *et al.* The binding interface database (BID): a compilation of amino acid hot spots in protein interfaces. *Bioinformatics (Oxford, England)* **19**, 1453-1454 (2003).

221 Kumar, M. D. & Gromiha, M. M. PINT: Protein-protein Interactions Thermodynamic Database. *Nucleic Acids Res* **34**, D195-198, doi:10.1093/nar/gkj017 (2006).

222 Moal, I. H. & Fernandez-Recio, J. SKEMPI: a Structural Kinetic and Energetic database of Mutant Protein Interactions and its use in empirical models. *Bioinformatics (Oxford, England)* **28**, 2600-2607, doi:10.1093/bioinformatics/bts489 (2012).

223 Berman, H. M. *et al.* The Protein Data Bank. *Nucleic acids research* **28**, 235-242 (2000).

224 Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421, doi:10.1186/1471-2105-10-421 (2009).

225 Xiao, N., Cao, D. S., Zhu, M. F. & Xu, Q. S. protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. *Bioinformatics (Oxford, England)* **31**, 1857-1859, doi:10.1093/bioinformatics/btv042 (2015).

226 Du, P., Gu, S. & Jiao, Y. PseAAC-General: fast building various modes of general form of Chou's pseudo-amino acid composition for large-scale protein datasets. *International journal of molecular sciences* **15**, 3495-3506, doi:10.3390/ijms15033495 (2014).

227 van Westen, G. J. P., Wegner, J. K., IJzerman, A. P., van Vlijmen, H. W. T. & Bender, A. Proteochemometric modeling as a tool to design selective compounds and for extrapolating to novel targets. *Medchemcomm* **2**, 16-30, doi:10.1039/c0md00165a (2011).

228 Lin, H. The modified Mahalanobis Discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition. *Journal of theoretical biology* **252**, 350-356, doi:10.1016/j.jtbi.2008.02.004 (2008).

229 Ding, H., Luo, L. & Lin, H. Prediction of cell wall lytic enzymes using Chou's amphiphilic pseudo amino acid composition. *Protein Pept Lett* **16**, 351-355 (2009).

230 Lin, H. & Ding, H. Predicting ion channels and their types by the dipeptide mode of pseudo amino acid composition. *Journal of theoretical biology* **269**, 64-69, doi:10.1016/j.jtbi.2010.10.019 (2011).

231 Ding, H., Liu, L., Guo, F. B., Huang, J. & Lin, H. Identify Golgi protein types with modified Mahalanobis discriminant algorithm and pseudo amino acid composition. *Protein Pept Lett* **18**, 58-63 (2011).

232 Ding, H. *et al.* iCTX-type: a sequence-based predictor for identifying the types of conotoxins in targeting ion channels. *BioMed research international* **2014**, 286419, doi:10.1155/2014/286419 (2014).

233 R: A Language and Environment for Statistical Computing (Vienna, Austria, 2013).

234 Kuhn, M. Building Predictive Models in R Using the *caret* package. *Journal of Statistical Software* **28**, 1-28 (2008).

235 Abdi, H. & Williams, L. J. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics* **2**, 433-459, doi:10.1002/wics.101 (2010).

236 Valentini, G. & Masulli, F. in *Neural Nets: 13th Italian Workshop on Neural Nets, WIRN VIETRI 2002 Vietri sul Mare, Italy, May 30 – June 1, 2002 Revised Papers.* (eds Maria Marinaro & Roberto Tagliaferri) 3-20 (Springer Berlin Heidelberg).

237 Xu, D., Tsai, C. J. & Nussinov, R. Hydrogen bonds and salt bridges across protein-protein interfaces. *Protein Eng* **10**, 999-1012 (1997).

238 Tsai, C. J., Lin, S. L., Wolfson, H. J. & Nussinov, R. Studies of protein-protein interfaces: a statistical analysis of the hydrophobic effect. *Protein Sci* **6** (1997).

239 Landgraf, R., Xenarios, I. & Eisenberg, D. Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. *J Mol Biol* **307**, 1487-1502, doi:10.1006/jmbi.2001.4540 (2001).

240 Yan, C., Wu, F., Jernigan, R. L., Dobbs, D. & Honavar, V. Characterization of protein-protein interfaces. *The protein journal* **27**, 59-70, doi:10.1007/s10930-007-9108-x (2008).

241 Liu, Q., Ren, J., Song, J. & Li, J. Co-Occurring Atomic Contacts for the Characterization of Protein Binding Hot Spots. *PLoS One* **10**, e0144486, doi:10.1371/journal.pone.0144486 (2015).

242 Keskin, O., Ma, B., Rogale, K., Gunasekaran, K. & Nussinov, R. Protein-protein interactions: organization, cooperativity and mapping in a bottom-up Systems Biology approach. *Physical biology* **2**, S24-35, doi:10.1088/1478-3975/2/2/S03 (2005).

243 Suzek, B. E. *et al.* UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics (Oxford, England)* **31**, 926-932, doi:10.1093/bioinformatics/btu739 (2015).

244 Wickam, H. *ggplot2: Elegant Graphics for Data Analysis.*, (Springer-Verlag, 2009).

245 Inc., P. T. *Collaborative data science*, <https://plot.ly/> (2015).

246 Inc., R. *Easy web applications in R.*, <http://www.rstudio.com/shiny/> (2013).

247 Candel, A. & Parmar, V. *Deep Learning with H2O*. (H2O.ai Inc., 2015).

248 Grant, B. J., Rodrigues, A. P., ElSawy, K. M., McCammon, J. A. & Caves, L. S. Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics (Oxford, England)* **22**, 2695-2696, doi:10.1093/bioinformatics/btl461 (2006).

249 Nielsen, M. *Neural Networks and Deep Learning*. (2015).

250 LeCun, Y., Bottou, L., Orr, G. B. & Müller, K. R. in *Neural Networks: Tricks of the Trade* (eds Genevieve B. Orr & Klaus-Robert Müller) 9-50 (Springer Berlin Heidelberg, 1998).

251 GoodFellow, I. J., Warde-Farley, D., Mirza, M., Courville, A. & Bengio, Y. Maxout Networks. *JMLR WCP* **28**, 1319-1327 (2013).

252 Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: Synthetic minority over-sampling technique. *J Artif Intell Res* **16**, 321-357 (2002).

253 Blagus, R. & Lusa, L. SMOTE for high-dimensional class-imbalanced data. *Bmc Bioinformatics* **14**, doi:Artn 106

10.1186/1471-2105-14-106 (2013).

254 The UniProt, C. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* **45**, D158-D169, doi:10.1093/nar/gkw1099 (2017).

255 Kang, Y. *et al.* Crystal structure of rhodopsin bound to arrestin by femtosecond X-ray laser. *Nature* **523**, 561-567, doi:10.1038/nature14656 (2015).

256 Kruse, A. C. *et al.* Structure and dynamics of the M3 muscarinic acetylcholine receptor. *Nature* **482**, 552-556, doi:10.1038/nature10867 (2012).

257 Bahar, I., Lezon, T. R., Bakan, A. & Shrivastava, I. H. Normal mode analysis of biomolecular structures: functional mechanisms of membrane proteins. *Chemical reviews* **110**, 1463-1497, doi:10.1021/cr900095e (2010).

258 Munteanu, C. R. *et al.* Solvent accessible surface area-based hot-spot detection methods for protein-protein and protein-nucleic acid interfaces. *J Chem Inf Model* **55**, 1077-1086, doi:10.1021/ci500760m (2015).

259 Kim, D. E., Chivian, D. & Baker, D. Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res* **32**, W526-531, doi:10.1093/nar/gkh468 (2004).

260 Zhu, X. & Mitchell, J. C. KFC2: a knowledge-based hot spot prediction method based on interface solvation, atomic density, and plasticity features. *Proteins* **79**, 2671-2683, doi:10.1002/prot.23094 (2011).

261 Martins, J. M., Ramos, R. M., Pimenta, A. C. & Moreira, I. S. Solvent-accessible surface area: How well can be applied to hot-spot detection? *Proteins* **82**, 479-490, doi:10.1002/prot.24413 (2014).

262 Zhanhua, C., Gan, J. G., Lei, L., Sakharkar, M. K. & Kangueane, P. Protein subunit interfaces: heterodimers versus homodimers. *Bioinformation* **1**, 28-39 (2005).

263 Jones, S. Computational and structural characterisation of protein associations. *Advances in experimental medicine and biology* **747**, 42-54, doi:10.1007/978-1-4614-3229-6_3 (2012).

264 Halperin, I., Wolfson, H. & Nussinov, R. Protein-protein interactions; coupling of structurally conserved residues and of hot spots across interfaces. Implications for docking. *Structure* **12**, 1027-1038, doi:10.1016/j.str.2004.04.009 (2004).

265 Li, X., Keskin, O., Ma, B., Nussinov, R. & Liang, J. Protein-protein interactions: hot spots and structurally conserved residues often locate in complemented pockets that pre-organized in the unbound states: implications for docking. *J Mol Biol* **344**, 781-795, doi:10.1016/j.jmb.2004.09.051 (2004).

266 Carbonell, P., Nussinov, R. & del Sol, A. Energetic determinants of protein binding specificity: insights into protein interaction networks. *Proteomics* **9**, 1744-1753, doi:10.1002/pmic.200800425 (2009).

267 Fleming, P. J. & Richards, F. M. Protein packing: dependence on protein size, secondary structure and amino acid composition. *J Mol Biol* **299**, 487-498, doi:10.1006/jmbi.2000.3750 (2000).

268 Yogurtcu, O. N., Erdemli, S. B., Nussinov, R., Turkay, M. & Keskin, O. Restricted mobility of conserved residues in protein-protein interfaces in molecular simulations. *Biophys J* **94**, 3475-3485, doi:10.1529/biophysj.107.114835 (2008).

269 Jones, S. & Thornton, J. M. Searching for functional sites in protein structures. *Current opinion in chemical biology* **8**, 3-7, doi:10.1016/j.cbpa.2003.11.001 (2004).

270 Janin, J., Bahadur, R. P. & Chakrabarti, P. Protein-protein interaction and quaternary structure. *Q Rev Biophys* **41**, 133-180, doi:10.1017/S0033583508004708 (2008).

271 Ma, B., Elkayam, T., Wolfson, H. & Nussinov, R. Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proc Natl Acad Sci U S A* **100**, 5772-5777, doi:10.1073/pnas.1030237100 (2003).

272 Saldano, T. E., Monzon, A. M., Parisi, G. & Fernandez-Alberti, S. Evolutionary Conserved Positions Define Protein Conformational Diversity. *PLoS Comput Biol* **12**, e1004775, doi:10.1371/journal.pcbi.1004775 (2016).

273 Moreira, I. S. The Role of Water Occlusion for the Definition of a Protein Binding Hot-Spot. *Current topics in medicinal chemistry* **15**, 2068-2079 (2015).

274 Moreira, I. S., Ramos, R. M., Martins, J. M., Fernandes, P. A. & Ramos, M. J. Are hot-spots occluded from water? *Journal of biomolecular structure & dynamics* **32**, 186-197, doi:10.1080/07391102.2012.758598 (2014).

275 Melo, R., Fieldhouse, R., Melo, A., Correia, J. D. G. & Cordeiro, M. N. D. S. A Machine-Learning Approach for Hot-Spot Detection at Protein-Protein Interfaces. *IJMS*, 1-16 (2016).

276 Hu, Z., Ma, B., Wolfson, H. & Nussinov, R. Conservation of polar residues as hot spots at protein interfaces. *Proteins* **39**, 331-342 (2000).

277 Jones, S. & Thornton, J. M. Analysis of protein-protein interaction sites using surface patches. *J Mol Biol* **272** (1997).

278 Jones, S. & Thornton, J. M. Prediction of protein-protein interaction sites using patch analysis. *J Mol Biol* **272**, 133-143, doi:10.1006/jmbi.1997.1233 (1997).

279 Xu, D., Lin, S. L. & Nussinov, R. Protein binding versus protein folding: the role of hydrophilic bridges in protein associations. *J Mol Biol* **265**, 68-84, doi:10.1006/jmbi.1996.0712 (1997).

280 Sibi, P., Jones, S. A. & Siddarth, P. Analysis of Different Activation Functions Using Back Propagation Neural Networks. *Journal of Theoretical and Applied Information Technology* **47**, 1264-1268 (2013).

281 Heaton, J. *Introduction to Neural Networks for Java, 2nd Edition*. (Heaton Research, Inc., 2008).

282 Zou, W., Li, Y. & Tang, A. Effects of the number of hidden nodes used in a structured-based neural network on the reliability of image classification. *Neural Computing and Applications* **18**, 249-260, doi:10.1007/s00521-008-0177-3 (2009).

283 Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research* **15**, 1929-1958 (2014).

284 Sensoy, O., Almeida, J. G., Shabbir, J., Moreira, I. S. & Morra, G. in *Molecular Biology Protocols* (Submitted).

285 Ozcan, O., Uyar, A., Doruker, P. & Akten, E. D. Effect of intracellular loop 3 on intrinsic dynamics of human beta2-adrenergic receptor. *BMC structural biology* **13**, 29, doi:10.1186/1472-6807-13-29 (2013).

286 Pydi, S. P., Singh, N., Upadhyaya, J., Bhullar, R. P. & Chelikani, P. The third intracellular loop plays a critical role in bitter taste receptor activation. *Biochim Biophys Acta* **1838**, 231-236, doi:10.1016/j.bbamem.2013.08.009 (2014).

287 Sun, B. *et al.* Crystal structure of the adenosine A2A receptor bound to an antagonist reveals a potential allosteric pocket. *Proc Natl Acad Sci U S A* **114**, 2066-2071, doi:10.1073/pnas.1621423114 (2017).

288 Gomez-Mouton, C. *et al.* Filamin A interaction with the CXCR4 third intracellular loop regulates endocytosis and signaling of WT and WHIM-like receptors. *Blood* **125**, 1116-1125, doi:10.1182/blood-2014-09-601807 (2015).

289 Almeida, J. G., Bonvin, A. M. J. J. & Moreira, I. S. SpotOnDB: a global assessment of protein-protein interfaces and binding hot-spots (In preparation).

290 Preto, A. J. *et al.* Understanding the Binding Selectivity of G-protein Coupled Receptors Toward G-proteins and Arrestins: Application to the Dopamine Receptor Family. (In preparation).

291 Shirian, J., Sharabi, O. & Shifman, J. M. Cold Spots in Protein Binding. *Trends in biochemical sciences* **41**, 739-745, doi:10.1016/j.tibs.2016.07.002 (2016).

292 Miller, S., Lesk, A. M., Janin, J. & Chothia, C. The accessible surface area and stability of oligomeric proteins. *Nature* **328**, 834-836, doi:10.1038/328834a0 (1987).

293 Sun, T., Zhou, B., Lai, L. & Pei, J. Sequence-based prediction of protein protein interaction using a deep-learning algorithm. *BMC Bioinformatics* **18**, 277, doi:10.1186/s12859-017-1700-2 (2017).

294 Gomes, J., Ramsundar, B., Feingberg, E. N. & Pande, V. S. Atomic Convolutional Networks for Predicting Protein-Ligand Binding Affinity. *arXiv*, doi:arXiv:1703.10603 (2017).

295 Ma, J. Usefulness and limitations of normal mode analysis in modeling dynamics of biomolecular complexes. *Structure* **13**, 373-380, doi:10.1016/j.str.2005.02.002 (2005).

F. List of Abbreviations

AAC – Amino Acid Composition
AMBER – Assisted Model Building with Energy Refinement
AUROC – Area Under Receiver Operating Curve
ASEdb – Alanine Scanning Energetics database
BID – Binding Interface Database
BLAST – Basic Local Alignment Search Tool
BLOSUM – BLOcks Substitution Matrix
BW – Ballesteros-Weinstein
CASP – Critical Assessment of Protein Structure
cAMP – Cyclic Adenosine Monophosphate
CATH – Class, Architecture, Topology, and Homologous superfamily
CD-HIT – Cluster Database with High Identity Tolerance
CHARMM – Chemistry at HARvard Macromolecular Mechanics
CLC – Chloride Conducting Channel
Cons – Conservation Scores
CPORT – Consensus Prediction Of interface Residues in Transient complexes
Cryo-EM – Cryo-electron Microscopy
CS-BLAST – Context Specific-Basic Local Alignment Search Tool
ECL1 – Extracellular Loop 1
ECL2 – Extracellular Loop 2
ECL3 – Extracellular Loop 3
E. Factor – Enrichment Factor
FASTA – Fast Adaptive Shrinkage/Thresholding Algorithm
FDR – False Discovery Rate
FN – False Negative
FNR – False Negative Rate
FP – False Positive
FSSP – Families of Structuraly Similar Proteins
GROMACS – GROningen MAchine for Chemical Simulations
HADDOCK – High Ambiguity Driven protein-protein DOCKing
H-bonds – Hydrogen Bonds
HMM – Hidden Markov Model
HS – Hot-spot
HX8 – Helix 8
ICL1 – Intracellular Loop 1
ICL2 – Intracellular Loop 2
ICL3 – Intracellular Loop 3
MANOVA – Multiple Analysis of Variance
MCC – Mathew's Correlation Coefficient
MD – Molecular Dynamics
mfDCA – mean-Field Direct Coupling Analysis
MP – Membrane Protein
MSA – Multiple Sequence Alignment
MUSCLE – Multiple Sequence Comparison by Log-Expectation
NMA – Normal Mode Analysis
NPV – Negative Predictive Value
NS – Null-spots
PAAC – Pseudo-Amino Acid Composition
PAM – Point Accepted Mutation

PDB – Protein Data Bank
PCA – Principal Component Analysis
PCM – ProteoChemometric Modelling
PINT – Protein-protein Interactions Thermodynamic database
PPIPP – Protein-Protein Interaction Prediction Platform
PPV – Positive Predictive Value
PSI-BLAST – Position-Specific Iterative Basic Local Alignment Search Tool
PTM – Post-Translational Modifications
PMP – Peripheral Membrane Protein
PSSM – Position-Specific Scoring Matrix
RF – Random Forests
RMSD – Root-Mean-Square-Deviation
SASA – Solvent Accessible Surface Area
sd – Standard Deviation
SKEMPI – Structural database of Kinetics and Energetics of Mutant Protein Interactions
SVM – Support Vector Machine
TM – Transmembrane
TMH – Transmembrane Helix
TMH1 – Transmembrane Helix 1
TMH2 – Transmembrane Helix 2
TMH3 – Transmembrane Helix 3
TMH4 – Transmembrane Helix 4
TMH5 – Transmembrane Helix 5
TMH6 – Transmembrane Helix 6
TMH7 – Transmembrane Helix 7
TN – True Negative
TP – True Positive
UniProt – Universal Protein Resource
VMD – Visual Molecular Dynamics
WHISCY – WHat Information Does Surface Conservation Yield?

G. Annex

Annex 1 - This table comprehends the machine learning algorithms mentioned along this thesis work with a short explanation for each.

Algorithm	Explanation
Artificial Neural Network (ANN)	An ANN is a machine-learning algorithm inspired in how neurons connect and interact with each other. Essentially, it is composed of layers of neurons, which output information to all neurons in the following layer and receive information from all neurons in the previous layer. Each connection has two trainable parameters: weight (w) and bias (b), which function as follows: $$O = w * i(O') + b,$$ where O is the output for the current layer and i is a function of the output of the previous layer. Three types of layers are essential to ANNs – the input layer, which takes direct input from the data, the output layer, which provides the prediction, and one or more hidden layers, which take data from a layer to the other. To train an ANN, using the most popular training algorithm – back-propagation – weights are adjusted iteratively depending on the error [1].
Bagging algorithms	Bagging is short for bootstrap aggregating and it is not a machine-learning algorithm *per se* but rather a meta-algorithm which uses other algorithms to achieve a better performance. Instead of training a machine-learning model using the entire training set, it creates several random subsets with replacement and trains various identical models using these subsets. Next, it averages the parameters of the resulting models to get a final, more robust model.
Convolutional Neural Network (CNN)	CNNs are a type of deep ANN which as gained a great deal of popularity in image processing and high-throughput gene analysis. Its main feature is the *convolution* of the data – a grid (bidimensional for images and unidimensional for genes) takes as input several pixels or DNA bases and, by progressively reducing the number of hidden layers, is able to reduce the number of dimensions of the input data [1].
Deep-learning	Deep learning can be considered more easily as an ANN architecture than actual machine-learning algorithm. When an ANN has more than one hidden layer, it can be considered as having a deep architecture, thus being referred to as deep-learning. Deep-learning has spanned a field of its own when discussing machine-learning, with heavy investigation being pursued on both theoretical and practical aspects, particularly in bioinformatics [1-19]. However, besides heavy investigation, its theoretical foundation is still lacking – understanding what hyperparameters are best, such as the number of hidden layers and the number of neurons in each hidden layer, is not well understood – and it can be rather difficult to determine what is it that makes deep learning a highly successful technique. Furthermore, it tends to require heavy amounts of data, making it hardly applicable to most biological problems where data retrieval can be challenging.
Discriminant Analysis	Discriminant analysis which aims to determine whether a set of independent variables are able to predict a dependent variable. To do so, it combines various noncorrelated functions of independent variables.

Hidden Markov Models (HMM)	A Markov model is composed of several Markov processes – state transitions that depend solely on the current state and not on past states – in a chain or network of possible states. HMMs are similar processes but the sequence of state transitions is unknown – it is hidden, thus the term. The reason why they are so sought in computational biology is that they are able to address three distinct and highly relevant problems – scoring a sequence in a multiple sequence alignment by determining the probability of an HMM generating a sequence, finding the optimal state sequence for a HMM to generate a sequence and finding what is the proper structure and parameters for a HMM to be able to solve a problem as a machine-learning model [20].
k-means clustering	k-means clustering is a machine-learning-based process to find the best possible fit for a given number of k clusters. It determines the best position of the centroid (average position) of all clusters. Through iterations, the squared error function is calculated: $SE(V) = \sum_{i=1}^{c}\sum_{j=1}^{c_i}(\left\|\left\|x_i - v_j\right\|\right\|)^2,$ where ||x$_i$-x$_j$|| is the Euclidean distance between x$_i$ and v$_j$, c$_i$ is the number of observations in cluster i and c is the number of centroids. By minimizing this function, the centroids with the least distance to each individual observation will be obtained.
k-nearest neighbours	The k-nearest neighbours algorithm uses Euclidian distance to attribute a class to a new observation. To do so, it calculates the distance to the nearest k labelled observations and, using a majority vote or other sort of operation (a logistic regression, for example) assigns the most likely class to the unlabelled observation. A common procedure is to consider labelled observations which are nearest to be more decisive than those that are distant. It is considered an instance-based (lazy) algorithm because all computation is done upon classification time [21].
Logistic regression	Logistic regressions are regression models utilized for categorical dependent variables. For multivariate cases, the output of a logistic regression works as follows: $O(x) = \frac{e^{a_1+a_2X_1+a_3X_2+\cdots+a_nX_n}}{1+e^{a_1+a_2X_1+a_3X_2+\cdots+a_nX_n}},$ Where $X_1, X_2, \ldots, X_n$ and $a_1, a_2, \ldots, a_n$ are the variables and their coefficients, respectively [22].
Random Forest (RF)	RFs are ensembles of binary decision trees (BDTs). BDTs provide one of two outputs based on the input. This decision can be triggered by simple Boolean decisions (True or False) or by more complex activation functions [23].
Support Vector Machine (SVM)	SVMs are non-probabilistic binary linear classifiers – they do not provide a probability if an object belongs to either class, they assign it to a single class. It features an n-1-dimensional hyperplane in an n-dimensional space dividing the data into two different classes. It became highly popular because it is very reliable and computationally inexpensive due to its simple construct – it only needs to determine which observations are the margins to draw the hyperplane. The optimal margins are the observations which maximize the distance from the hyperplane thus allowing it to accurately separate the two classes with little information [24].
Synthetic Minority Over-sampling Technique (SMOTE)	SMOTE is a technique which uses the k-nearest neighbours algorithm to generate new entries in the dataset. To do so it draws lines between this entry and all its k-nearest neighbours. Then, synthetic data points are generated along these lines [25]. While this technique is

	effective in low-dimension spaces, high-dimensionality has proven to decrease the performance of SMOTE [26].

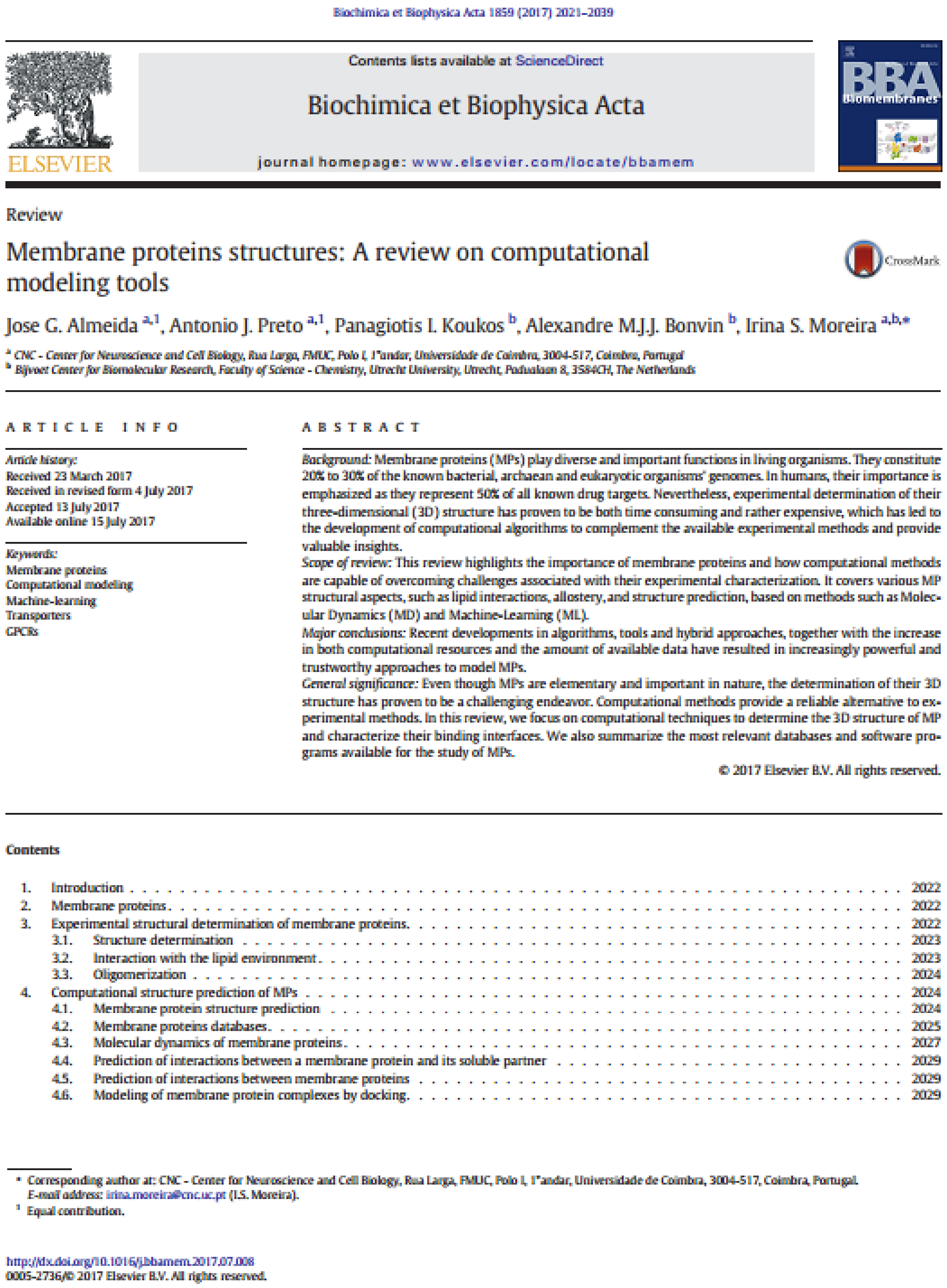
Annex 2 - This table comprehends an explanation of all columns in the dataset used to construct the SpotOnDB.

Column number	Explanation
1	A unique identifier for each residue entry
2	Protein Data Bank [27] entry chains considered for each PPI4DOCK [28] complex, separated by an underscore
3:5	Chain, residue number and residue name
6	Prediction of a residue as HS (1) or NS (0) according to the SpotOn algorithm [29]
7:18	Features related to Solvent Accessible Surface Area (SASA) [30,31]
19:58	Position-Specific Scoring Matrix (PSSM) values and PSSM proportions as calculated by PSI-BLAST [32]
59, 60	Number of atoms within 2.5 and 4.0 Å
61	Number of hydrogen bonds being formed by the residue
62	Number of hydrophobic interactions with the residue
63	Number of π-π interactions (aromatic) with the residue
64	Number of T-stacking interactions (aromatic) with the residue
65	Number of cation-π interactions (aromatic) with the residue
66	Number of salt bridges formed by the residue
67:86	Number of monomeric interfacial residues
87	Total number of interfacial residues
88	Total ΔSASA [30,31]
89:168	Amphiphilic Pseudo-Amino Acid composition (features extracted with the protr package [33] from R)
169:188	Amino Acid Composition (features extracted with the protr package [33] from R)
189:888	Scales-based descriptors derived by 20+ classes of 2D and 3D descriptors, including topological chemical descriptors, weighted holistic invariant molecular descriptors, vectors of hydrophobic, steric and electronic properties, among others (features extracted with the protr package [33] from R)
889:891	Raw b-factor values for the whole residue, backbone atoms and sidechain atoms
892:894	Normalized b-factor values for the whole residue, backbone atoms and sidechain atoms according to [34]
895	Normalized conservation scores from Consurf [35]
896:901	Number of intramonomer, intermonomer and HS neighbours at 8 Å and 10 Å

Annex 3 - Papers published during this thesis work.

Reference	**Abstract**
Almeida JG, Preto AJ, Koukos P, Bonvin AJJM, Moreira IS, Membrane proteins structures: a review on computational modeling tools, BBA Biomembranes 1859, 10, 2021-2039 (2017) (Review article) [37]	**Background** Membrane proteins (MPs) play diverse and important functions in living organisms. They constitute 20% to 30% of the known bacterial, archaean and eukaryotic organisms' genomes. In humans, their importance is emphasized as they represent 50% of all known drug targets. Nevertheless, experimental determination of their three-dimensional (3D) structure has proven to be both time consuming and rather expensive, which has led to the development of computational algorithms to complement the available experimental methods and provide valuable insights. **Scope of review** This review highlights the importance of membrane proteins and how computational methods are capable of overcoming challenges associated with their experimental characterization. It covers various MP structural aspects, such as lipid interactions, allostery, and structure prediction, based on methods such as Molecular Dynamics (MD) and Machine-Learning (ML). **Major conclusions** Recent developments in algorithms, tools and hybrid approaches, together with the increase in both computational resources and the amount of available data have resulted in increasingly powerful and trustworthy approaches to model MPs. **General significance** Even though MPs are elementary and important in nature, the determination of their 3D structure has proven to be a challenging endeavor. Computational methods provide a reliable alternative to experimental methods. In this review, we focus on computational techniques to determine the 3D structure of MP and characterize their binding interfaces. We also summarize the most relevant databases and software programs available for the study of MPs.
Lemos A, Melo R, Preto AJ, Almeida JG, Moreira IS, Cordeiro MNDS, In silico studies targeting G-protein coupled receptors for drug research against Parkinson's disease, Current Neuropharmacology, submitted (Book chapter) [38]	Parkinson's Disease (PD) is a long-term neurodegenerative brain disorder that mainly affects the motor system. The causes are still unknown, and even though currently there is no cure, several therapeutic options are available to manage its symptoms. The development of novel anti-parkinsonian agents and an understanding of their proper and optimal use are, indeed, highly demanding. For the last decades, L-3,4-DihydrOxyPhenylAlanine or levodopa (L-DOPA) has been the gold-standard therapy for the symptomatic treatment of motor dysfunctions associated to PD. However, the development of dyskinesias and motor fluctuations (*wearing-off* and *on-off* phenomena) associated to long-term L-DOPA replacement therapy have limited its antiparkinsonian efficacy. The investigation for non-dopaminergic therapies has been largely explored as an attempt to counteract the motor side effects associated to dopamine replacement therapy. Being one of the largest cell membrane protein families, G-Protein-Coupled Receptors (GPCRs) have become a relevant target for drug discovery focused in a wide range of therapeutic areas, including Central Nervous System (CNS) diseases. The modulation of specific GPCRs potentially implicated in PD, excluding dopamine receptors, may provide promising non-dopaminergic therapeutic alternatives for symptomatic treatment of PD. In this review, we focused on the impact of specific GPCR subclasses, including dopamine receptors, adenosine receptors, muscarinic acetylcholine receptors, metabotropic glutamate receptors, and 5-hydroxytryptamine receptors, on the pathophysiology of PD and the importance of structure- and ligand-based *in silico* approaches for the development of small molecules to target these receptors.
Moreira IS, Koukos P, Melo R, Almeida JG, Preto AJ, Schaarschmidt J, Trellet M, Gumus ZH, Costa J, Bonvin AMJJ, SpotOn: a web server for protein-protein binding hot-spots, Scientific Reports 7, 8007 (2017) (Scientific article) [29]	We present SpotOn, a web server to identify and classify interfacial residues as Hot-Spots (HS) and Null-Spots (NS). SpotON implements a robust algorithm with a demonstrated accuracy of И.9ɔ and sensitivity of И.98 on an independent test set. The predictor was developed using an ensemble machine learning approach with up-sampling of the minor class. It was trained on ɔ⍰ complexes using various features, based on both protein ⍰D structure and sequence. The SpotOn web interface is freely available at: http://milou.science.uu.nl/services/SPOTON/.
Almeida JG, Bonvin AMJJ, Moreira IS, SpotOnDB: a global assessment of protein-protein interfaces and binding hot-spots (in preparation)	Understanding protein-protein interfaces is crucial to explain complex formation and determine the foundations that governs biological networks. One key aspect is the existence and prevalence of Hot-Spots (HS), residues which, upon alanine mutation, negatively impact the formation of protein-protein complexes. While several protein complexes have been individually studied regarding their interface and the existence of HS, studies comprising high amounts of data would be highly valuable in revealing relevant patterns. In this work, we use our computational pipeline, SpotOn, to determine several structural and sequence-related features of all predicted HS in the complexes of the non-redundant database PPI4DOCK (3.746 HS for 1.403 complexes with 66.710 interfacial residues). The resulting big data analysis, which is available as an interactive online database at

	http://milou.science.uu.nl/services/SPOTONDB/, provides insights into HS structural and physico-chemical characteristics in protein-protein interfaces.
Sensoy O, Almeida JG, Shabbir J, Moreira IS, Morra G, Computational studies of G-protein coupled receptor complexes: structure and dynamics, Molecular Biology Protocols, accepted (Book chapter) [36]	G-protein coupled receptors (GPCRs) are ubiquitously expressed transmembrane proteins associated with a wide range of diseases such as Alzheimer's, Parkinson, schizophrenia and also implicated in several abnormal heart conditions. As such, this family of receptors is regarded as excellent drug targets. However, due to the high number of intracellular signaling partners, these receptors have a complex interaction networks and it becomes challenging to modulate their function. Experimentally determined structures give detailed information on the salient structural properties of these signaling complexes but they are far away from providing mechanistic insights into the underlying process. This chapter presents some of computational tools, namely molecular dynamics, molecular docking and molecular modeling and related analyses methods that have been used to complement experimental findings.
Preto, A.J., Almeida J.G., Melo A., Kurkcuoglu Z., Melo R., Telle M., Melo A., Natalia M.N.D.S., Morra G., Sensoy O., Bonvin A.M.J.J., Moreira I.S. Understanding the Binding Selectivity of G-protein Coupled Receptors Toward G-proteins and Arrestins: Application to the Dopamine Receptor Family (in preparation)	Elucidation of crystal structures pertaining to ternary complexes of 2-Adrenergic Receptor (2-AR) and Rhodopsin bound to Gs and Arrestin-1, respectively, has enhanced our understanding in the molecular determinants responsible for selective coupling of G-protein-coupled-receptor (GPCR)-G-protein or GPCR-Arrestin (Arr) complexes. Nevertheless, these are single examples of an immense number of possible interfaces established between these molecular systems. In this study, we focus on catecholamine-bound GPCRs, in particular the Dopamine Receptor (DR) family, in order to bring new insights into the physiological and pharmacological properties of these important drug targets. A variety of computational methods were applied to investigate the putative interactions between the protein interfaces of all members of the DR family (D1R, D2R, D3R, D4R and D5R) and the protein interfaces of their binding partners (Arrs: Arr-2, Arr-3; G-protein: Gq, Gz, Gt2, Gi1, Gi2, Gi3, GsS, Go, GsL). To effectively compare DR-partner interactions, we calculated solvent accessibility, structural conservation, residue composition and propensity, packing density, interfacial residue mobility, and residue pairing. We have also employed elastic network model to understand their dynamics upon coupling, and the difference of behaviour between the DR binding partners. Elucidation of pharmacologically relevant interactions between DR complexes and their binding partners at the molecular level allows us not only to unlock the determinants of the functional selectivity, but also expedite designing of more selective and hence safer therapeutic molecules for treatment of GPCR-mediated diseases.



Biochimica et Biophysica Acta 1859 (2017) 2021–2039

Contents lists available at ScienceDirect

Biochimica et Biophysica Acta

journal homepage: www.elsevier.com/locate/bbamem

Review

Membrane proteins structures: A review on computational modeling tools

Jose G. Almeida [a,1], Antonio J. Preto [a,1], Panagiotis I. Koukos [b], Alexandre M.J.J. Bonvin [b], Irina S. Moreira [a,b,*]

[a] CNC - Center for Neuroscience and Cell Biology, Rua Larga, FMUC, Polo I, 1°andar, Universidade de Coimbra, 3004-517, Coimbra, Portugal
[b] Bijvoet Center for Biomolecular Research, Faculty of Science - Chemistry, Utrecht University, Utrecht, Padualaan 8, 3584CH, The Netherlands

ARTICLE INFO

Article history:
Received 23 March 2017
Received in revised form 4 July 2017
Accepted 13 July 2017
Available online 15 July 2017

Keywords:
Membrane proteins
Computational modeling
Machine-learning
Transporters
GPCRs

ABSTRACT

Background: Membrane proteins (MPs) play diverse and important functions in living organisms. They constitute 20% to 30% of the known bacterial, archaean and eukaryotic organisms' genomes. In humans, their importance is emphasized as they represent 50% of all known drug targets. Nevertheless, experimental determination of their three-dimensional (3D) structure has proven to be both time consuming and rather expensive, which has led to the development of computational algorithms to complement the available experimental methods and provide valuable insights.

Scope of review: This review highlights the importance of membrane proteins and how computational methods are capable of overcoming challenges associated with their experimental characterization. It covers various MP structural aspects, such as lipid interactions, allostery, and structure prediction, based on methods such as Molecular Dynamics (MD) and Machine-Learning (ML).

Major conclusions: Recent developments in algorithms, tools and hybrid approaches, together with the increase in both computational resources and the amount of available data have resulted in increasingly powerful and trustworthy approaches to model MPs.

General significance: Even though MPs are elementary and important in nature, the determination of their 3D structure has proven to be a challenging endeavor. Computational methods provide a reliable alternative to experimental methods. In this review, we focus on computational techniques to determine the 3D structure of MP and characterize their binding interfaces. We also summarize the most relevant databases and software programs available for the study of MPs.

Contents

[*] Corresponding author at: CNC - Center for Neuroscience and Cell Biology, Rua Larga, FMUC, Polo I, 1°andar, Universidade de Coimbra, 3004-517, Coimbra, Portugal.
E-mail address: irina.moreira@cnc.uc.pt (I.S. Moreira).

[1] Equal contribution.

http://dx.doi.org/10.1016/j.bbamem.2017.07.008
0005-2736/© 2017 Elsevier B.V. All rights reserved.

In silico studies in the drug research against Parkinson's disease

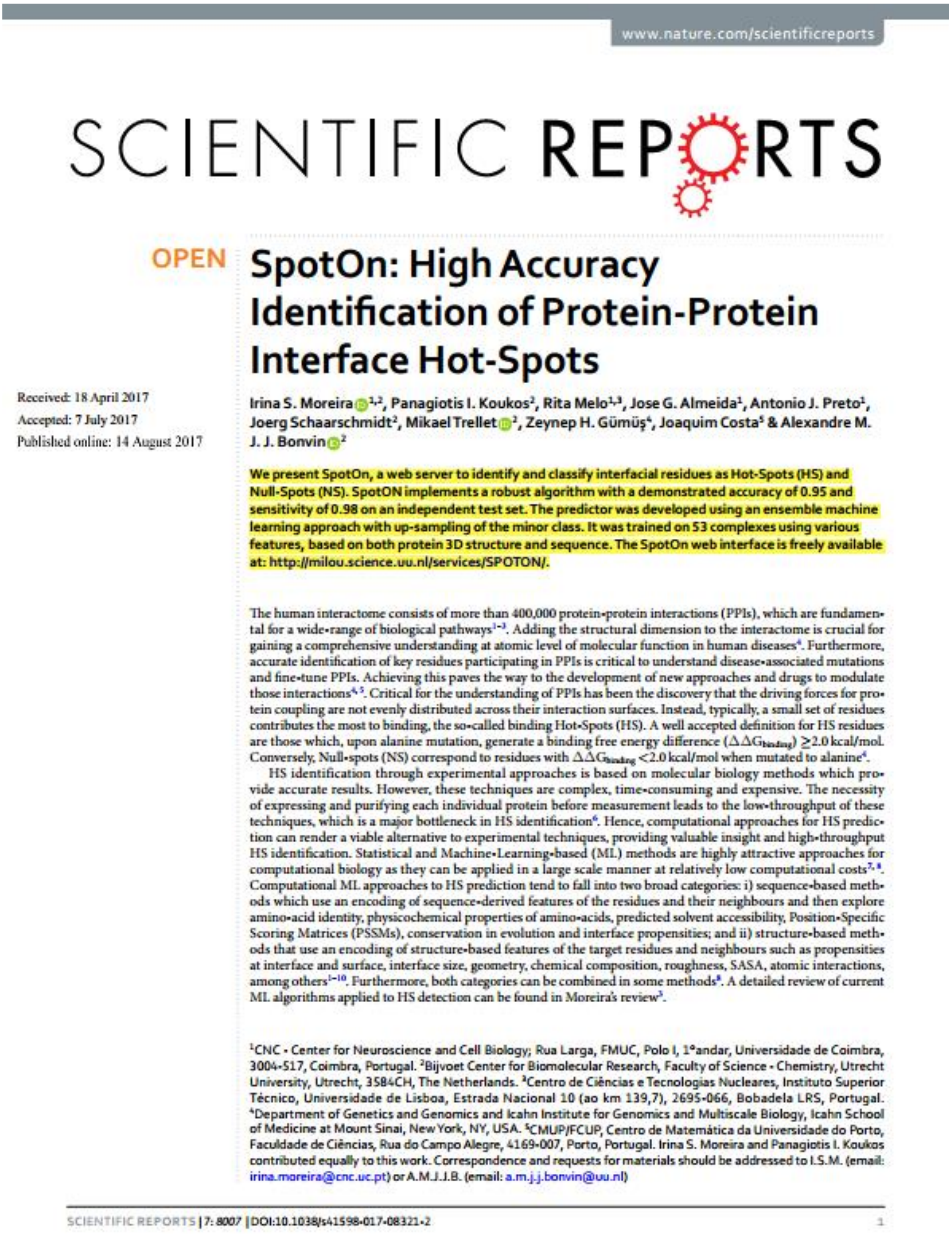
Agostinho Lemos[a], Rita Melo[b,c], Antonio J. Preto[b], Jose G. Almeida[b], Irina S. Moreira[*b,d], M. Natália D. S. Cordeiro[*a]

[a]LAQV/REQUIMTE, Department of Chemistry and Biochemistry, Faculty of Sciences, University of Porto, Rua do Campo Alegre s/n, 4169-007 Porto, Portugal; [b]CNC - Center for Neuroscience and Cell Biology, Faculty of Medicine, University of Coimbra, Rua Larga, 3004-517 Coimbra, Portugal; [c]Centro de Ciências e Tecnologias Nucleares, Instituto Superior Técnico, Universidade de Lisboa, Estrada Nacional 10 (ao km 139,7), 2695-066 Bobadela LRS, Portugal; [d]Bijvoet Center for Biomolecular Research, Faculty of Science - Chemistry, Utrecht University, Utrecht, 3584CH, The Netherlands

* Address correspondence to these authors at: (MNDSC) LAQV@REQUIMTE/Department of Chemistry and Biochemistry, University of Porto, 4169-007 Porto, Portugal; Fax: +351 220402659; E-mail: ncordeir@fc.up.pt; (ISM) CNC - Center for Neuroscience and Cell Biology, Faculty of Medicine, University of Coimbra, Rua Larga, 3004-517 Coimbra, Portugal; Fax: +351 304502930; E-mail: irina.moreira@cnc.uc.pt

Abstract

Parkinson's Disease (PD) is a long-term neurodegenative brain disorder that mainly affects the motor system. The cause is still unknown, and even though currently there is no cure, several therapeutic options are available to manage its symptoms. Thus, the development of novel anti-parkinsonian agents and an understanding of their proper and optimal use are highly demanding. For the last decades, L-3,4-dihydroxyphenylalanine or levodopa (L-DOPA) has been the gold-standard therapy for the treatment of motor dysfunctions. However, the development of dyskinesias and motor fluctuations (*wearing-off* and *on-off* phenomena) associated to long-term L-DOPA replacement therapy have limited the anti-parkinsonian efficacy of L-DOPA. The investigation for non-dopaminergic therapies has been largely explored as an attempt to counteract the motor side effects associated to dopamine replacement therapy. Being one of the largest cell membrane protein families, G-Protein-Coupled Receptors (GPCRs) have become a relevant target for drug discovery focused in a wide range of therapeutic areas, including Central Nervous System (CNS) diseases. The modulation of specific GPCRs potentially implicated in PD, excluding dopamine receptors, may provide promising non-dopaminergic therapeutic alternatives for symptomatic treatment of PD. In this review, we focus on the impact of specific GPCR subclasses, including dopamine receptors, adenosine receptors, muscarinic acetylcholine receptors, metabotropic glutamate receptors, and 5-hydroxytryptamine receptors, on the pathophysiology of PD and the



SCIENTIFIC REPORTS

OPEN

SpotOn: High Accuracy Identification of Protein-Protein Interface Hot-Spots

Received: 18 April 2017
Accepted: 7 July 2017
Published online: 14 August 2017

Irina S. Moreira[1,2], Panagiotis I. Koukos[2], Rita Melo[1,3], Jose G. Almeida[1], Antonio J. Preto[1], Joerg Schaarschmidt[2], Mikael Trellet[2], Zeynep H. Gümüş[4], Joaquim Costa[5] & Alexandre M. J. J. Bonvin[2]

We present SpotOn, a web server to identify and classify interfacial residues as Hot-Spots (HS) and Null-Spots (NS). SpotON implements a robust algorithm with a demonstrated accuracy of 0.95 and sensitivity of 0.98 on an independent test set. The predictor was developed using an ensemble machine learning approach with up-sampling of the minor class. It was trained on 53 complexes using various features, based on both protein 3D structure and sequence. The SpotOn web interface is freely available at: http://milou.science.uu.nl/services/SPOTON/.

The human interactome consists of more than 400,000 protein-protein interactions (PPIs), which are fundamental for a wide-range of biological pathways[1–3]. Adding the structural dimension to the interactome is crucial for gaining a comprehensive understanding at atomic level of molecular function in human diseases[4]. Furthermore, accurate identification of key residues participating in PPIs is critical to understand disease-associated mutations and fine-tune PPIs. Achieving this paves the way to the development of new approaches and drugs to modulate those interactions[4,5]. Critical for the understanding of PPIs has been the discovery that the driving forces for protein coupling are not evenly distributed across their interaction surfaces. Instead, typically, a small set of residues contributes the most to binding, the so-called binding Hot-Spots (HS). A well accepted definition for HS residues are those which, upon alanine mutation, generate a binding free energy difference ($\Delta\Delta G_{binding}$) $\geq$2.0 kcal/mol. Conversely, Null-spots (NS) correspond to residues with $\Delta\Delta G_{binding}$ <2.0 kcal/mol when mutated to alanine[4].

HS identification through experimental approaches is based on molecular biology methods which provide accurate results. However, these techniques are complex, time-consuming and expensive. The necessity of expressing and purifying each individual protein before measurement leads to the low-throughput of these techniques, which is a major bottleneck in HS identification[6]. Hence, computational approaches for HS prediction can render a viable alternative to experimental techniques, providing valuable insight and high-throughput HS identification. Statistical and Machine-Learning-based (ML) methods are highly attractive approaches for computational biology as they can be applied in a large scale manner at relatively low computational costs[7,8]. Computational ML approaches to HS prediction tend to fall into two broad categories: i) sequence-based methods which use an encoding of sequence-derived features of the residues and their neighbours and then explore amino-acid identity, physicochemical properties of amino-acids, predicted solvent accessibility, Position-Specific Scoring Matrices (PSSMs), conservation in evolution and interface propensities; and ii) structure-based methods that use an encoding of structure-based features of the target residues and neighbours such as propensities at interface and surface, interface size, geometry, chemical composition, roughness, SASA, atomic interactions, among others[1–10]. Furthermore, both categories can be combined in some methods[8]. A detailed review of current ML algorithms applied to HS detection can be found in Moreira's review[5].

[1]CNC - Center for Neuroscience and Cell Biology; Rua Larga, FMUC, Polo I, 1°andar, Universidade de Coimbra, 3004-517, Coimbra, Portugal. [2]Bijvoet Center for Biomolecular Research, Faculty of Science - Chemistry, Utrecht University, Utrecht, 3584CH, The Netherlands. [3]Centro de Ciências e Tecnologias Nucleares, Instituto Superior Técnico, Universidade de Lisboa, Estrada Nacional 10 (ao km 139,7), 2695-066, Bobadela LRS, Portugal. [4]Department of Genetics and Genomics and Icahn Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [5]CMUP/FCUP, Centro de Matemática da Universidade do Porto, Faculdade de Ciências, Rua do Campo Alegre, 4169-007, Porto, Portugal. Irina S. Moreira and Panagiotis I. Koukos contributed equally to this work. Correspondence and requests for materials should be addressed to I.S.M. (email: irina.moreira@cnc.uc.pt) or A.M.J.J.B. (email: a.m.j.j.bonvin@uu.nl)

SpotOnDB: a global assessment of protein-protein interfaces and binding hot-spots

Jose G. Almeida[1], Alexandre M.J.J. Bonvin[2], Irina S. Moreira[1,2,*]

[1] CNC - Center for Neuroscience and Cell Biology; Rua Larga, FMUC, Polo I, 1ºandar, Universidade de Coimbra, 3004-517, Coimbra, Portugal.

[2] Bijvoet Center for Biomolecular Research, Faculty of Science - Chemistry, Utrecht University, Utrecht, 3584CH, the Netherlands

Abstract

Understanding protein-protein interfaces is crucial to explain complex formation and determine the foundations that governs biological networks. One key aspect is the existence and prevalence of Hot-Spots (HS), residues which, upon alanine mutation, negatively impact the formation of protein-protein complexes. While several protein complexes have been individually studied regarding their interface and the existence of HS, studies comprising high amounts of data would be highly valuable in revealing relevant patterns. In this work, we use our computational pipeline, SpotOn, to determine several structural and sequence-related features of all predicted HS in the complexes of the non-redundant database PPI4DOCK (3.746 HS for 1.403 complexes with 66.710 interfacial residues). The resulting big data analysis, which is available as an interactive online database at http://milou.science.uu.nl/services/SPOTONDB/, provides insights into HS structural and physico-chemical characteristics in protein-protein interfaces.

Keywords

Protein Complexes; Big data; Functional and structural Features; Evolutionary Features

Introduction

Protein-protein complexes are fundamental for a variety of biological functions ranging from hormone-receptor interactions [1] to innate immunity [2, 3]. One of the key features of a Protein-Protein-Interface (PPI) is its rich and diverse energetic landscape, featuring sites of high importance for complex formation (Hot-Spots: HS) [4], little importance (Null-Spots: NS) and, more recently, sites which are thought to be

COMPUTATIONAL STUDIES OF G-PROTEIN COUPLED RECEPTOR COMPLEXES: STRUCTURE AND DYNAMICS

Ozge Sensoy[1], Jose G. Almeida[2], Javeria Shabbir[1], Irina S. Moreira[2,3], Giulia Morra[4,5,*]

1 Istanbul Medipol University, The School of Engineering and Natural Sciences, 34810, Istanbul, Turkey

2 CNC - Center for Neuroscience and Cell Biology; Rua Larga, FMUC, Polo I, 1ºandar, Universidade de Coimbra, 3004-517, Coimbra, Portugal.

3 Bijvoet Center for Biomolecular Research, Faculty of Science - Chemistry, Utrecht University, Utrecht, 3584CH, The Netherlands

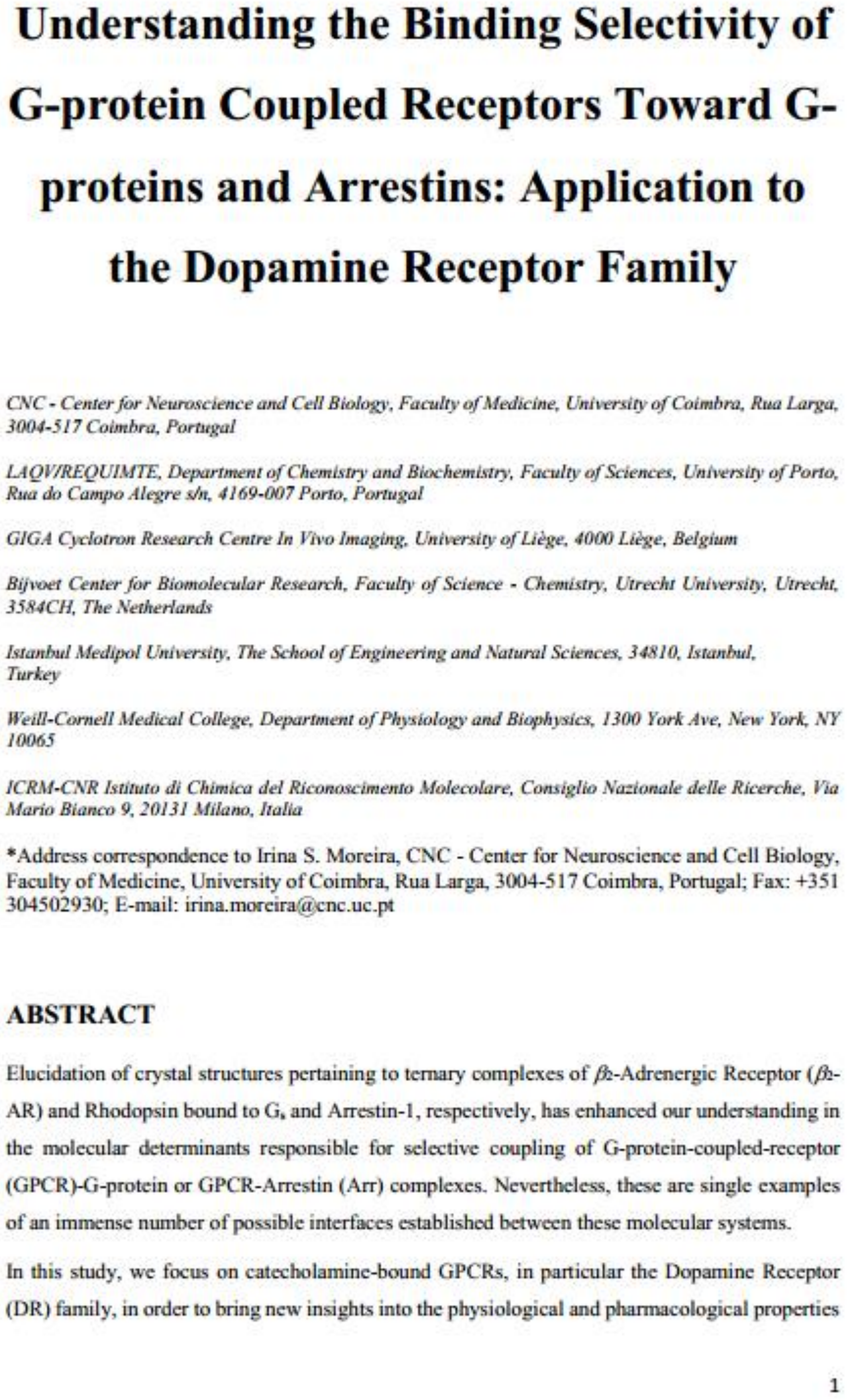
4 Weill-Cornell Medical College, Department of Physiology and Biophysics, 1300 York Ave, New York, NY 10065

5 ICRM-CNR Istituto di Chimica del Riconoscimento Molecolare, Consiglio Nazionale delle Ricerche, Via Mario Bianco 9, 20131 Milano, Italia

ABSTRACT

G-protein coupled receptors (GPCRs) are ubiquitously expressed transmembrane proteins associated with a wide range of diseases such as Alzheimer's, Parkinson, schizophrenia and also implicated in in several abnormal heart conditions. As such, this family of receptors is regarded as excellent drug targets. However, due to the high number of intracellular signaling partners, these receptors have a complex interaction networks and it becomes challenging to modulate their function.

Experimentally determined structures give detailed information on the salient structural properties of these signaling complexes but they are far away from



Understanding the Binding Selectivity of G-protein Coupled Receptors Toward G-proteins and Arrestins: Application to the Dopamine Receptor Family

CNC - Center for Neuroscience and Cell Biology, Faculty of Medicine, University of Coimbra, Rua Larga, 3004-517 Coimbra, Portugal

LAQV/REQUIMTE, Department of Chemistry and Biochemistry, Faculty of Sciences, University of Porto, Rua do Campo Alegre s/n, 4169-007 Porto, Portugal

GIGA Cyclotron Research Centre In Vivo Imaging, University of Liège, 4000 Liège, Belgium

Bijvoet Center for Biomolecular Research, Faculty of Science - Chemistry, Utrecht University, Utrecht, 3584CH, The Netherlands

Istanbul Medipol University, The School of Engineering and Natural Sciences, 34810, Istanbul, Turkey

Weill-Cornell Medical College, Department of Physiology and Biophysics, 1300 York Ave, New York, NY 10065

ICRM-CNR Istituto di Chimica del Riconoscimento Molecolare, Consiglio Nazionale delle Ricerche, Via Mario Bianco 9, 20131 Milano, Italia

*Address correspondence to Irina S. Moreira, CNC - Center for Neuroscience and Cell Biology, Faculty of Medicine, University of Coimbra, Rua Larga, 3004-517 Coimbra, Portugal; Fax: +351 304502930; E-mail: irina.moreira@cnc.uc.pt

ABSTRACT

Elucidation of crystal structures pertaining to ternary complexes of β_2-Adrenergic Receptor (β_2-AR) and Rhodopsin bound to G$_s$ and Arrestin-1, respectively, has enhanced our understanding in the molecular determinants responsible for selective coupling of G-protein-coupled-receptor (GPCR)-G-protein or GPCR-Arrestin (Arr) complexes. Nevertheless, these are single examples of an immense number of possible interfaces established between these molecular systems.

In this study, we focus on catecholamine-bound GPCRs, in particular the Dopamine Receptor (DR) family, in order to bring new insights into the physiological and pharmacological properties

H. Annex References

1 Nielsen, M. *Neural Networks and Deep Learning*. (2015).

2 Hinton, G. E., Osindero, S. & Teh, Y.-W. A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation* **18**, 1527-1554, doi:10.1162/neco.2006.18.7.1527 (2006).

3 Salakhutdinov, R. & Hinton, G. E. Deep Boltzmann Machines. *Proceedings of The 12th International Conference on Artificial Intelligence and Statics*, 448-455, doi:10.1109/CVPR.2009.5206577 (2009).

4 Deepa, S. N. & Devi, B. A. Modified Radial Basis Function Network for Brain Tumor Classification. *Swarm, Evolutionary, and Memetic Computing, Pt I* **7076**, 366-371 (2011).

5 Spencer, M., Eickholt, J. & Jianlin, C. A Deep Learning Network Approach to ab initio Protein Secondary Structure Prediction. *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM* **12**, 103-112, doi:10.1109/TCBB.2014.2343960 (2015).

6 Heffernan, R. *et al.* Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. *Sci Rep* **5**, 11476, doi:10.1038/srep11476 (2015).

7 Carneiro, G., Nascimento, J. & Bailey, A. P. Unregistered Multiview Mammogram Analysis with Pre-trained Deep Learning Models. *Medical Image Computing and Computer-Assisted Intervention* **9351**, 652-660 (2015).

8 Liang, M., Li, Z., Chen, T. & Zeng, J. Integrative Data Analysis of Multi-Platform Cancer Data with a Multimodal Deep Learning Approach. *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM* **12**, 928-937, doi:10.1109/TCBB.2014.2377729 (2015).

9 Hua, K. L., Hsu, C. H., Hidayati, S. C., Cheng, W. H. & Chen, Y. J. Computer-aided classification of lung nodules on computed tomography images via deep learning technique. *OncoTargets and therapy* **8**, 2015-2022, doi:10.2147/OTT.S80733 (2015).

10 Wang, S., Peng, J., Ma, J. & Xu, J. Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields. *Sci Rep* **6**, 18962, doi:10.1038/srep18962 (2016).

11 Pang, S. & Yang, X. Deep Convolutional Extreme Learning Machine and Its Application in Handwritten Digit Classification. *Computational intelligence and neuroscience* **2016**, 3049632, doi:10.1155/2016/3049632 (2016).

12 Shimizu, R. *et al.* in *2016 International SoC Design Conference (ISOCC).* 191-192.

13 Nie, D., Zhang, H., Adeli, E., Liu, L. & Shen, D. 3D Deep Learning for Multi-modal Imaging-Guided Survival Time Prediction of Brain Tumor Patients. *Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention* **9901**, 212-220, doi:10.1007/978-3-319-46723-8_25 (2016).

14 Yuan, Y. *et al.* DeepGene: an advanced cancer type classifier based on deep learning and somatic point mutations. *BMC Bioinformatics* **17**, 476, doi:10.1186/s12859-016-1334-9 (2016).

15 Vougas, K. N. *et al.* Deep Learning and Association Rule Mining for Predicting Drug Response in Cancer. *bioRxiv*, doi:10.1101/070490 (2016).

16 Kelley, D. R., Snoek, J. & Rinn, J. L. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res* **26**, 990-999, doi:10.1101/gr.200535.115 (2016).

17 Kadurin, A. *et al.* The cornucopia of meaningful leads: Applying deep adversarial autoencoders for new molecule development in oncology. *Oncotarget*, doi:10.18632/oncotarget.14073 (2016).

18 Angermueller, C., Parnamaa, T., Parts, L. & Stegle, O. Deep learning for computational biology. *Molecular systems biology* **12**, 878, doi:10.15252/msb.20156651 (2016).

19 Wang, S., Sun, S., Li, Z., Zhang, R. & Xu, J. Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. *PLoS Comput Biol* **13**, e1005324, doi:10.1371/journal.pcbi.1005324 (2017).

20 Eddy, S. R. Hidden Markov models. *Curr Opin Struct Biol* **6**, 361-365 (1996).

21 Peterson, L. E. K-nearest neighbor. *Scholarpedia* **4**, 1883, doi:doi::10.4249/scholarpedia.1883 (2009).

22 Hosmer, D. W. & Wang, C. Y. Stepwise Logistic Regression. *Biometrics* **34**, 158-158 (1978).

23 Ho, T. K. Vol. 1 278-282 vol.271 (1995).

24 Drew Conway, J. M. W. *Machine Learning for Hackers*. (1989).

25 Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: Synthetic minority over-sampling technique. *J Artif Intell Res* **16**, 321-357 (2002).

26 Blagus, R. & Lusa, L. SMOTE for high-dimensional class-imbalanced data. *Bmc Bioinformatics* **14**, doi:Artn 106

10.1186/1471-2105-14-106 (2013).

27 Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Research* **28**, 235-242 (2000).

28 Yu, J. & Guerois, R. PPI4DOCK: large scale assessment of the use of homology models in free docking over more than 1000 realistic targets. *Bioinformatics (Oxford, England)* **32**, 3760-3767, doi:10.1093/bioinformatics/btw533 (2016).

29 Moreira, I. S. *et al.* SpotOn: a web server for protein-protein binding hot-spots. *Sci. Rep. (accepted)* (2017).

30 Munteanu, C. R. *et al.* Solvent accessible surface area-based hot-spot detection methods for protein-protein and protein-nucleic acid interfaces. *J Chem Inf Model* **55**, 1077-1086, doi:10.1021/ci500760m (2015).

31 Melo, R., Fieldhouse, R., Melo, A., Correia, J. D. G. & Cordeiro, M. N. D. S. A Machine-Learning Approach for Hot-Spot Detection at Protein-Protein Interfaces. *IJMS*, 1-16 (2016).

32 Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25** (1997).

33 Xiao, N., Cao, D. S., Zhu, M. F. & Xu, Q. S. protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. *Bioinformatics (Oxford, England)* **31**, 1857-1859, doi:10.1093/bioinformatics/btv042 (2015).

34 Liu, Q., Li, Z. & Li, J. Use B-factor related features for accurate classification between protein binding interfaces and crystal packing contacts. *BMC Bioinformatics* **15 Suppl 16**, S3, doi:10.1186/1471-2105-15-S16-S3 (2014).

35 Ashkenazy, H. *et al.* ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Res* **44**, W344-350, doi:10.1093/nar/gkw408 (2016).

36 Sensoy, O., Almeida, J. G., Shabbir, J., Moreira, I. S. & Morra, G. in *Molecular Biology Protocols* (Submitted).

37 Almeida, J. G., Preto, A. J., Koukos, P. I., Bonvin, A. & Moreira, I. S. MEMBRANE PROTEINS STRUCTURES: A review on computational modeling tools. *Biochim Biophys Acta*, doi:10.1016/j.bbamem.2017.07.008 (2017).

38 Lemos, A. *et al.* In silico studies targeting G-protein coupled receptors for drug research against Parkinson’s disease. *Current Neuropharmacology* (Submitted).