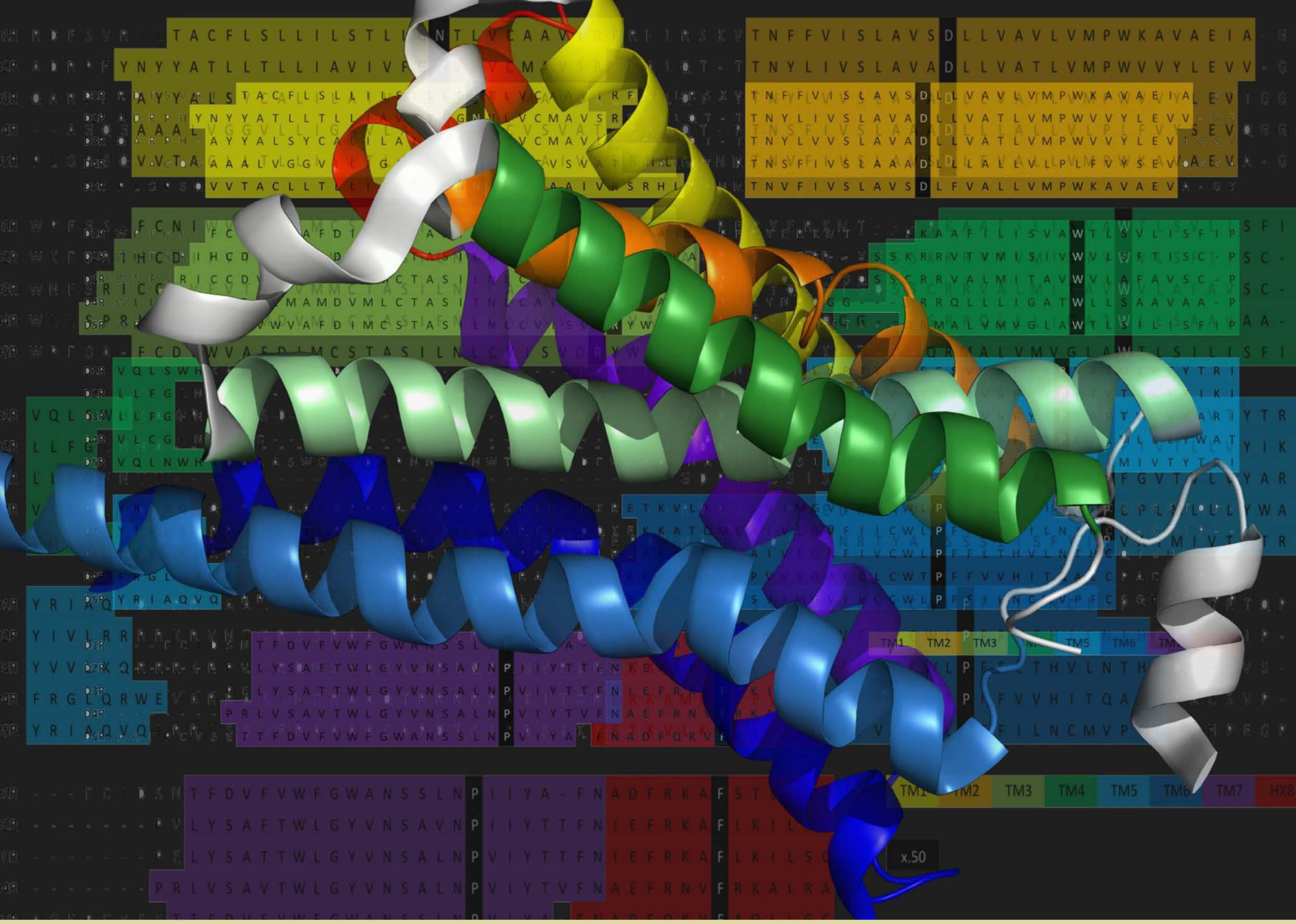


António José Preto Martins Gomes

A bioinformatics approach for the understanding of membrane protein complexes

Dissertação de Mestrado em Bioquímica, orientada por Professora Doutora Irina Moreira
e apresentada ao Departamento de Ciências da Vida da Faculdade de Ciências e Tecnologia da Universidade de Coimbra

Agosto 2017

• U C •

Universidade de Coimbra

António José Preto Martins Gomes

A bioinformatics approach for the understanding of membrane protein complexes

Dissertação apresentada para provas de Mestrado em Bioquímica

Orientadora: Professora Doutora Irina S. Moreira

Co-orientadora: Professora Doutora Ana Luísa Carvalho

Setembro 2017

Universidade de Coimbra

· U C ·

Universidade de Coimbra

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

First of all, I would like to thank my mother, for giving me the opportunity to study what I want, and permanently being supportive, as well as my sister. I would also like to thank my godfather, since he was my example and motivation figure into attempting scientific research work.

I thank Irina Moreira for the amazing opportunity of working with daily purpose with an amazing group. I would also like to thank Agostinho Lemos and Rita Melo for being helpful and thoughtful throughout the whole journey. I will also thank Joel Arrais, since he gave me the opportunity, while in Bachelor’s Degree, to get in contact with an area I now want to explore more than ever.

My friends of ever, José Pedro, João Pedro, João Manuel, Miguel Oliveira, Henrique Ferrolho and Guilherme Routar, because they have always been there, through the best and the worst.

The people I got to know in Coimbra, who I cherish deeply, Bernardo Carvalho, Ana Rita, André Areal, Margarida Almeida, David Bidarra, Gonçalo Teixeira, Pedro Filipe, João Gonçalo, Nuno Piedade, Nuno Chicória, Cláudio Costa, Daniela Rosa, Guilherme Frias, João Coelho, David Navega.

Finally, I thank specially José Guilherme Almeida, since he was present throughout a whole year and we got to share work, a Junior Enterprise and the start of a life long friendship.

RESUMO

Os problemas das ciências da vida remetem inúmeras vezes para a caracterização da estrutura e função das moléculas biológicas, nomeadamente, as proteínas. A caracterização estrutural de proteínas solúveis, proteínas membranares e complexos proteicos diversos tem vindo a beneficiar largamente de contribuições da área computacional. Particularmente no caso das proteínas membranares, pela sua quase ubiquidade de funções e pela dificuldade do seu estudo de forma experimental, a aplicação de métodos computacionais de previsão de estrutura e caracterização de interface permite o seu estudo de forma alargada e detalhada, sempre que possível apoiando-se em dados experimentais.

Nesta tese de mestrado dividimos o trabalho em 2 vertentes: i) a compreensão de um exemplo típico de proteína membranar – complexos acoplados a proteína G (GPCR) e ii) o desenvolvimento de metodologias para caracterização de interações proteicas. São assim expostos resultados de modelação de complexos entre GPCRs, arrestinas e proteínas G, progredindo para a caracterização de área de superfície, proximidade de resíduos, conservação de resíduos, ligações de hidrogénio, entre outras características. Desta forma, identificam-se subestruturas, regiões, padrões de resíduos e resíduos específicos determinantes para a formação de complexos entre os GPCRs e outras proteínas. A informação foi sintetizada sob a forma gráfica e disponibilizada online (http://45.32.153.74/gpcr/). Na segunda parte dedicamo-nos à aplicação de algoritmos de *Machine-Learning* de forma a corretamente classificarmos os resíduos da interface como cruciais a nível estrutural e funcional da proteína, os *Hot-Spots*. Com base em dados experimentais e recolha de descritores destes resíduos, foi construído um modelo de previsão de *Hot-spots*, implementado num portal de acesso livre (http://milou.science.uu.nl/cgi/services/SPOTON/spoton/). Procuramos também perceber como a inclusão de descritores coevolutivos das proteínas influencia a performance das metodologias desenvolvidas.

PALAVRAS-CHAVE

Proteínas membranares; Interfaces proteína-proteína; *hot-spots*; modelação computacional; *machine-learning*, portais de web; GPCRs; coevolução

ABSTRACT

In Life science's it is crucial to characterize the structure and function of biological molecules, in particular, proteins. The structural characterization of soluble proteins, membrane proteins and diverse protein complexes, as of late, has benefitted from computational approaches' contributions. These would be essential in the case of membrane proteins. Due to their near functional ubiquity and the difficulty of their experimental study, the employment of structure prediction and interface characterization computational methods allows for broad and in-depth comprehension, which fundamental if we realize membrane proteins are key targets in the pharmaceutical industry.

In this master thesis, the works was divided in two different but interconnect tasks: i) the understanding of a typical example of membrane protein: G-coupled protein receptor complexes (GPCR) and ii) the methodological development for protein interaction characterization. As a results of the first task we list all conclusions retrieved from the modelling of complexes between GPCRs, arrestins and G-proteins with the full assessment of the surface area inter-residual distance, residue conservation, hydrogen bonds, among other characteristics. Subsequently, we proceed to identify substructures, regions, residue patterns and specific residues' relevant for complex formation between GPCRs and other proteins. The data was summarized under graphical display and made available online (http://45.32.153.74/gpcr/). In the second part, we focused on Machine-Learning algorithms' deployment in order to correctly classify protein interfacial residues, which are crucial at a structural and functional level: Hot-Spots. Based on experimental data and residue feature collection, a Hot-Spot prediction model was built, available at a free-access portal (http://milou.science.uu.nl/cgi/services/SPOTON/spoton/). We also sought to understand how the inclusion of coevolutionary features influences the performance of the developed methodologies.

KEYWORDS

Membrane proteins; protein-protein interfaces; hot-spots; computational modeling; machine-learning, web-portals; GPCRs, coevolution

Table Index

Figure Index

Equation Index

LIST OF ABBREVIATIONS

AAC: Amino Acid Composition

AAindex: Amino Acid index database

aPAAC: amphiphilic Pseudo Amino Acid Composition

ANN: Artificial Neural Network

Arrs: Arrestins

ATP: Adenosine TriPhosphate

AUROC: Area Under Receiving Operating Characteristic

BID: Binding Interface Database

BLOSUM: BLOcks Substitution Matrix

CNS: Central Nervous System

COCOMAPS: bioCOmplexes COntact MAPS

Cryo – EM: Cryo – Electron Microscopy

D1-5R: Dopamine Receptors 1 to 5

DCA: Direct Coupling-Analysis

EBI: European Bioinformatics Institute

ECL: ExtraCellular Loop

EPR: Electron Paramagnetic Resonance

FDR: False Discovery Rate

GPCR: G-Protein Coupled Receptor

HADDOCK: High Ambiguity Driven biomolecular DOCKing

HB: Hydrogen Bonds

HMM: Hidden Markov Model

HS: Hot-Spots

HVR: High Variability Regions

HX8: non-membrane embedded HeliX on GPCRs

ICL: IntraCellular Loop

kNN: k-Nearest Neighbours

logit: logistic regression

MARS: Multivariate Adaptive Regression Splines

McBASC: McLachlan-Based Substitution Correlation

mfDCA: mean-field Direct Coupling-Analysis

MI: Mutual Information

ML: Machine Learning

MSA: Multiple Sequence Alignment

MP: Membrane Proteins

NMR: Nuclear Magnetic Resonance

NPV: Negative Predictive Value

NS: Null-Spot

ORF: Oblique Random Forest

PAAC: Pseudo Amino Acid Composition

PCA: Principal Component Analysis

PDA: Penalised Discriminant Analysis

PCM: ProteoChemoMetric

PINT: Protein-protein INteractions Thermodynamic database

PLR: Penalised Logistic Regression

PMP: Peripherally Membrane-bound Proteins

PPI: Protein-Protein Interactions

PPSM: Position-Specific Scoring Matrix

PPV: Positive Predictive Value

PSICOV: Protein Sparse Inverse COVariance

PTM: Post Translational Modifications

RF: Random Forest

RMSD: Root-Mean-Square Deviation

SASA: Solvent Accessible Surface Area

SB: Salt Bridges

SKEMPI: Structural Kinetic and Energetic database of Mutation Protein Interactions

SMOTE: Synthetic Minority Over-sampling Technique

ssNMR: solid-state Nuclear Magnetic Resonance

SVM: Support Vector Machine

TM: TransMembrane

TNR: True Negative Rate

TPR: True Positive Rate

VMD: Visual Molecular Dynamics

1. INTRODUCTION

Proteins are biomolecules essential for life, performing a broad array of tasks essential for organism maintenance. Proteins can interact with other proteins, molecules and, in some cases, smaller particles. In order to understand this process, it is necessary to clarify their building blocks – amino acids – as well as the way these occupy space under different conditions: protein structure, which is highly related to their function.

Proteins' structures are flexible, and in fact, their mobility is an essential characteristic that allows them to be so important in the tasks they complete [1]. Protein conformations, acquired through folding processes, are essential for their function [2], for this reason, deep understanding on proteins' structure under different states and conformations is essential in order to intervene in their functional behaviors, which is a desirable outcome either for disease treatment or cell enhancement.

1.1 Protein interface and Hot-spots

Protein-based coupling occurs by specific interfacial residues with chemical-physical properties different from the remaining of their surface. As such, interfaces, protein core and non-interface surfaces have different amino acid composition, physicochemical properties, secondary structure which, ultimately, different solvent accessibility [3]. The establishment of key interactions fundamental from a structural and functional point of view makes them a prime study target, since they tend to mediate the proteins' biological activity [4]. Protein interfaces are also commonly participants on PPIs of different types such as the formation of homodimeric, heterotrimeric, enzyme-inhibitor complexes, among many other [5]. To perform these tasks, protein interfaces require certain characteristics, as a minimum accessible surface area, hydrophobic profile, solvation potential and protrusion [6]. Protein interface characterization requires several metrics, which motivates the use of ML approaches for their recognition. Nevertheless, some such as evolutionary conservation and occlusion from the solvent seem to be particularly relevant [4].

Although PPI can vary in size, there are particular residues that seems to be mainly responsible for the actual coupling: the so-called HS. Therefore, HS are highly conserved residues that tend to be particularly relevant for the establishment of interactions, contributing more significantly to binding affinity than other residues [7]. Overall, a HS is defined as a residue whose mutation to alanine decreases binding free energy ($\Delta\Delta G_{binding}$) in at least 2.0 kcal mol $^{-1}$ [8], if this criteria is not met, the residue is designated as a Null-Spot. Several approaches for HS prediction have been proposed, and more recently ML-based methods seem to have a higher degree of sucess [9, 10] [11-13].

1.2 Membrane proteins

Membrane proteins (MP) are proteins embedded in lipid environments, frequently lipid bilayers, that perform a large number of tasks [14]. These proteins act frequently as messengers between the intracellular and the extracellular environment, which makes them indispensable for cell life function and, consequently, for organism maintenance [15]. Assuming an indispensable role on the communication between cells and organs, they also feature important functions such as cell cycle life regulation (conversely, also apoptosis) [16], ion and molecule transport, immune system molecule recognition and energy transduction [17-19]. Ultimately, these and other functions give rise to many biological functions, such as sugar blood regulation through insulin identification by tyrosine kinases [20] and neuronal communication, necessary and constituent for brain functioning [21] . Having in common the association with the lipid environment, not all MPs do so in the same way. While some permeate the membrane (intrinsic/transmembrane - TM), others are peripherally membrane-bound proteins (PMP) [22-24]. In both cases, the lipid environment can have significant structural changes and greatly determine MPs mobility along the membrane [25].

The same MP can have its substructures differently characterized; those that occupy space on the outer side of the membrane are called extracellular while the ones inside the cell are intracellular. The residues spanning inside the membrane define intramembrane structures. All these structures can vary among MPs and perform different functions [26]. The residue content of the MPs' structures varies depending on their membrane relative location, particularly due to the electronegativity profiles of the environment; sections of the protein inside of the membrane tend to be rich in hydrophobic residues, matching the lipid chemical profile, while those outside of the membrane tend to have more hydrophilic residues, since they are in closest contact with water molecules. Thus, TM proteins are usually referred to as being amphipathic, which describes their irregularity of chemical profiles along their structures [27].

Furthermore, although the primary structure of the protein is highly determinant, the structure and function of many TM proteins depend also on PTM, which are modifications that introduce changes non-dependent on the residue content alone, hence, after the translation process is concluded, examples are phosphorylation and glycosylation, both acting by adding groups to certain residues of the protein [28]. Regarding the secondary structure, the two major recurrent protein structure motifs in MPs are TM α-helices [29], repeatedly crossing the membranes in α-helical bundles and β-strands arranged into super-secondary structures known as β -barrels [30].

The most functionally relevant intrinsic MPs are typically split into ion channels, membrane receptors and transporters [17, 31]. Ion channels generate a connection between the extracellular and intracellular space that is susceptible to the passage of ions, and crosses both a physical barrier and the electrochemical gradient inbound to the separation by the membrane. Their structure can be modulated by the TM electrochemical potential, the binding of ligands, and mechanical stress and/or changes in the local lipid environment [32]. Transporters can move molecules or ions across the membrane, however differently than channels, usually working through conformational changes, being very important in the transport of ligands against electrochemical gradient, by using ATP breakage energy [33]. Membrane receptors comprise a vast amount of proteins, among which the superfamily of GPCRs [26], these, due the scope of this work, will further be individually explored.

1.1.1 G-protein coupled receptors

GPCRs are membrane receptors responsible for many different functions, belonging to one of the largest superfamilies of membrane associated proteins [34]. GPCRs share a typical pattern consisting of seven TM helixes (TM1-7) connected by three ICLs in the cytoplasm and three ECLs on the outer side of the lipid membrane. GPCRs terminate with an helix that spans parallelly to the membrane (HX8) [35], which has been shown to participate in modulating the interaction between the receptor and its intracellular partners such as Postsynaptic density protein, Drosophila disc large tumor suppressor, and Zonula occludens-1 protein (PDZ) domain-containinG-proteins [36, 37]. The amino and the carboxyl termini of class A GPCRs reside in the extracellular and intracellular part of the cell, respectively [38, 39].

An example of a GPCR structure can be seen in Figure 1, produced with PyMOL [40] from the PDB entry 3SN6 [41]. Three High Variability Regions (HVR) have been identified: between TM5 and TM6 (ICL3) and at the N- and C-terminal regions [39] [42]. Even though GPCRs share high structural similarity, their ligands can range from a photon to a protein [38]. GPCRs can receive distinct stimuli, having roles on metabolic, neuronal, hormonal and immunological functions, as well as in cell growth and cell death [43].

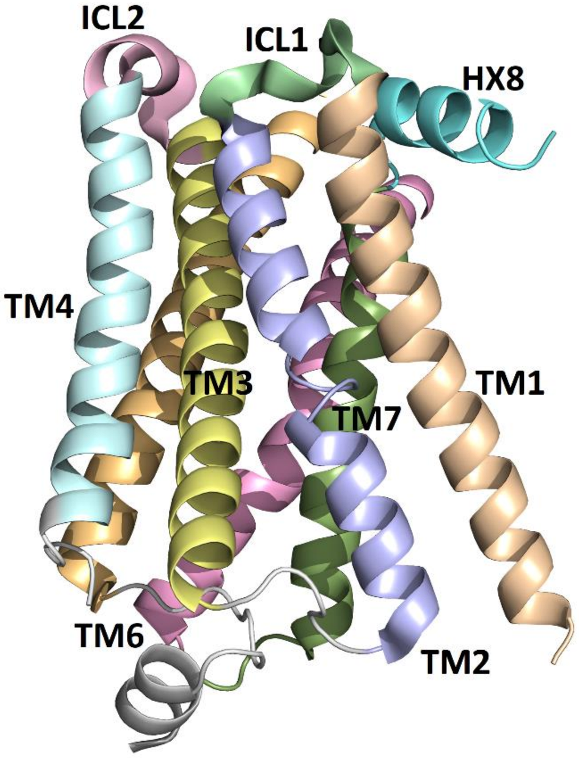


Figure 1: GPCR model example

GPCRs are commonly divided in five families, regarding their sequence similarity on the TM (most conserved regions), these families are: rhodopsin (class A), secretin (B1), adhesion (B2), glutamate (C) and frizzled/taste (F) [35]. A large number of GPCRs are olfactory receptors. GPCRdb [44] is a database that amasses a great amount of updated information on GPCRs, namely among the referred families, also providing several tools useful on the understanding of such receptors as well as their interactions with both orthosteric and allosteric ligands.

GPCRs play a central role in a large variety of cellular mechanisms in human physiology and disease and are the targets of 40% of all commercialized drug targets. As such, they are the subject of major efforts towards understanding their function and signaling selectivity [45]. New insights have been provided by recent GPCR structures in selected conformations, stabilized by a variety of ligands with pharmacologically distinct properties (agonists, inverse agonists, etc.), by nanobodies mimicking signal transducers [46], and in some cases by full heterotrimeric G-proteins (GTP-bindinG-protein) [47, 48]. GPCRs have similar intracellular binding partners such as G-proteins, arrestins and GPCR-interacting G-proteins (GIPs), membrane-inserted GPCR-binding G-proteins [49]. These play an important part on the structural rearrangement of the GPCR structure, and, consequently, their activation state and function. Arrestins, for instance, are responsible for the desensitization of GPCRs [50]. The binding of these proteins greatly increases the difficulty of G-protein binding to GPCRs and provokes the internalization of the complexes [51]. G-proteins, on the other hand, are the heterotrimeric proteins (constituted by three subunits) that, through the coupling with GPCRs, allow the triggering of signaling cascades that propagate into the cell [52]. The lipid membrane environment also has an active role in modulating GPCR structure and function. For example, interaction with cholesterol significantly changes GPCRs conformational flexibility [53] 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. Constitutive internalization of GPCRs, a crucial cellular function responsible for receptor regulation, is regulated by GPCR interactions and can be clathrin-dependent or clathrin-independent, stressing the large array of interactions and the versatility of GPCRs [54]. Trafficking of GPCRs, which can be agonist dependent or independent, commonly displays an important role on the signaling routes these receptors are involved in [54].

Dopamine receptors are class A GPCRs present in many neurons in the central nervous system (CNS), reason why their understanding is commonly important for the comprehension and treatment of several neurological disorders. These receptors are highly specific for dopamine also interacting with other related ligands that exert their physiological and pharmacological effects through the activation of five distinct but closely related subtypes of DR complexes, which are divided into two major subclasses (D1-like receptors: D1R and D5R; D2-like receptors: D2R, D3R, and D4R), based on their ligand and G-protein-subtype specificity, anatomical distribution and physiological effects [55-57]. When considering drug design against GPCRs, studying the differences between the active and inactive state [58] is important, as is the case for Dopamine receptors 2 and 4 (D2R and D4R, respectively) [59], as these changes can be determinant in drug/receptor interactions, and therefore crucial for the design of new drugs [60, 61].

1.1.2 Lipid environment

When considering MPs, the lipid environment is essential in defining their structure and function, often significantly changing the proteins' properties [62]. MPs' association with the membrane is what makes the task to study them harder than soluble proteins. Having that in mind, although most MP structures are not easily determined, it is useful to note that some MPs can retain their structure and function while in soluble form. The construction of fusion proteins [63], and other strategies are employed to overcome the difficulties of their study. When this is not possible, detergents can be used to solubilize the expressed proteins [64] by extracting them from the membrane, ideally without affecting their structure.

Membrane domains, such as lipid rafts, can change significantly the structure and function of some proteins as these domains have different properties (namely high glycosphingolipids content) [65, 66]. In lipid rafts, solvent extraction can be less effective, since these are more effective at retaining MPs than other lipid membrane domains. This works either by surrounding the protein with a tighter and more ordered lipid packing, or by other mechanisms, such as anchoring [67]. Furthermore, even when not considering lipid rafts or lipid raft-like domains, other lipid structures and molecular organizations (depending on factors such as temperature, pressure, lipid composition and other proteins) can influence the membrane structure, which, in turn, can affect membrane-inserted proteins. This is usually referred to as lipid polymorphism, to which distinct lipid phases are associated, and which has been observed to play a role in G-protein structure and function [68]. Some intrinsic protein properties such as hydrophobicity, van der Waals interactions, prosthetic groups, among others, can play a major role in the interaction between the protein and the membrane. Hydrophobic mismatch, for instance, occurs when the thickness of the bilayer's hydrophobic section does not correlate with the length of the hydrophobic residues of the membrane, generating a mismatch, as characterized for example by calorimetry [69], NMR [70] and fluorimetry techniques [62, 71-73].

Further changes in the membrane can occur upon insertion and formation of dimers or even high-order oligomers, for example, which contributes towards the complexity of MP-membrane interactions. Other relevant changes are the insertion of peripheral groups (adding a step to the usual two step model considered for MPs' inclusion and dimerization/oligomerization) such as prosthetic groups, more elaborate protein folding, generation of new binding surfaces or portioning of space away from the lipid. This can be studied through a combination of kinetic analysis and NMR [62, 72, 74].

1.3 Experimental membrane protein structure determination

The study of MPs is highly reliant on the available structures, which, due to the influence of the membrane, are difficult to attain [75, 76]. Such influence is expressed through various specific factors, such as cholesterol content [77, 78] and hydrophobic thickness of the lipid bilayer [62, 65, 67, 68, 79-82]. The membrane-embedded sections of the protein are hard to determine since the membrane induces changes on the structure, which diminish drastically the accessibility to the methods commonly used to experimentally determine protein structures on soluble proteins. Another difficulty is the expression of MPs in laboratory systems in such a way that the structure is similar to that of the actual proteins. An indicator of the relevance of this problem is that only 4.193 structures of membrane proteins (or rather mainly of extracellular sub-domains) can be found among the 131.485

determined protein structures deposited at the PDB [83], which adds up to around 1% of the total protein structures available.

Adding up to the problems referred, new difficulties arise when considering each of the three main methods used for protein structure experimental determination: X-ray crystallography, NMR and Cryo – EM. X-ray crystallography solved structures amount to the larger number of protein structures determined by the same method. These experiments require a large amount of time to prepare and optimize. Establishing proper crystallization conditions is the main challenge, particularly regarding membrane proteins and, when such is achieved, further optimization is required [84].

Considering the protein to be expressed as a soluble protein, detergent/solvent can be added to induce lipid-like transformations on the MP, preferably with the use of different 3D continuous lipid phases (allowing the protein to freely flow) [85]. Distinct detergents, with different hydrophobicity properties, can be used depending on the protein's properties [86]. The choice of the detergents can be time and resource consuming, with no guaranteed results [86, 87]. The use of detergents leads to micelle-like structure formation, which is not an accurate representation of the bilayer environment and can result in deformations in structure. Some approaches to overcome these problems include the inclusion of MPs in nanodiscs – detergent free membrane-like structures stabilized by polymers or proteins, which allow for liquid-state NMR studies [88] – and the lipid cubic phase method [89]. The latter works by isolating a biological membrane with the target protein and solubilizing it with detergent. The resulting micelle is purified and homogenized with monoacylglycerol, and contains a bilayer with the target protein [90].

Another approach is the use of antibody fragments to stabilize the protein structure [91]. The latter often results in more stable crystals, but the MP conformation might differ from its native state due to the additional interactions with the antibody fragments. The previous approaches, although closing the gap from the difference in the expressed protein to the native structure, do not prevent data collection and analysis from being difficult, as the variability of crystals and their conditions (hydrophobic protein regions camouflaged by hydrophobic solvent, making it difficult to assess the transmembrane MP structure) might prevent automated and stable data acquisition and processing [84].

Differently from X-ray crystallography, NMR spectroscopy does not rely on the incidence and measurement of X-rays, but rather nuclear spin derived positional calculations. The difference of approach makes it possible to measure more accurately the membrane-embedded part of the protein, due to the possibility of adjusting the measurements to the spins of characteristically MP associated residue atoms, as well as previously marking the proteins with radioactive isotopes. However, this approach has low sensitivity, size limitation and does not measure the intrinsic motions of the system under investigation as accurately as would be preferable.

Soluble NMR, performed with similar sample preparation as in X-ray crystallography, faces sample preparation issues, as well as spectral crowding, which arises from the large amount of atoms emitting signals that can interfere with each other [92]. Nonetheless, NMR has proven useful to study the dynamics (relative population and conformation of different states, exchange rates, internal motions) of MPs undergoing conformational changes [93]. Recently, new techniques such as solid state NMR (ssNMR) have provided much better results when compared to liquid phase NMR, as there is no molecular weight cap, allowing for the study of biological systems in which the protein is much closer to its native conditions [94]. However, this does not prevent spectral crowding, since a lot more undesirable signals are bound to be amassed in the measurement. Compared to X-ray crystallography, NMR, and in particular ssNMR, has the great advantage of allowing the study of MP in an actual membrane environment and not in a "detergent simulation" of a membrane [95-97].

MP structure determination has also been conducted using paramagnetic tags, a technique focused on labelling MPs with atoms which's spectral signal is known, so that they can later be analyzed with NMR [98] and/or Electron Paramagnetic resonance (EPR) [99]. Recently, it has even been demonstrated that MPs can be studied by ssNMR in their native cellular environment [100]. The two previous approaches, particularly X-ray crystallography, already have a large amount of determined structures to account for. Cryo-electron microscopy (Cryo-EM) is a more recent approach that employs the imaging of radiation-sensitive entities – cells, viruses and macromolecules – under cryogenic conditions using a transmission electron microscope, hence having a much broader sample size limit than both NMR and X-ray crystallography [101]. Unlike X-ray crystallography, it does not require crystallization. Its main drawback is the relatively low resolution for membrane proteins when compared to X-ray structures.

1.4 Membrane protein structure computational prediction methods

The difficulties arising in the experimental determination of MP structures result in large time and resources expenses, with, sometimes, no guarantees of result. Computational methods, also used for soluble protein structure prediction and having a large contribution on drug design and discovery [102], overall, are an approach to be considered for MP structure prediction. The application of these computational approaches to MP is still recent and based on the adaptation of known protocols for soluble proteins. For most cases, a simulation of the solvent must be employed, as such, this is one of the factors that must be greatly altered for MP structure prediction.

More than simply predicting the structure of MP, it is of great interest to predict complex formation between MP and soluble proteins and MP-MP, in these cases, it is again noted that, when soluble proteins are involved, the task is easier, whereas MP-MP complex formation, even when considering homodimers, is harder to conduct. Therefore, to predict MP structure, it can be helpful to consider cytosolic/extracellular partner interactions, since their interacting motifs are easier to study than those responsible for MP-MP interaction. By doing so, part of the cytosolic/extracellular regions of a protein can be determined, making it easier to identify membrane spans. Experimentally determined MP structures in different conformations may also help in achieving more accurate predictions, as some predictions might recreate only one conformation, while important interactions are also taking place in other possible conformations. Furthermore, membrane lipid composition should be considered when attempting to predict the structure of a MP as it affects, not only the conformation, but also often the activation state of membrane-embedded proteins [103].

MP tridimensional structure prediction is highly based on the availability of similar protein structures to the one that is to be predicted. If similar proteins exist, homology models are used. On the contrary, *de novo* methods are employed, in order to build models of the protein without template proteins associated [104]. Knowing both the sequence of the protein and the structure of a homologue, homology modeling provides the best results within a reasonable time-frame. Some methods have been developed specifically for MP modelling, namely MEMOIR (Membrane protein modelling pipeline), [105] which can model the 3D structure of a protein of known sequence provided there are available homologous MPs with determined 3D structures, and MEDELLER [106], taking the name from the previous and more general installment MODELLER [107] (non-specific to MPs), which has provided interesting results due 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, thus rendering the models more accurate but 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 [108], a web implementation of TMFoldRec [109]. Upon topology prediction, systematic sequence to structure alignment is performed, resulting in the selection of several templates that are ordered according to energy and reliability. Rosetta has also been widely applied to MP prediction [110]. 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 [110]. *De novo* methods, on the other hand, are employed on the absence of structural homologues, and make use of known determined features of the protein to be determined, such as secondary structure, topology information and substructure (helices and loops) information. Furthermore, approaches employing co-evolution information and ML have become increasingly more successful [111].

ML approaches work by training mathematical or logical models on a computer that, afterwards, can make use of the model to predict unknown instances of similar characteristics. The dataset is made up of instances, or samples, for which are known certain features. Regarding the instances, one or more target values (classes) are selected; the model is then trained to predict this classes for new, unknown instances. ML approaches can be supervised, if the dataset used for training has known output classes, or unsupervised, if there is no information regarding the classes. Usually, supervised learning methods tend to run faster and more accurately. Overall, ML can be defined as the automatic extraction of information from data by efficient algorithms, to discover patterns and correlations and build predictive models. ML involves the creation of algorithms that improve their own performance when undertaking a certain task based on their own experience [112]. These approaches aim to be statistically consistent, computationally efficient, and simple to implement and interpret. The choice of a ML algorithm for a specific problem should be made in light of its characteristics, deep familiarity with the theoretical foundations of the field, data source and prediction performance [113]. Dataset construction, comprising feature selection and extraction are major milestones and can condition the performance, otherwise, problems such as overfitting and underfitting can arise, although automatized approaches to avoid these issues exist and are currently being further researched. Also, the performance evaluation metrics also need to be attended to accordingly with the method and problem in question.

Regarding G-proteins, some of the referred features can be derived from the sequence, others arise from different efforts. Several groups have made advances in the number and type of features available, for instance PsiPred [114] is a broadly utilized platform for secondary structure prediction that utilizes PSSMs as inputs to an ANN approach. However, this is hardly specific for MPs. Adding hydrophobicity scales to the prediction of secondary structures, that can also be used as features, and should yield better results [115]. Initially, the utilized scales were focused on ranking single amino acids or small peptides [116], more recent advances in hydrophobicity scales include the energy of amino acids in fully folded proteins, such as the hydrophobicity scale developed by White and von Heijne [117], which was shown to deliver the best results along with scales such as the Unified Hydrophobicity Scale [116]. Other possible features to take into account are the regions of the protein that actually face the membrane, cytosolic or extracellular sides, and which are the motifs responsible for interactions, whether they are membrane-protein interactions or secondary structure-secondary structure interactions [118].

MP topology prediction by ML techniques can take into account the referred features. If such is done, it can then progress from predicting secondary structures to tertiary structures and even super-secondary structures. They are also used to predict the TM protein segments, nowadays often making use of direct residue coevolution features, which are then translated into residue-residue contacts [119-121]. A few methods managed to combine various sources of information to predict TM α-helices and α-helical bundles, as well as β-barrels. OCTOPUS [122] may be one of the most complex ML approaches for TM α-helical spans, as it combines four different ANNs – membrane, interface, loops and globular residues – through a HMM. HMMs consist of a set of sequential states, whose progress is dependent on the confirmation of the current state [123]. TMs were also predicted using SVMs: Memsat-SVM [124-126]. BOCTOPUS [127], developed by the same group as OCTOPUS, allows to predict β-barrels. BOCTOPUS combines local predictions through SVMs and a HMM to combine all local SVM predictions.

Evolutionary conservation of residues, and coevolution [128, 129] are also a growingly utilized feature in protein interface prediction [130-133]. Coevolution, concerning G-proteins, aims at assessing evolutionary conservation of protein sequences and functions. Computationally, this gives rise to coevolution-scores, regarding the proteins, its' residues and sometimes inter-residue interactions, that can be of use structure prediction and refinement. These are based on the conservation of amino acids at the interfaces, as it relies on scoring residues or residues pairs, depending many times on MSA methods [119-121, 134-143].

2. METHODOLOGY

2.1 Homology Modelling

Homology modelling comprehends a process in which experimental information of a protein is used to, as accurately as possible, determine the tridimensional (3D) structure of another protein from sequence, of which this information is unknown [144]. Homology modelling approaches can make use of experimental information usually retrieved from X-ray and NMR experiments. The main problem with implementing homology modelling techniques is the search and availability of templates. Most software approaches already implement a search method, usually based on sequence similarity, in order to find the most suitable candidates. However, it is also generally possible to provide user designated structures as templates. Other problems arise from the need to predict particular secondary substructures, either loops, helixes or alpha sheets with accuracy, since they might interact particularly, in which case their prediction as correct as possible is needed. For such, some approaches focus on predicting these structures in particular, while other approaches focus on protein structure refinement after initial assessment. MPs, in particular, due to their membrane associated parts, are difficult to predict, since there are fewer templates than in other cases and *de novo* prediction is also harder, since the solute is not only different, but also not homogeneous, which implicates major changes in the structure [145].

The prediction of complexes, in particular those involving one or more MP structures, adds complexity to the problem, since it is not enough to match the monomers together, even admitting their individual structure was correctly predicted, but their interaction patterns and structural modifications need to be taken into account. For such, structure refinement is a major resource [146].

2.2 Multiple Sequence Alignment

MSA is an indispensable tool on protein structure computational biology. MSA relies in the alignment of sequences and matching of residues from different proteins' sequences, while inserting gaps in order to maintain the most conserved areas. Global approaches can be used with an overall alignment of sequences but local optimization algorithms are more commonly utilized. These, among an alignment of sequences, start with subalignments of smaller portions of the sequences and expand until the full sequences are portrayed [147]. To pair up the residues from the different sequences in the most optimized order possible, as well as introduce gaps in the sequences where there is no homologue residue, algorithms are employed that attribute likelihood scores to the residues, and between the sequences. From these scores the alignments can be built and/or progressively adjusted, depending on the method. The method employed by Clustal Omega [148] follows this procedure with characteristics that make it both accurate but also able to consider larger sequences, a problem that is present in most exclusively progressive methods.

Clustal Omega [148], as some of the most recent approaches, builds HMM to the alignments that allow the attribution of likelihood scores to each residue of each aligned sequence. HMMs is a statistical approach, frequently used as a ML device, that emerged from Bayesian approaches. HMMs work by building probability profiles of target classes, in this case, the residue at each position, these residues are then tested in 'hidden layers' and the final product is evaluated, the key factor is that the process can be repeated after one alignment is achieved, which allows for the inclusion of consistency in HMMs, granting higher accuracy rates as well as the possibility of aligning large sequences [149]. The online approach often stores the HMMs profiles in order to later save computational time without losing accuracy.

2.3 Coevolution

Coevolutionary information is associated with phylogenetic relations and refers to the evolution between pairs of organisms or biomolecules. Therefore, it is dependent on homology similarly between the assessed, in this case, protein sequences. Coevolutionary information, generally, is a measure of residue conservation. Conserved residues, on the other hand, are usually related to important structures elements that can be essential for protein function performance [150]. Sequence related coevolution information is usually derived after performing MSA on a set of homologues and becomes relevant often in interdependent amino acid frequencies, at similar spots of the alignment but also regarding similar patterns of amino acid substitutions. Some sequence coevolutionary measurement approaches available today are:

- McLachlan-Based Substitution Correlation (McBASC) calculates similarity by linear correlation using the information of PSSMs;
- Mutual Information (MI) measures dependency and covariance between variables by calculating the ratio between their joint occurrence probability and their independent occurrence probability [151], given in equation 1. In his equation, *i* and *j* stand for residue amino acids at different positions; *x* and y are, respectively to the query sequence positional numbers *i* and *j*, their similarity representatives. In some cases, this metric can be subjected to Average Product Correction (APC) as described in equation 2. APC averages (indicated by avg())the columns (a,b) and divides it by the average total MI as previously calculated by equation 1 [120];

Equation 1: Mutual Information formula

$$MI(i,j) = \sum_x \sum_y Pij(x,y)\log(\frac{Pij(x,y)}{Pi(x)Pj(y)})$$

Equation 2: Average Product Correction formula

$$APC(a,b) = \frac{MI\big(a,avg(x)\big)MI(b,avg(x))}{avg(MI)}$$

- Chi-square, as MI, describes coupling probabilities. However, instead of using a logarithmical ratio, uses the mathematical square [120], as described in equation 3. In the equation 3, *i* and *j* stand for residue amino acids at different positions; *x* and *y* are, respectively to the query sequence positional numbers *i* and *j*, their similarity representatives.

Equation 3: Chi-square formula

$$x^2(i,j) = \sum_x \sum_y \frac{(Pij(x,y) - Pi(x)Pj(y))^2}{Pi(x)Pj(y)}$$

- Pearson correlation considers the effective sum of alignments where both positions are not gaps (*N_{eff})* and similarity scores (*Sil*, *Sij*, *Si*, *Sj*, originated from a PSSM matrix) for the possible positions (the different residues assessed, for the given position). It is also based on the weights derived of the division of different values depending on the frequency of states (states being indicated by W) per position and the standard deviation (σ) [120] of the amino acids at given positions *(i, j),* as described in equation 4, in which *i* and *j* stand for residue amino acids at different positions, the *l* indicator always refers to the amino acid at the aligned sequence.

Equation 4: Pearson correlation formula

$$r(i,j) = \frac{1}{Neff}\sum_l \frac{Wsl(Sil - avg(Sl))(Sjl - avg(Sj))}{\sigma i \sigma j}$$

- Joint Shannon Entropy (equation 5), similar to MI and Chi-square, is used to define conservation [120]. In equation 5, *i* and *j* stand for amino acids at different positions; *x* and *y* are, respectively to the query sequence positional numbers *i* and *j*, their similarity representatives.

Equation 5: Joint Shannon entropy formula

$$S(i,j) = -\sum_x \sum_y Pij(x,y)\log(Pij(x,y))$$

- Direct-Coupling Analysis (DCA) calculates the frequencies of residue couplings by assessing the amount of times a pair of residues is present in each alignment (the designation “pair of residues” is used to describe two residues at specific locations) and by calculating the frequency of the individual residues in each location. From this information, a covariance matrix is calculated, reporting on residue coupling conservation [152]. In equation 6 [153] the couplings, *DI(i,j),* are calculated considering probabilities of pairs of specific residues at specific positions, *Pij(Ai,Aj),* where *i* and *j a*re the positions and *Ai* and *Aj* the specific residues at those positions. Additionally, it takes into consideration *Ai* and *Aj'*s relative frequencies to the positions (*fi* and *fj).*

Equation 6: Direct-Coupling Analysis formula

$$DI\ (i,j) =\ \sum_{Ai,Aj} Pij(Ai,Aj)\ln(\frac{Pij(Ai,Aj)}{fi(Ai)fj(Aj)})$$

- mean-field Direct-Coupling Analysis (mfDCA) uses an approach similar to the one used in DCA by combining the maximum entropy principle in order to minimize the biasing of the model [152];
- Protein Sparse Inverse COVariance (PSICOV) starts by building a covariance matrix in which directly coupled sites are inferred according and from which covariance scores are calculated, according to equation 7 [151]. S(a,b,i,j), the covariance score, depends on the amino acids type (represented by a and b) and the residue positions (represented by I and j) of a pair of residues. To calculate S, a sample matrix is built with a size of n residues, and the scores are calculated taking into account binary variables represented by x, that indicate the absence of presence of the amino acids of type a or b on the position I or j, respectively.

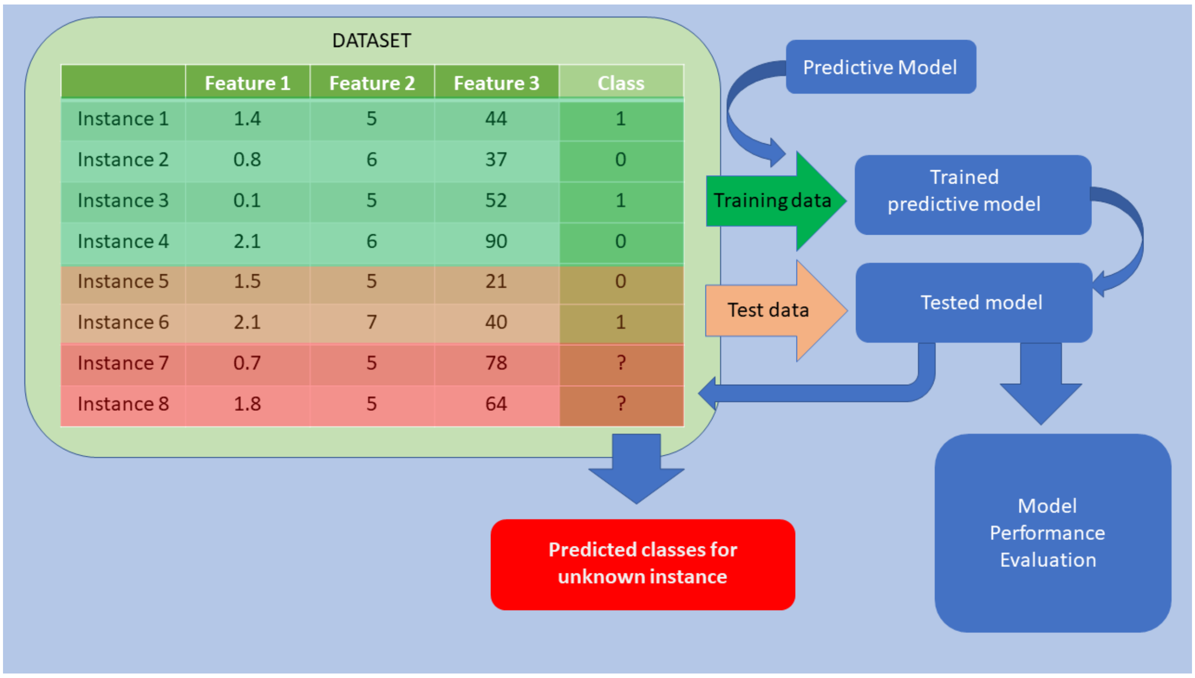
Equation 7: Protein Sparse Inverse COVariance

$$S(a,b,i,j) = \frac{1}{n}\sum_{k=1}^{n}(x(i,a,k) - avg(x(i,a))(x(i,b,k) - avg(x(i,b))$$

Other metrics are available as well as adaptations of these metrics and more regular statistical models, such as chi-square adapted to the data under scope.

2.4 Machine Learning

ML stands as an entire field of computer science that makes use of tools from mathematics, information theory, statistics, informatics among many other areas. Its employment is spreading worldwide and makes its stance as a useful approach in many other fields of study. Regarding computational biology, in particular protein structure, it has given proof of being able to give an important contribution to the field in many different prisms [154-157]. ML has the purpose of enabling the machine to learn from data in order to later predict unknown outcomes or perform tasks. Its typical workflow consists of a set of steps as displayed in Figure 2. According to a recent comprehensive review [158] and demonstrated by a series of recent publications [159-161] to establish a really useful computational tool for a biological system, we need to consider the following procedures: (i) construct or select a valid benchmark dataset to train and test the model; (ii) formulate the biological samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be analysed; (iii) introduce or develop a powerful algorithm (or engine) to operate the analysis; (iv) properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the statistical method and (v) establish a user-friendly web-server for the method that is accessible to the public.



DATASET

	Feature 1	Feature 2	Feature 3	Class
Instance 1	1.4	5	44	1
Instance 2	0.8	6	37	0
Instance 3	0.1	5	52	1
Instance 4	2.1	6	90	0
Instance 5	1.5	5	21	0
Instance 6	2.1	7	40	1
Instance 7	0.7	5	78	?
Instance 8	1.8	5	64	?

Figure 2: Machine-Learning basic workflow

ML is applied to a dataset comprised of instances, which are described by features. The number of instances is the number of samples available to feed to the ML model so that it can learn. The features that describe the instances are essential in this process, since they will help generate a model that can distinguish between new, unknown instances. ML models can be categorized into unsupervised, semi-supervised and supervised learning. Supervised learning is employed on a dataset of which final classes are known and, usually, if the data is available, is the approach chosen. Unsupervised learning, on the other hand, is used on the absence of knowledge on the outcome, it can be useful to build artificial partitions for the available data, by employing clustering algorithms. Semi-supervised learning is the term sometimes used to refer models that can either accept data with or without known outcomes. In this thesis we focus mainly on supervised learning. Given a dataset, the model was trained by experimenting various predictive models. However, not all the instances of the dataset were used for this step (training step) and some are set aside for later testing (testing set). This partitioning usually is randomized and a typical 70%-30% rule is followed. The performance of the different algorithms is assessed in the testing set. The final model can be used to predict the outcome of new instances.

2.4.1 Data pre-processing

A dataset should be comprised and described by features as independent and non-redundant as possible, guaranteeing they convey different and useful information. Several steps performed on the dataset, previously to the model training, aim to ensure this and are considered data pre-processing [162, 163]. So, data pre-processing steps aim at maximizing the value of the data in order to build the best predictive model possible while minimizing computational cost. These steps tend to prevent common issues such as underfitting and overfitting [164].

Underfitting happens when the dataset information is not enough for the model to capture the trend of the data and effectively categorize its' instances accordingly. Underfitting usually happens due to small amounts of instances or features with low variance. This can lead to biasing and predictive models unable to properly work with unknown samples. In summary, these models are not generalized and are too simple for their purpose [165]. The generation of new independent non-redundant features is a path to counter underfitting which can be achieved in several different ways, being highly dependent on the case. If possible, adding new instances is also a viable approach, but depends on the accessibility of some specifics types of data.

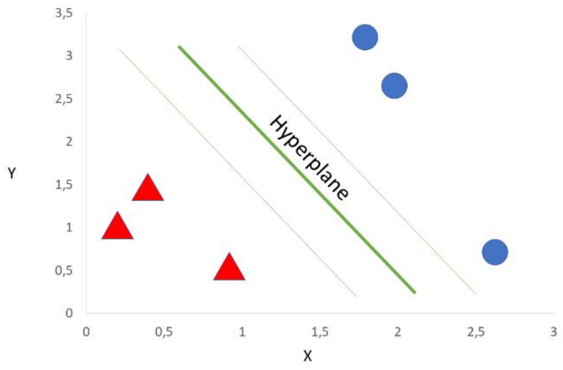
Overfitting, on the other hand, usually occurs in the opposite sense: the model has excessive amounts of information and fits too much on the data, leading to a similar generalization problem. In both cases, poor predictions are usually generated from the models, although overfitting is usually both more common and harder to deal with than underfitting. Whereas, elimination of instances is not a viable option, feature selection is usually the approach considered to counter overfitting. Feature selection is a process in which the features are evaluated accordingly with their contribution to the model. If they are not correlated with the class, they should be eliminated. Similarly, if they are redundant among each other, the most redundant are eliminated [166]. Feature selection methods can vary greatly, some of the most simple are MI [167] and PCA [168]. PCA works by performing an orthogonal transformation on the data, forcing it into a set of linearly uncorrelated values whose dimensionality is equal or inferior to the number of features, thus excluding the components with less variance. This is one approach to face the so-called 'curse of dimensionality', which is a very common issue caused by a high number of data variables (features/dimensions) [169].

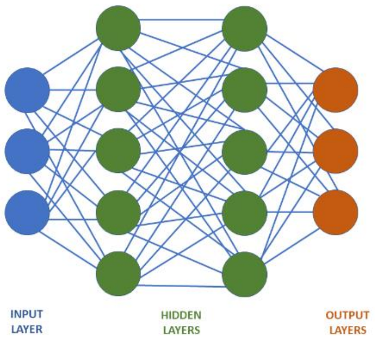
In order to maximize the use of the dataset and guarantee the best dataset possible, cross-validation is commonly performed. Cross-validation prevents biased models, by swapping and testing the instances in different combinations. Cross-validation performs resampling so that the data can then fit and be evaluated on the model. Some of the data is used to train the model, the rest is used to test it. By doing this, it is decreased the risk of biasing of the data and increased the probability of picking the best performing model, out of one method [169, 170]. Other pre-processing steps can be added, for instance, the scaling/normalization of the data: the subtraction of the average to the value, divided by the standard deviation. This helps prevent the different features from drifting in variability too much [171, 172].

2.4.2 Models

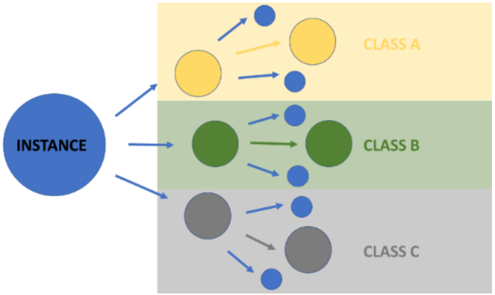
ML models can be, as stated before, of different types. They generally take different amounts of time to train and deliver different results with different performances. Table 1 points out a few basic ML models from which many others can be generated, as well as combined. Although many prediction models do well on their own, there are several approaches that allow the combination of different models. However, it is not straightforward that the combination of different approaches leads to better performance. Bagging, or bootstrap aggregating, is one approach to combine several models, and it works by generating different versions of a single predictor and employ these on a combined predictor [173]. The several models are evaluated separately and ranked depending on their performance. The rank then determines their contribution to the combined model. The replicates are trained on the same dataset, however subjected to bootstrap (random sampling) [174], in order not to repeat the runs. The bootstrapping allows the extraction of more possible combinations of the dataset (not only regarding instances, but also features) to train the several models, which then can be combined. Costing more computational time than the single models, it is essential that these approaches do not perform worse than the individual models, for the same reason, they usually work better in models that can have higher variance in their predictions [175]. Dimensionality reduction is another factor to take into account, in order to improve the cost/performance ratio of bagging models. Boosting is another example of ensemble model and focuses on building a strong prediction model from several weaker ones. A boosting ensemble does not necessarily pick a set of the best prediction models and combines them, rather, it selects the models that, although might not be the ones to perform better individually, have high variance, as a set. Boosting considers the individual models as estimators and builds a function that attempts to minimize the loss of prediction value among them, by attributing different weights to each model [176]. The size of the datasets used on boosting matters, as it tends to be less effective for high dimensionality cases, however, it usually can still outperform the individual models [177]. Combined or ensemble models have evolved greatly since bagging and boosting final forms, overall, more recent approaches continue to become more used and reliable [178].

Table 1: Machine Learning models' examples

<table>
<tr><td>ML model</td><td>Type</td><td>Model key information</td><td>Reference</td></tr>
<tr><td>Naïve Bayes (NB)</td><td>Classification</td><td><i>Equation 8: Bayes' Theorem</i>
$$P(A\backslash B) = \frac{P(B\backslash A) * P(A)}{P(B)}$$</td><td>[179]</td></tr>
<tr><td colspan="4">Description: Based on Equation 1, a NB model calculates the likelihood of the instance belonging to class (A) conditionally to each feature B. In the equation only one feature is represented, however, the formula can be easily adapted for several features.</td></tr>
<tr><td>k-Nearest Neighbours (kNN)</td><td>Classification and regression</td><td><i>Equation 9: Euclidean distance</i>
$$D(p,q) = \sqrt{(p1 - q1)^2 + \cdots + (pn - qn)^2}$$</td><td>[180]</td></tr>
<tr><td colspan="4">Description: A kNN model generates the classes based on the closest instances, the distance is measured with the formula presented in equation 2. New instances are labelled accordingly with their distances to the clusters.</td></tr>
<tr><td>Support Vector Machines (SVM)</td><td>Classifier</td><td>

<i>Figure 3: SVM visual example</i></td><td>[181]</td></tr>
<tr><td colspan="4">Description: A SVM model works by building a hyperplane that divides the instances (circles and triangles) according to proximity, the more the margins of the plane and the groupings, the more accurate the model.</td></tr>
<tr><td>Artificial Neural Network (ANN)</td><td>Classifier</td><td>

<i>Figure 4: ANN visual representation</i></td><td>[182]</td></tr>
</table>

<table>
<tr><td colspan="4">Description: Based on their mammal brain homologues, ANNs imitate neuronal networks. On the simplest form, there are three types of layers: input (containing dataset information), hidden (derived from the input through calculated associations) and output (the probability of the instance belonging to each the classes. The nodes of one layer, represented by circles in the figure are connected by edges to the node of the next layer, in this case, each node from one layer is connected to all the nodes of the next layer. The connection is made through activation functions with a given threshold that, as in neurons when passed, will trigger a response on the next neuron.</td></tr>
<tr><td>Random Forest (RF)</td><td>Classification and regression</td><td>
<em>Figure 5: RF visual example</em></td><td>[183]</td></tr>
<tr><td colspan="4">Description: As depicted in the figure, a RF is an ensemble of decision trees (represented by the coloured boxes) through which an instance is subjected, only progressing in the decision if given parameters are met. Depending on the path taken, the instance will belong to different classes.</td></tr>
<tr><td>Logistic regression (logit)</td><td>Classification and Regression</td><td><em>Equation 10: Logistic regression general formula</em>
$\ln\left(\frac{Pi}{1-Pi}\right) = \beta 0 + \cdots + \beta 1X1, i + \beta mXm, i$</td><td>[184]</td></tr>
<tr><td colspan="4">Description: Logistic regression is a regression model on a categorical variable (can only assume 0 or 1 as values). It considers several variables, weights them and returns a prediction based on the different contributions of each variable. This is shown on Equation 4, where β stipulates the contribution of the several variables X.</td></tr>
</table>

2.4.3 Performance Evaluation

Performance evaluation in ML is utterly important, in order to compare models and choose the best and, depending on the case, less time-consuming model. For this, it is usually considered a confusion matrix, as exemplified in table 2. This table enables the comparison between what is measured by the model and the actual values, regarding the testing instances put aside for this purpose. The different metrics on the cells of the matrix are True Negative (TN), False Positive (FP), False Negative (FN) and True Positive (TP). These metrics can be associated with rates to inform on the performance of the model, these rates are: accuracy, True Positive Rate (TPR or sensitivity), True Negative Rate (TNR or specificity), Positive Predictive Value (PPV or precision), Negative Predictive Value (NPV), False Discovery Rate (FDR), False Negative Rate (FNR) and F1-score (Equations 9 to 16).

Table 2: Confusion matrix

Predicted Actual	No	Yes
No	True Negative (TN)	False Positive (FP)
Yes	False Negative (FN)	True Positive (TP)

Equation 11: Accuracy formula

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

Equation 12: Sensitivity formula

$$TPR = \frac{TP}{TP + FN}$$

Equation 13: Specificity formula

$$TNR = \frac{TN}{FP + TN}$$

Equation 14: Precision formula

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

Equation 15: Negative Predictive Value Equation

$$NPV = \frac{FP}{FP + TN}$$

Equation 16: False Discovery Rate Equation

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

Equation 17: False Negative Rate Equation

$$FNR = \frac{FN}{FN + TP} = 1 - TPR$$

Equation 18: F1 - score

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

Apart from the metrics displayed, a very common metric is the Area Under Receiving Operating Characteristic (AUROC), which is the area under the curve described by the probability of the classifier ranking a true instance (consider T, in equation 17) higher than a randomly chosen negative one. The area is calculated with the integral below, and is represented as the area below a curve marked by the different assessed data points, as in the example figure 6, plotted using pROC of the R package [185].

Equation 19: AUROC formula

$$AUROC = \int_{-\infty}^{\infty} TPR(T)\big(-FPR'(T)\big)dT$$

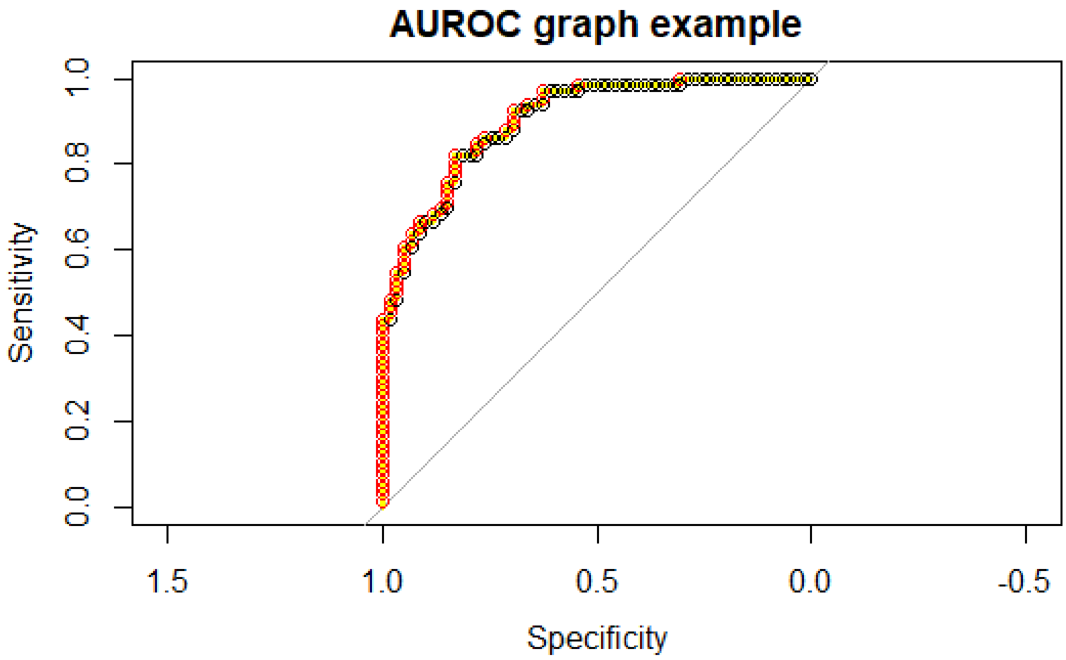


Figure 6: Example AUROC graph

2.4.4 Confounding variables

Confounding variables are variables correlated both with the response (the class variable) and with the input features. Confounding variables can suggest a correlation or causality [186], and lead to decreasing model performance [187]. Two variables are confounders when it is hard to impossible to separate their effect from each other. Since they carry information that has no causal connection to the class variable, they simultaneously introduce predictive error and make themselves nearly impossible to identify, prior to experiment. Following this, the best way to exclude confounding variables is definitely to not include them at all, since feature selection models based on variance (which comprise most of the available feature selection approaches) will not be able to rule them out. However, this is not possible in all datasets or experiments.

3. METHODS AND MATERIALS

In the next sections are described the methods and results for two different pipelines of this thesis work. The first, in which are included the subsections from 3.1 to 3.4, focus on application of a variety of computational techniques for the building of complexes involving MP in general and GPCR in particular as well as their respective analysis. The purpose of this pipeline was to identify GPCR (in particular dopamine receptors from 1 to 5) patterns when in interaction with arrestins and G-proteins. These patterns should allow the discrimination of the most relevant substructures and residues as well as a more comprehensive understanding of all the protein-protein interfaces involved. In this study, we used several software programs and packages, and web platforms, namely MODELLER [188], High Ambiguity Driven protein-protein DOCKing (HADDOCK) [189], bio3d [190], InterProSurf [191], BioCOmplexes COntact MAPS (COCOMAPS) [192], Conservation Surface mapping (ConSurf) [193, 194], and EVolutionary Fold (EVFold) [153] and Elastic Network Modelling [195] for a thorough and comprehensive analysis of protein-protein interfaces in various DR-Arr complexes (D_1R-Arr-2, D_1R-Arr-3, D_2R-Arr-2, D_2R-Arr-3, D_3R-Arr-2, D_3R-Arr-3, D_4R-Arr-2, D_4R-Arr-3, D_5R-Arr-2, D_5R-Arr-3) as well as DR-G-protein complexes (D_1R-G_{i1}, D_1R-G_{i2}, D_1R-G_o, D_1R-G_{sL}, D_1R-G_{sS}, D_2R-G_{i1}, D_2R-G_{i2}, D_2R-G_{i3}, D_2R-G_o, D_2R-G_z, D_3R-G_{i1}, D_3R-G_{i2}, D_3R-G_{i3}, D_3R-G_q, D_3R-G_{sL}, D_3R-G_{sS}, D_3R-G_z, D_4R-G_{oB}, D_4R-G_{t2}, D_4R-G_z, D_5R-G_{sL}, D_5R-G_{sS}, D_5R-G_z). For all complexes, 3D homology models were used to assess a variety of evolutionary-based (conservation and co-evolution), structure-based (intermolecular interactions, salt bridges, hydrogen bonds, solvent accessibility), and dynamic-based (fluctuations and cross-correlation) to understand the molecular determinants responsible for binding specificity of DR complexes to their cognate G-protein and Arr subtypes.

The second pipeline, from subsection 3.5 to 3.8, describes the methods and materials used for the prediction of HS by ML approaches with and without the use of coevolutionary information. We have begun by applying ML to soluble proteins as these are not only essential systems but can be regarded as a proof of concept for further exploration of these algorithms to MP.

3.1 Homology Modelling

All proteins were constructed by homology modelling using the MODELLER package [188, 196], which allows the construction of 3D models from the amino acid sequence of a protein by means of the alignment with one or more known protein structure (template) that are likely to resemble the structure of the target sequence. The methodology helps to surpass the limits imposed by the scarcity of experimental structures of GPCRs available. The template PDB-ID: 3SN6 [197] was chosen for active-form of DR with UniProt sequence IDs P21728, P14416, P35462, P21917, and P21918 for D_1R, D_2R, D_3R, D_4R, and D_5R, respectively. An ALA_n linker was added to connect TM5 and TM6, which were modeled with extended helical segment (beyond the membrane) up to the linker, making the intracellular extension of TM5 and TM6 similar to that observed in the crystal structure of the β_2AR-G_s complex (PDB-ID: 3SN6 [197]). Clustal Omega program [198] was used for Multiple Sequence Alignment (MSA) of the five FASTA sequences of DR complexes. The crystal structure of the β_2AR-G_s complex (PDB-ID: 3SN6) [41], and of human rhodopsin-visual Arrestin complex (PDB-ID: 4ZWJ) [199] were used as templates for the construction of the 3D models of G-proteins and Arrs, respectively, using the MODELLER package [188, 196]. The accession codes of query sequences of G_q, G_z, G_{t2}, G_{i1}, G_{i2}, G_{i3}, G_{sS}, G_o, G_{sL} and G_{oB} used for homology modeling were P50148, P19086, P50149, P63096, P04899, P08754, P63092, P04971, GI:20147687, and GI20147683, respectively. The accession codes of query sequences of Arr-2 and Arr-3 proteins used for homology modeling were P49407-1 and P49407-B, respectively. One hundred models were created for each query sequence and the G-protein and Arr models with the lowest Discreet Optimized Protein Energy (DOPE) score were selected out of the ten models with the highest score for the MODELLER objective function. As for the DR complexes, the model which featured the highest intramembrane domain-ICL3 distance was selected out of the ten models with the highest MODELLER objective function.

3.2 Complex Building and Refinement

Structure refinement was performed with HADDOCK [200], which is a web platform able to perform protein structure refinement in an explicit solvent representation. To construct 3D models of DR complexes-G-protein complexes, the models of DR complexes and G-proteins were aligned based on the crystal structure of the β_2AR-G_s complex (PDB-ID: 3SN6) [41], and the models DR complexes and Arrestins were aligned based on the crystal structure of human rhodopsin-visual Arrestin complex (PDB-ID: 4ZWJ) [199]. These complexes were submitted to the HADDOCK server and the best model attained for each protein-protein complex was used in subsequent analyses. The final structures are listed in table 3.

Table 3: Final complexes after modelling procedures

GPCR	Complexes
D1R	D1R-ARR2, D1R-ARR3, D1R-Gi1, D1R-Gi2, D1R-Go, D1R-Gs(lo), D1R-Gs(sh)
D2R	D2R-ARR2, D2R-ARR3, D2R-Gi1, D2R-Gi2, D2R-Gi3, D2R-Go, D2R-Gz
D3R	D3R-ARR2, D3R-ARR3, D3R-Gi1, D3R-Gi2, D3R-Gi3, D3R-Gq, D3R-Gs(lo), D3R-Gs(sh), D3R-Gz
D4R	D4R-ARR2, D4R-ARR3, D4R-Gob, D4R-Gt2, D4R-Gz
D5R	D5R-ARR2, D5R-ARR3, D5R-Gs(lo), D5R-Gs(sh), D5R-Gz

3.3 Sequence alignment

Protein sequence alignment is needed for many of the steps involving both achieving and analyzing results. In this case, a first sequence alignment has already been described, when homology modelling was conducted. This was performed using the embedded functions the MODELLER [107] software provides. For the analysis steps to be displayed forward, the alignment performed was with EBI's online available tool, Clustal Omega [148]. This is a MSA tool that allows for the alignment of protein amino acid sequences presuming a certain degree on evolutionary similarity between them. For this purpose, the alignments considered were made with the evaluated sequences as described in table 4.

Table 4: Alignments made with the sequences, to be further used on upcoming steps

Alignment	Sequences aligned
Arrestins	ARR2, ARR3
Dopamine Receptor	D1R, D2R, D3R, D4R, D5R
G-Protein	Gq, Gz, Gt2, Gi1, Gi2, Gi3, Gs(sh), Gs(lo), Go, Gob

3.4 Structure analysis and interface characterization

The structural analysis of the complexes obtained by the methods described in the previous steps was conducted by assessing characteristics such as interface characterization, HB, SB, RMSD, distance between residues and coevolutionary features. Amino acid content analysis is important for the understanding of the interfaces,. In table 5 are listed the several amino acid groups, according to residue physicochemical properties (dependent on the amino acid side groups), as well as their triple and single letter codes, in respective order, relatively to the amino acid name presented on the 'Amino Acids' column.

Table 5: Amino acids groups, by physicochemical properties

Group	Amino Acids	Amino Acids (triple letter code)	Amino Acids (single letter code)
Nonpolar aliphatic	Glycine, Alanine, Valine, Leucine, Methionine, Isoleucine	GLY, ALA, VAL, LEU, MET, ILE	G, A, V, L, M, I
Polar uncharged	Serine, Threonine, Cysteine, Proline, Asparagine, Glutamine	SER, THR, CYS, PRO, ASN, GLN	S, T, C, P, N, Q
Positively charged	Lysine, Arginine, Histidine	LYS, ARG, HIS	K, R, H
Negatively charged	Aspartate, Glutamate	ASP, GLU	D, E
Aromatic	Phenylalanine, Tyrosine, Tryptophan	PHE, TYR, TRP	F, Y, W

3.4.1 Interface characterization, Hydrogen Bonds and Salt Bridges

Interface characterization aims to assess whole complex characteristics as well as determine important interfacial residues. One of the tools used to make interface characterization was COCOMAPS [192]. This tool is free-available online that takes pdb files as input [192] and retrieves scores regarding each of the chains (the GPCR and the G-protein/arrestin) as well as the complex. The different characteristics evaluated were:

- Interacting residues: as those in a cut-off distance inferior to 8Å, as default by the website;
- Buried and Surface Area: the buried and surface area of both the complex and the individual residues. For the complex they are also discriminated as polar or non-polar;
- Interfacial residues: defined on the basis of the buried surface area upon complex formation;
- HB: hydrogen bonds stablished between interacting residues;
- Physicochemical nature of the interacting residues: hydrophobic-hydrophilic, hydrophobic-hydrophobic, hydrophilic-hydrophilic and hydrophilic-hydrophobic.

Interface characterization was also performed with the aid of the Intersurf [201] webserver. This server is similar to COCOMAPS [192] in the input process. The definition of interface comes upon measuring the distances between complex geometrical points. Intersurf [201] also outputs the numbers of surface and buried atoms and polar (determined regarding their SASA, apolar and total energy per area; this information is available for both the chains and the complex). VMD [202], a software used for protein representation, was employed to assess the SB between the GPCRs and G-proteins/Arrestins of the different complexes.

3.4.2 Root-Mean-Square Deviation and Inter-residual distances

RMSD calculations was performed on the substructures of the GPCRs using PyMOL [40] and the .pdb files representing the refined structures, against the templates of the corresponding GPCR: β-2 adrenergic receptor for the dopamine receptors complexed with G proteins [203] and rhodopsin receptor for the complexes with the arrestins [204] . The RMSD was performed, not on the overall structure, but on each of the substructures: ICLs, TMs and HX8 (ECLs are on the outer part of the membrane whereas both arrestins and G-proteins interact with GPCRs through the inner side of the membrane). The RMSD was performed upon superimposition of the substructures of the models and their correspondents on the templates according to the numberings in table 6.

Table 6: The starting and ending number residue for each dopamine receptor substructure

		TM1	ICL1	TM2	ECL1	TM3	ICL2	TM4	ECL2	TM5	ICL3	TM6	ECL3	TM7	HX8
D1R	From	9	35	42	71	78	109	121	148	173	206	234	262	279	302
	To	34	41	70	77	108	120	147	172	205	233	261	278	301	314
D2R	From	6	34	41	70	77	110	119	146	160	191	219	249	257	281
	To	33	40	69	76	109	118	145	159	190	218	248	256	280	293
D3R	From	6	32	40	68	76	110	121	148	162	194	222	253	261	285
	To	31	39	67	75	109	120	147	161	193	221	252	260	284	297
D4R	From	5	32	39	67	74	110	117	144	158	194	219	251	257	283
	To	31	38	66	73	109	116	143	157	193	218	250	256	282	295
D5R	From	7	38	42	70	78	110	120	149	188	222	248	280	296	318
	To	37	41	69	77	109	119	148	187	221	247	279	295	317	332
β-2 adrenergic Receptor	From	32	62	69	98	103	137	147	170	195	237	265	299	304	327
	To	61	68	97	102	136	146	169	194	236	264	298	303	326	341
Rhodopsin Receptor	From	34	66	71	101	106	141	149	169	199	237	241	279	284	307
	To	65	70	100	105	140	148	168	198	236	240	278	283	306	326

All inter-residual distance were measured by in-house scripts and listed using the Weinstein-Ballesteros numbering [205]. This numbering assigns a x.50 residue to each of the helixes and describes these as being the most conserved residues. The residues before are then counted from 50 to 1, in a descent manner, while the ones after proceed from 50, in a crescent progression. The x stands by the number of the helix, for example, the most conserved residue at TM5 is called 5.50. This numbering allows for an easier analysis regarding the conserved interacting residues. The scores retrieved with the python script, using the Biopython [206] module, were then used, with R language and, particularly, circlize package [207], to build graphs for each of the complexes, displaying the interacting residues at a cut-off value of 8 Å, accordingly to each relevant substructure of the GPCR in study.

3.4.3 Evolutionary information

The degree of positional conservation of specific amino acid residues has been linked to their importance in protein structure and function. Thus, the determination of the conserved amino acid positions among G-protein family members may uncover the relevance of each position to structure and function of the receptor. EVFold [153] and ConSurf-DataBase (ConSurf-DB) [193, 208] software programs were employed to probe evolutionarily conserved position-specific amino acids and to identify structurally and functionally important regions within the proteins.

ConSurf-DB is an alternative web-based server which determines evolutionary conservation profiles for amino acids at a given protein sequence. The ConSurf -DB methodology is typically performed in three steps. The first step consists of skimming through the PDB Repository to search for and compile a list of protein sequences based on the PDB entry and chain ID and selection of the non-redundant protein sequences. using Protein Sequence Culling Server (PISCES) [209]. Afterwards. a Multiple Sequence Alignment (MSA) is constructed for each protein using Multiple Alignment using Fast Fourier Transform (MAFFT) [210]. initially by executing a Context Specific-Basic Local Alignment Search Tool (CS-BLAST) search on SWISSPROT database [211]. The list of collected protein homologues is then filtered according to the coverage with minimum of 80% and sequence identity ranging between 60% and 95%. The remaining sequence homologues are re-filtered by using Cluster Database with High Identity Tolerance (CD-HIT) by using 90% threshold for the sequence identity clustering. The search for sequence homologues is only carried out after each iteration if a maximum number of hit sequences has not been achieved. If the number of hit sequences detected after CS-BLAST running on SWISSPROT database is inferior to 50 hits. the Context Specific Iterated-Basic Local Alignment Search Tool (CSI-BLAST) is rerun with three iterations on the Uniref90 database of proteins. which is larger than SWISSPROT database. If the number of hit sequences remains lower than 50%. the minimal percentage of sequence identity is gradually reduced for homologous sequences. Finally. the selected hit sequences will be aligned using MAFFT. Subsequently. the resulting MSA is used to reconstruct a phylogenetic tree and to calculate position-specific conservation scores with Rate4Site [212]. which maps the rate of evolution among homologous proteins. The results of position-specific conservation scores are represented as a discrete scale of nine coloured Rate4Site grades depending on the degree of conservation of position-specific amino acids. The conservation scores are exhibited on the protein sequence or structure and on the MSA for visualization.

EVfold makes use of EVCoupling information for predicting the 3D structure of the protein by taking into account differences and discriminates the residues with higher evolutionary conservation. In particular, by using the protein sequence. and comparing it to sequence of the other proteins in the same family whose structure is known. Evolutionary Couples. which are pairs of residues that consistently interact with each other. are determined. Subsequently. with this information. 3D structure of a given protein can be built depending on the rankings of these couplings (higher rankings are more determinant on the structuring of the protein Evolutionary Couplings (EC) are calculated using either mean-Field Direct Coupling Analysis (mfDCA) [213]. or Pseudo Likelihood Maximization (PLM) [214]. In the former. EC calculation is relatively rapid. whereas with PLM the process is much slower but more accurate. More than one protein sequence can be provided to enable the search of protein-protein interactions by EC information. Considering inter-protein residue interactions. sequence features also stand as a factor for the possibility of their interaction. The quality of the alignment must be assessed on this process. as well the number of EC's to be used. The output files will inform on the predicted structure (as well as compare it to root-mean-square deviation – based structures) and the ECs.

The amount of data generated in these analyses was too extensive to be clearly detailed and described in this thesis. We only restricted it to the main important conclusions and full information about the results can be found at (http://45.32.153.74/gpcr/). Static data visualization plots in this thesis were performed using *ggplot2* [215] and website construction was performed using *shiny* application [216] from R package.

3.5 Hot-Spot Dataset

Our final HS dataset was constituted of observations of a known class, the class being either HS or NS. As mentioned, a residue is considered a Hs depending on if, upon alanine mutation, it generates a binding free energy difference ($\Delta\Delta G_{binding}$) superior or equal to 2.0 kcal/mol; if $\Delta\Delta G_{binding}$ is inferior to 2.0 kcal/mol, the residue is considered to be a NS [217]. The $\Delta\Delta G_{binding}$ energy for 533 instances, each representing one residue, was collected from the existent four databases: ASEdb [8], BID [218], SKEMPI [219] and PINT [220]. For each of the residues of the dataset, we collected various structural and evolutionary-related features, in order to later subject them to ML algorithms. The features comprised a lot of different characteristics, some of them being relative to the residue in particular while others regarded the protein or substructure they were involved in. The features were split into non-coevolutionary features and coevolutionary features, and 2 different tasks were performed. First we have not used coevolution to train our model and in a second approach we have added these ones to access their role on the overall model performance. The main idea here was to access if the use of coevolution information, although more difficult to attain form an experimental point of view, could improve the performance of a HS detection method.

3.5.1 Non-coevolutionary features

Several different non-coevolutionary features were considered, ranging from sequence-based features to structure based ones, adding up do 868 non-coevolutionary features. It should be noted that these features can include evolutionary information, but not coevolutionary information.

Table 7: Non-coevolutionary features

Features	Description	Number	Reference
Solvent Accessible Surface Area (SASA)	Measurement of water accessibility of the residue take into consideration also monomer and complex SASA (hence also a hydrophobicity measurement)	10	[221]
Interface size	Total number of interface residues and SASA total variation	2	
Number of interface residues	Between the 20 amino acids, the counts of each at the protein's interface	20	
Number of protein-protein contacts	Contacts within 2.5 Å and 4.0 Å distance cut-offs	2	[222]
Number of intermolecular hydrophobic interactions	Number of HB and SD assessed with VMD	2	
Position-Specific Scoring Matrix (PSSM)	Residue frequency associated to full sequence or subsequence, by amino acids, using BLAST	20	[223]
Weighted observed percentages	Amino acid weighted (disregarding alignment gaps for percentage calculation) percentage at the interface, using BLAST	20	
Amino Acid Composition (AAC)	Frequency of each amino acid on the protein, achieved with R package:	792	[224]

	protr		
Pseudo Amino Acid Composition (PAAC)	AAC extended so that it includes ordered information on the amino acid content, achieved with R package: protr		
amphiphilic Pseudo Amino Acid Composition (aPAAC)	PAAC annotated on the hydrophobicity/hydrophilicity characteristics of the protein, achieved with R package: protr		
BLOcks SUbstitution Matrix (BLOSUM)	Scoring matrixes based on highly conserved regions from an observed alignment, in this case, of 62% similarity, achieved with R package: protr		
Protein Fingerprinting	Calculates amino acids specific descriptors from Amino Acid index database (AAindex) [225], these can be narrowed with PCA, achieved with R package: protr		
ProteoChemoMetric (PCM) modelling	2D and 3D modelling descriptors that can be narrowed with PCA, achieved with R package: protr		

3.5.2 Coevolutionary features

ConSurf [226] results for residues were considered, as considered in subsection 3.4.3, for each of the instances of this dataset. Several coevolutionary features were considered, provided that we could retrieve scores for individual residues. In particular, we added results from the following 4 web-servers: EVFold [227], EVComplex [228] CoeViz web-server [120], and InterEvScore [121].

EVFold [227] individual scores were considered, since the software retrieves conservation and coupling strength values for monomeric structures. We have also used EVComplex [228], a web-package of the same authors, which gives scores for the final complex structure. Regarding the CoeViz web-server [120], coevolution scores were calculated by four different metrics: chi-square, mutual information (with further average product correction), Pearson correlation and joint Shannon entropy. All these metrics were defined in previous chapter. The alignments were built by fetching sequences up to a defined percentage of sequence similarity (predefined as 90%), and aligned, with MSA, with the help of PSI-BLAST [229]. Thus, conservation was calculated for the protein upon comparison of residue conservation in similar sequences. However, more than simply calculate the conservation of residues, this web-server allowed the calculation of pairwise –residue conservation. The results retrieved are constituted of tables with an equal number of rows and columns, with the protein sequence's residues in both the vertical and the horizontal. These form a matrix with the number of columns and the number of rows equals to the length of the sequence, each cell being filled with the score regarding the conservation of the pair or, in the case of the diagonal, the residue itself.

InterEvScore [121] was built with the purpose of predicting PPI with the use of multi-body interactions and coevolutionary information. This software attributes scores coevolution based on 2 and 3-body potentials that are determined by residue interaction propensity, derived from interaction frequency. Regarding evolutionary information, MSA couples were derived from the information of InterEvol database [230], retrieving interacting residue couples, these were then used to derive new scores, including the previous 2 and 3-body potentials, while adding a conservation factor using the Rate4Site program [231], which, using empirical Bayesian estimation, assesses evolutionary rates along the MSAs.

These features were then added to the previous 868 features' dataset and ran in different combinations in order to better assess their individual contribution to the final model. In some cases, due to missing values of some coevolutionary features (as the amount of genomic sequence available is not equally for all systems), the observations for which the coevolution values were missing were excluded, as described in table 8.

Table 8: Datasets to be used along the works

Dataset name	Number of observations	Number of features	Consurf (1)	InterEV (13)	CoeViz (8)	EVFold (3)	EVComplex (2)
Original	533	868					
"Allrows"	533	890	X	X	X		
"Both"	55	895	X	X	X	X	X
"Complex"	157	892	X	X	X		X
"Fold"	264	893	X	X	X	X	
"Fold*"	264	872	X			X	

*This dataset was built after all the other runs, in order to assess the possibility of InterEV and CoeViz features being confounding variables

3.6 Dataset Pre-processing

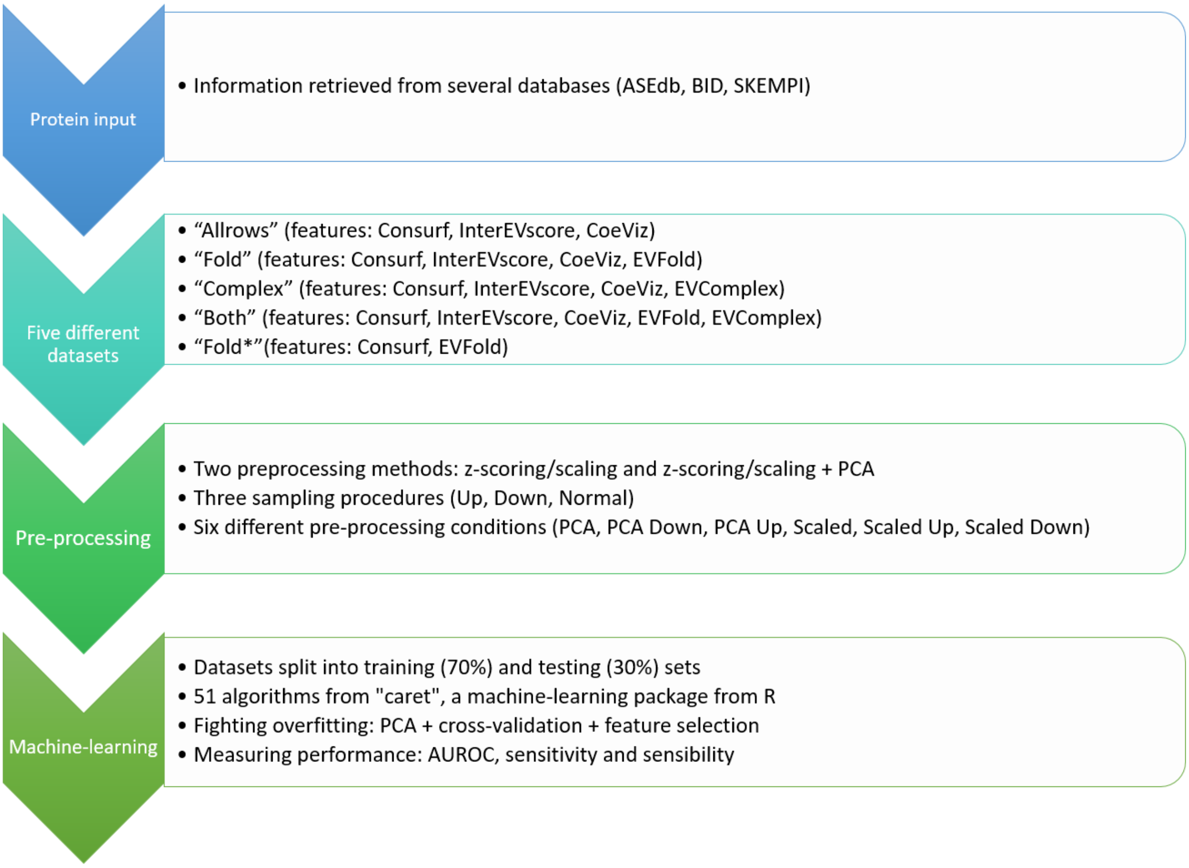
Dataset preprocessing was employed in order to prevent overfitting of the ML models over the dataset. Since the amount of features is larger than the amount of instances, this is likely to affect the training of the model. Also, the amount of HS is much lower than that of NS (an unbalanced dataset), which can raise important sensitivity issues. First, scaling of the data was performed, for each of the datasets, in order to normalize them (the subtraction of each value by its average value was split by the deviation). We used different approaches to assess the influence of the unbalanced taste in the final results. In particular, we randomly increased the HS occurrences, by replicating the existing ones, to match the amount of NS (up sampling) or by decreasing the amount of NS to match that of HS, by eliminating NS (down sampling). Finally, all the datasets were duplicated and half of them were subjected to PCA. Also, at least half of the tested algorithms had in-build feature selections methods and we cannot forget that all algorithms were tested in an independent test set retrieved randomly of the original database. All these preprocessing steps were performed using R scripts that made use of caret package [232] The final dataset for the non-coevolution example can be found at Moreira *el al. [13],* whereas the dataset that comprises coevolution data is listed in table 9.

Table 9: Final Datasets involving coevolution before ML training

Dataset Name	Dataset	PCA	Sampling	Scaling
Scaled Allrows	Allrows	no	Regular	Yes
Scaled Allrows up		no	Up	Yes
Scaled Allrows down		no	Down	Yes
PCA Allrows		Yes	Regular	Yes
PCA Allrows up		Yes	Up	Yes
PCA Allrows down		yes	Down	Yes
Scaled Both	Both	no	Regular	Yes
Scaled Both up		no	Up	Yes
Scaled Both down		no	Down	Yes
PCA Both		Yes	Regular	Yes
PCA Both up		Yes	Up	Yes
PCA Both down		yes	Down	Yes
Scaled Complex	Complex	no	Regular	Yes
Scaled Complex up		no	Up	Yes
Scaled Complex down		no	Down	Yes
PCA Complex		Yes	Regular	Yes
PCA Complex up		Yes	Up	Yes
PCA Complex down		yes	Down	Yes
Scaled Fold	Fold	no	Regular	Yes
Scaled Fold up		no	Up	Yes
Scaled Fold down		no	Down	Yes
PCA Fold		Yes	Regular	Yes
PCA Fold up		Yes	Up	Yes
PCA Fold down		yes	Down	Yes

3.7 Machine Learning Models training, testing and evaluation

The models were trained using R scripts that explored the functions and ML models of R package caret [232]. The models evaluated were (according to their names in the caret package documentation): 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. The evaluated models were previously subjected to a clustering process in order to separate them in five different clusters. To perform the hierarchical clustering, the information available on the model's characteristics was used to calculate the JACCARD coefficient and the complete aggregation scheme. The resulting models were trained on the randomized 70% of the instances from the two cases: with and without evolution related features, while the remaining 30% were used for the testing. The performance of the models was evaluated with the metrics AUROC, accuracy, TPR, TNR, PPV, NPV, FDR and F1-score. (already described in previous sections). In particular, the best performing models of each cluster were considered, for each of the tested datasets. The purpose of the clustering was to create later on an ensemble model. In particular, we used logistic regression, to combine the best performance model of each cluster into a final ensemble classifier. Overall, the pipeline describing the implementation for HS prediction with ML, making use of coevolutionary features, is described in figure 7.



- Information retrieved from several databases (ASEdb, BID, SKEMPI)

- "Allrows" (features: Consurf, InterEVscore, CoeViz)
- "Fold" (features: Consurf, InterEVscore, CoeViz, EVFold)
- "Complex" (features: Consurf, InterEVscore, CoeViz, EVComplex)
- "Both" (features: Consurf, InterEVscore, CoeViz, EVFold, EVComplex)
- "Fold*"(features: Consurf, EVFold)

- Two preprocessing methods: z-scoring/scaling and z-scoring/scaling + PCA
- Three sampling procedures (Up, Down, Normal)
- Six different pre-processing conditions (PCA, PCA Down, PCA Up, Scaled, Scaled Up, Scaled Down)

- Datasets split into training (70%) and testing (30%) sets
- 51 algorithms from "caret", a machine-learning package from R
- Fighting overfitting: PCA + cross-validation + feature selection
- Measuring performance: AUROC, sensitivity and sensibility

Figure 7: Overall pipeline for HS prediction with coevolutionary features

3.8 SPOTON web-site

The predictor for HS without coevolutionary information was implemented in a new and user-friendly web-server, "SpotOn" (Hot SPOTs ON protein complexes), which is freely available at:

http://milou.science.uu.nl/services/SPOTON/ . This implementation was done in collaboration with researchers from Utrecht University (Netherlands), Mount Sinai (USA) and Oporto University [13].

4. RESULTS AND DISCUSSION

The results attained were split into two different sections: i) the study of GPCR coupling and their structural characterization, and ii) HS identification without and with the involvement of coevolution features.

4.1 GPCR

GPCRs are practically ubiquitous proteins and drug targets, making them of high interest when dealing with a wide range of emerging diseases as Parkinson Disease (PD). The design of receptor subtype ligands that interact with the orthosteric binding site of GPCRs has proven to be ineffective, specifically for muscarinic acetylcholine receptors and metabotropic glutamate receptors, because of the high homology across binding sites of different GPCR subtypes, leading to a decreased subtype selectivity and specificity and unfavourable side effect profiles. Taking this into account, allosteric modulators are preferable to target subtype specific GPCRs by interacting with a protein region that is both larger and more diverse. Experimentally, these structure-based drug design methologies have the advantage of understanding drug-GPCR interactions at a molecular level, which is vital for the development of new and reliable pharmacophore models. Nevertheless, the drug design of GPCR modulators based on orthosteric or allosteric binding site requires prior structural data information, something that it is scarce for the majority of GPCRs. In fact, future drugs acting on GPCRs are likely to rely on ligand-based computational methodologies to tackle missing structural data information. Overall, these *in silico* approaches have been extremely relevant in early stages of drug discovery, particularly in lead optimization of drug candidates, in order to determine the most favourable molecular modifications for the identification of more potent and subtype selective GPCR modulators targeting PD. Another aspect of extremely importance in drug discovery process of GPCR modulators resides in their pharmacokinetic and toxicological profile since usually drug candidates with a favourable pharmacodynamic profile fail to advance at late stages of drug discovery process due to their unfavourable pharmacokinetic properties and toxicity. A drug design strategy that perfectly combines favourable pharmacodynamic properties of small molecule GPCR modulators with encouraging pharmacokinetic properties (e.g. blood-brain barrier permeability, brain exposure, *etc*) is crucial for the development of promising anti-parkinsonian agents with potential clinical efficacy. Due to highly relevance of understanding the computational approaches applied to the study of GPCRs role in neurologic diseases I have also been involved in the research and write of a bibliographic review on the subject, which will be part of a special issue on the topic at the Current Neuropharmacology journal. In a parallel work, I tried to better understand and characterize the differential coupling of dopamine receptors in particular. To better and more promptly display the most significant results, a fully dynamic webserver http://45.32.153.74/gpcr/ was constructed. Throughout the RESULTS section, references to the website contents will be made when appropriate.

4.1.1 Alignments

We started by performing the sequence alignments of all proteins involved in these pathways: DR, G-proteins and Arrestins, which are available in the website under the 'MONOMER' tab. GPCR's alignment, is particularly relevant due to elevated sequence similarly between crucial structural elements involved in activation and function.

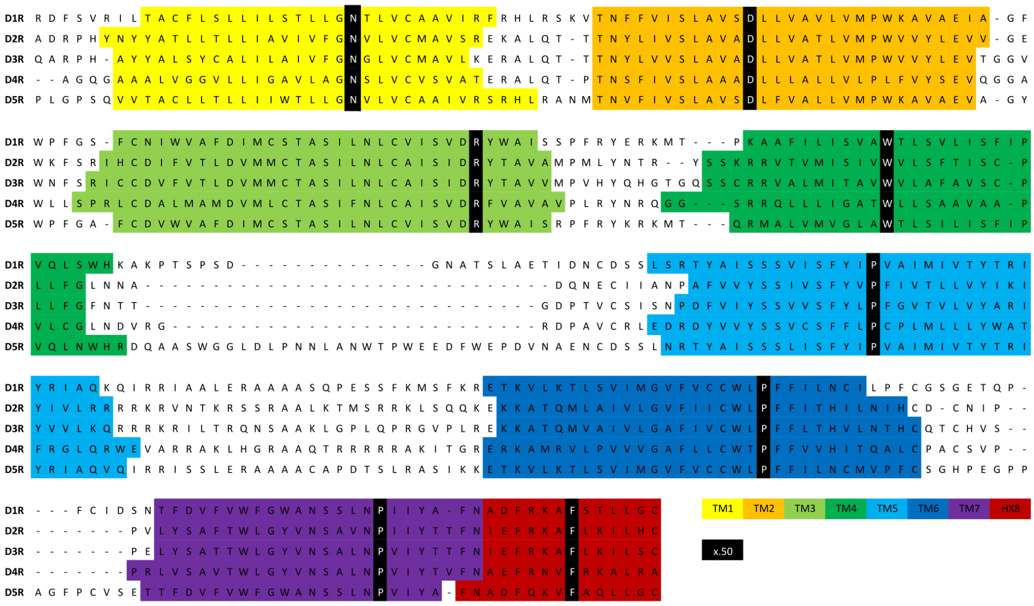


Figure 8: D1-5R sequence alignment

Figure 8 illustrates the alignments of D1R to D5R used for further analysis. The x.50 residues of all TM are colored in black. As mentioned, the ICL1 was considered to include all the residues ranging between TM1 and TM2, ECL1 between TM2 and TM3, ICL2 between TM3 and TM4, ECL2 between TM4 and TM5, ICL3 between TM5 and TM6 and, finally, ECL3 between TM6 and TM7.

4.1.2 Complex interface characterization

Interface characterization was performed using the methods and materials presented. The results' summary tables are presented on the annexed tables 1 to 6. The results on those tables were most of the used results for the graphical display on the website. Until the end of subsection 4.1.2 will be presented the some representative graphics, however, the full graphical display is available at the webserver. In this particular subsection, the information will the assessed regarding the whole complex, in order to later infer on the important sub structural differences and similarities, between the complexes.

4.1.2A Amino acid content

Amino acid content was assessed with originally built Python scripts, in order to achieve a first overall look at possible patterns distinguishable between the complexes, either between the DR – Arrestin and DR-G-protein complexes, but also regarding the particular monomeric dopamine receptors, G-proteins and arrestins involved. The interacting residues at dopamine receptors and at the complexing partners were clearly different both considering particular residue, but also overall physicochemical properties. There is a peak on glutamate content for D2R's interface with G-proteins, while histidine, methionine, proline and tryptophan residues are completely absent.

The characterization by individual amino acid content at the D2R-G-protein complexes previously reffered concerned the GPCR interface. The same complexes are considered, however, now at the interface of the G-proteins. Can be noted that the highest content is of arginine residues, with considerable contents of proline and methionine. As happened at the GPCR interface, tryptophan and histidine residues are completely absent, as are aspartate, cysteine, glutamine, glutamate and glycine. The differences at the interface suggest that there are residues more relevant on the interfaces. Additionally, the differences at the peaking and lowest contents in both interfaces suggest complementarity, however, to understand this, it is clearer to look at the residue group percentages graphs present in the 'DR COMPLEX STRUCTURE' tab. The amino acid content at GPCR interface of D2R-G-proteins complexes has only two groups that have above average residue content percentage, these are acid negatively charged and nonpolar aliphatic residue groups. Regarding the same complexes, but at the G-proteins' interface, the groups above average are nonpolar aliphatic, basic positively charged and polar uncharged, while acid negatively charged residues are completely absent. Overall, nonpolar aliphatic, polar uncharged and nonpolar aromatic percentages are similar between the interfaces at D2R and G-proteins. The percentages regarding acid negatively charged and basic positively charged residue contents, however, are drastically different. This tendency remains fairly unchanged for D1R, D3R and D5R. Considering D4R. However, is shown that the acid negatively charged residues at G-proteins' interface add up to about 10% of the residues, matching this group with nonpolar aromatic, which does not happen for any other dopamine receptor. In the same complexes, at D4R's interface, the acid negatively charged group is now below average. The information assayed can also suggest that there are relevant differences regarding the G-proteins' interfaces, when interacting with dopamine receptors.

Considering the complexes involving D1R, D2R and D3R – G-proteins, the amino acid content by group at GPCR interface can also be analysed at the website. In comparison with what was previously said when referring the residue group percentages of all the D2R – G-protein complexes at GPCR interface (for all remaining dopamine receptors this happened similarly, apart from D4R), can be seen that, when in contact with Gi1, dopamine receptors' acid negatively residue content at the interface is, not only above average, but the higher value, topping nonpolar aliphatic residues. This happens when the complexes considered involve Gi1, Gi2 and Gi3. When considering the remaining G-proteins, as Gs(lo), is noticed that the acid negatively charged residues' percentage at GPCR's interface drops well below the average.

Analysing the interface of interaction of DR-Arr complexes, it was found that the interface is mostly defined by nonpolar aliphatic residues. The ARG and LEU are the most predominant residues of DR and Arrs, respectively, observed in the interface of DR-Arrs complexes. In DR complexes, the ARG residues involved in the interaction are present in TM3 of DR complexes when complexed to ARR2 and in ICL2 of all DR-Arrs, except for D3R-Arrs. Similarly to the case of DR-Arr complexes, the interface of DR-G-protein complexes is primarily defined by nonpolar aliphatic residues. For D1R-, D2R-, and D3R-G-protein complexes, the most predominant residue in the interface of protein-protein complexes is GLU of DR complexes, for D4R-G-protein complexes the most predominant residue is LEU of D4R, and for D5R-G-protein complexes the most predominant residue is ARG of D5R. Concerning D1R-G-protein complexes, the GLU residues involved in the interaction are present mainly in ICL2 (except for D1R-Gi2 complex) and in ICL3 (with the exception of D1R-Gs complexes). Also, the GLU residue can be observed in TM6 of D1R-Go complex.

4.1.2B Energy by area and atom positioning

Regarding the energy by area, as can be seen in the online server on the 'DR COMPLEX STRUCTURE' tab, selecting the 'Area/Energy' interfacial feature, there are no large variations between all the complexes considered. Should, nevertheless, be noted that Gs proteins (Gs(lo) and Gs(sh)) have higher values for both polar and apolar energy by area and, consequently, the total is also higher than in other complexes. The arrestins' complexes, on the other hand, always stand below the average energy by area. Regarding the positioning of the atoms (surface or buried), the pattern is very similar to that of energy by area, with no large variances among all complexes. Gs proteins have more surfaced and buried atoms, while Arrs have less.

4.1.2C Evolutionary conservation and structural features

Evolutionary conservation values (Consurf and EVFold) were identified to be similar for the interfaces in all complexes, however, it is to note that its values were above average, regarding the whole complexes. Consurf, with a per definition average value of 5 has, on the interface's residues, a value of usually above 6 (see Figure 9). Similarly, EVFold values for interface residues are slightly above average for interface residues. As can be also seen in Figure 9, the amount of HB/SB is around 16. This value varies for all the complexes but, most importantly, varies among substructures, as will be seen further ahead.

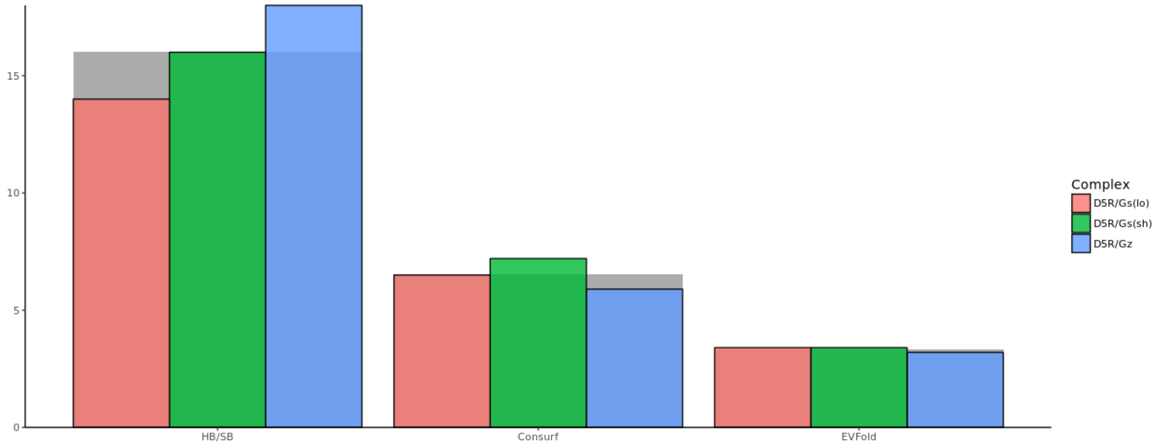


Figure 9: HB/SB, EVFold and Consurf values at the interface of GPCRs on D5R-G-protein complexes

4.1.3 Complex substructures interface characterization

In the previous subsection the whole complex interface relevant information was assessed. In the present subsection, information regarding the monomers is analysed. In particular, GPCRs' substructures ICL1, ICL2, ICL3 and HX8 were considered, as previous studies indicated them as the more prone to interact sub structures, in the complexes formed. As before, highlights will be given, since the website can be consulted for full analysis information.

4.1.3A Surface Area

The evaluation of surface area with both Intersurf and Cocomaps stands as a preliminary important characteristic on complex assessment, since it indicates the amount of residues and, overall, the importance of the substructure for complex formation.

In order to characterize the surface area of D3R-G-protein complexes, table 10 shows the range of values Intersurf and Cocomaps' surface area measure for each substructure, the same ranges being similar to all complexes involving G proteins. Regarding ICL1, can be seen that surface area is usually the smallest of the substructures assessed (Table 10). Gi1, Gi2 and Gi3 have close to none surface area at ICL1, while Go and Gob have absolutely none. Gs(lo), Gs(sh) and Gz, on the other hand, display higher values.

Table 10: Surface area range by substructure, as evaluated for the dopamine receptor – G-protein complexes

	Surface area metrics	ICL1	ICL2	ICL3	HX8
DR – G proteins complexes	Intersurf	0-20	350-500	300-550	0-90
	Cocomaps	0-50	250-500	250-500	20-140

ICL2 has less surface area on the complex involving Gq (D3R-Gq) than in any other of the complexes with D3R considered. Regarding ICL3, that for G-proteins has similar surface area values as those of ICL2, can be seen that in dopamine receptors in complex with Gs(lo) and Gs(sh) ICL2 tends to have higher values. Regarding Gi1, Gi2 and Gi3, D3R is the receptor that, when in complex, displays larger surface area.

Considering the HX8 surfaces areas on DR-G protein complexes, they are the second lowest, however, they peak substantially for D5R-G-protein complexes. Complexes involving Go, Gob and Gq proteins have close to no surface area at dopamine receptors' HX8, when interacting with the G-proteins. Gi1, Gi2 and Gi3 are the G proteins for which D2R's HX8 surface area values are higher. D5R, on the other hand has higher surface area values for Gs(lo), Gs(sh) and Gz related complexes. When considering arrestins' complexes, the surface area at ICL1 is always higher than for G proteins. At ICL2, ICL3 and HX8 arrestins' complexes almost always have higher surface areas, however, exceptions arise, particularly da complexes with Gs proteins. The heatmap present in Figure 10 depicts the surface areas evaluated by Intersurf, by substructure, for all the complexes, and aims at giving an overall picture of all the cases, while also clustering the complexes based on their surface area similarity. The data used to build is present in the annexed tables, and the individual graphical display is available at the webserver.

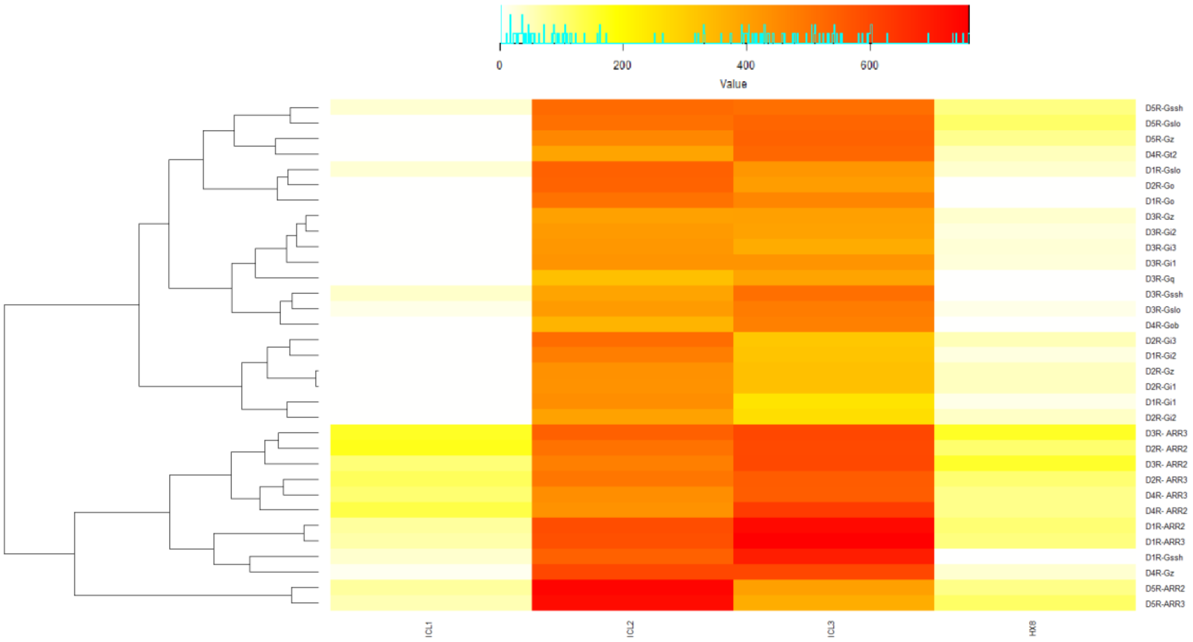


Figure 10: Heatmap reporting the interface surface area for ICL1, ICL2, ICL3 and HX8, at all the complexes analysed

4.1.3B Hydrogen Bonds and Salt Bridges

Hydrogen and Salt Bridges are main intervenients on the quaternary structure of proteins, for such reason, their identification is of utmost importance for the identification of the most relevant substructures on monomer interaction, regarding complexes.

HB/SB are present in all complexes, although they exhibit different patterns among the complexes' substructures. Regarding the tables annexed, can be seen that it is rare the case in which HB/SB are not present at either ICL1, ICL2, ICL3 and HX8 in a percentage above 90%. For such, once again, these substructures were considered to pinpoint the HB/SB location. At ICL1, the substructure in which HB/SB appear less commonly, in many complexes these are completely absent, however, at the complexes involving arrestins they appear more commonly than in the complexes involving G-proteins, still in very low amounts. Regarding ICL2, at D1R complexes there is a larger amount of HB/SB at Gi1, Gi2 and Go, however being lower at Gs(lo) and Gs(sh). Considering D2R complexes, the trend continues, appearing more HB/SB at the ICL2 of Gi1, Gi2 and Gi3. At D4R's ICL2 the highest amount of HB/SB occurs for the complex with Gt2, and at D5R in both complexes involving arrestins.

ICL3 has higher amounts of HB/SB in the complexes involving arrestins and D1R, however, the opposite occurs at D2R, where complexes with G proteins have more HB/SB. Regarding the G-proteins in complex with D1R, the amount of HB/SB at the ICL3 interfaces with Gi1, Gi2 and Go is lower than at

Gs(lo) and Gs(sh). In the complexes involving D2R, Go and Gz have the higher counts at ICL3. At D4R Gz, ICL3 has the higher amount of HB/SB, and arrestins have the lowest. T D5R’s ICL3, arrestins have the lowest amounts of HB/SB and Gs(lo) and Gs(sh) have the highest amounts of HB/SB. Due to the overall low amounts of HB/SB at HX8, and, contrary to ICL1, not particularly favouring any kind of complex, at this substructure there does not seem to be any major tendency to point out. A summary of all that was said is presented in the heatmap on Figure 11. All the data used to build it is present at the tables in the annexes. D4R, overall, displays more SB/HB with G-proteins. ICL3 is the substructure with the most HB/SB, followed by ICL2.

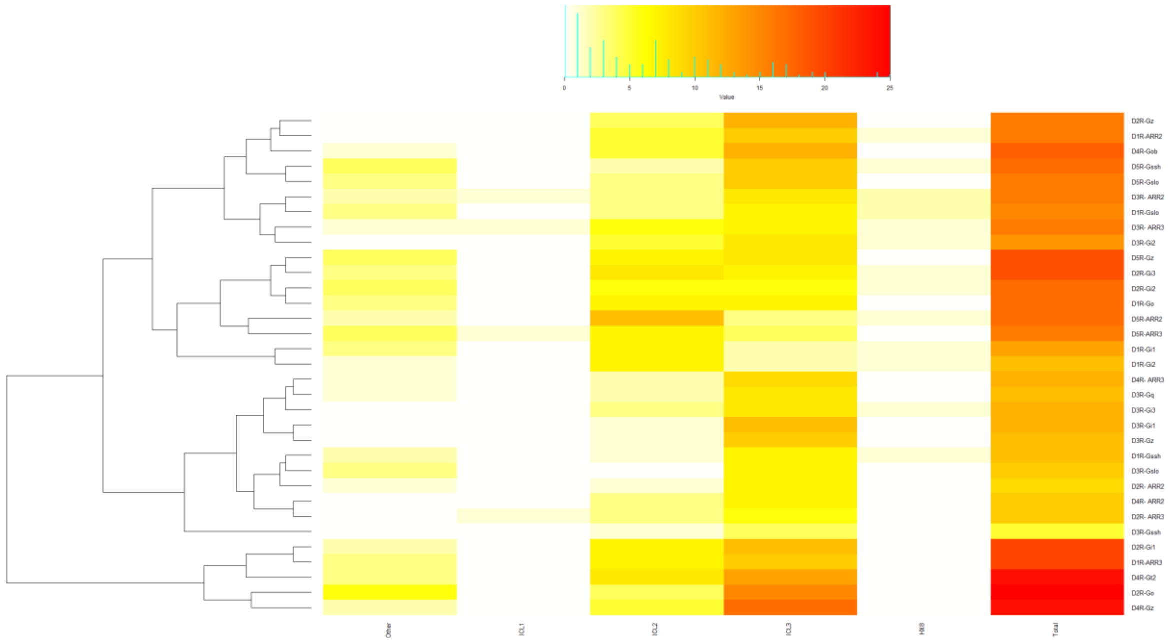


Figure 11: HB/SB heatmap displaying occurrence for ICL1, ICL2, ICL3, HX8, other locations and total, for all complexes

4.1.3C Interacting residues

Interacting residues can be assayed depending on their distance to other residues. The measurement of these pairs of residues was performed with the use of module biopython, via a python script, in which was included Weinstein numbering, at TMs. To visualize the interacting pairs of residues, the module circlize, by R software deploy, was used. All the graphs built are available on the webserver, on the tab ‘INTERPLOTS’. In the graphs produced, the residues were considered as interacting if they were standing at a distance below 8Å, as was considered at COCOMAPS to define interfacial residues. In the graphs, the bottom half of the circle is always considered to be the GPCR, whereas the upper half represents the residues of the G-protein or arrestin. The lines connecting the several sections indicate the residues pairing. GPCRs were showed as subdivided in their substructures, TMs, ICLs and HX8; the ECLs and the residues before TM1 were considered as ‘other’, since they are not expected to interact, as in fact they never did. Here will be presented some of these interaction plots, the remaining can be found at the webserver.

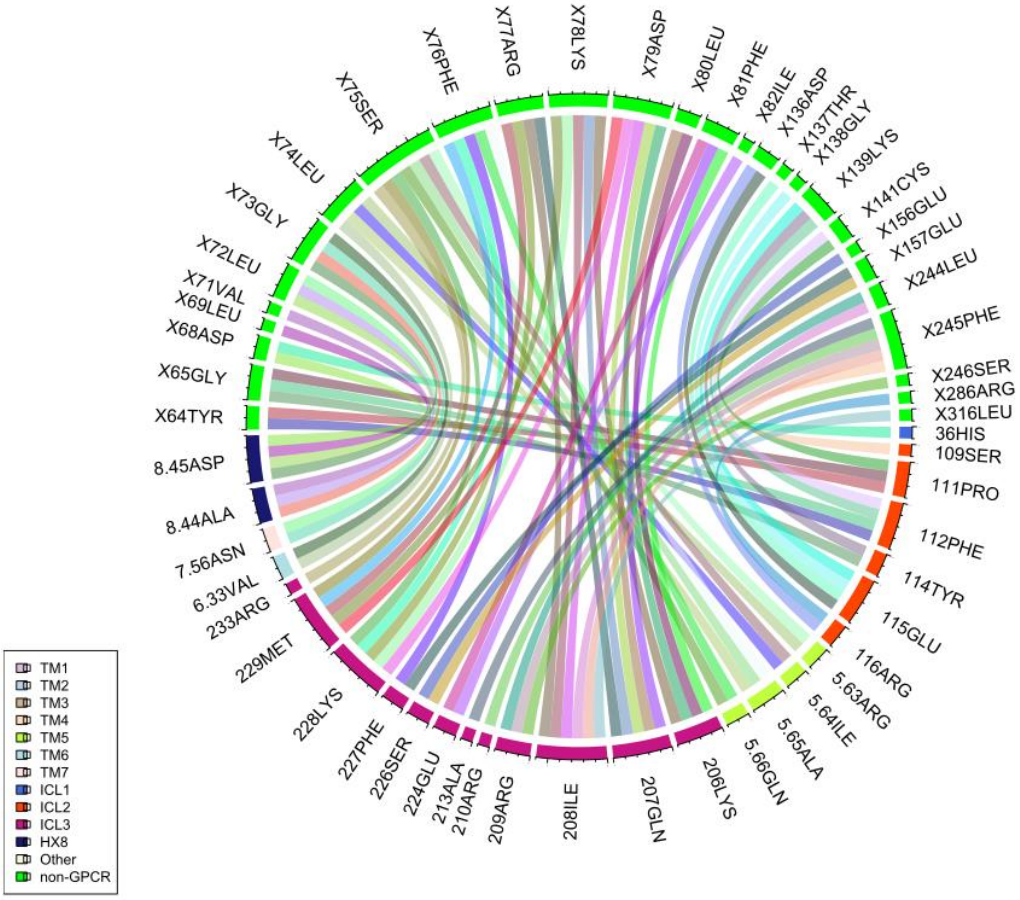


Figure 12: Interaction plot between ARR3 and D1R

Figure 12 shows the interacting residues between ARR3 and D1R. Although some differences between ARR2 and ARR3 were shown, when interacting with GPCRs, it is overall important to notice particularly the differences towards GPCRs complexes with G-proteins. Interacting residues at ICL1, when interacting with arrestins, although scarce, are a lot more frequent than when considering complexes with G-proteins. Furthermore, ICL2 and ICL3 seem to, in both cases, be highly populated with interacting residues. Another substructure had consistently a motif of residues, which was TM7/HX8, in which the same set of 3 or 4 residues, or physicochemical similar residues, was always present.

Regarding G-proteins' interactions with GPCRs can be pointed out large amounts of interactions at both ICL2 and ICL3, as well as the motif at TM7/HX8. However, the most striking founding was to notice similar interacting residue patterns with similar G-proteins, that did not appear in G-proteins not so similar. This allows to pinpoint not only regions but also residue patterns or singular residues as particularly important and selective for different G-proteins. The TM residues implicated on interactions, for all the complexes, are all close to the ICLs, which reinforces the importance of these substructures for the interaction with intracellular partners.

The structural features concerning the interaction between ARR2 and ARR3 and each of the five subtypes of DR complexes exhibit a plethora of common characteristics, more than simply identifying interactions between with the substructures, patterns can be unveiled. As seen before, both ARR2 and ARR3 interact with ICL2 and HX8 of all DR complexes. In ICL2 domain of DR complexes, Arrs bind to a four amino acid residue pattern composed of PRO, a nonpolar aromatic residue for D1-like receptors (PHE) or a nonpolar aliphatic residue for D2-like receptors (MET for D2R, VAL for D3R, and LEU for D4R), and two hydrophilic amino acid residues (GLU and ARG for D1R; ASN and THR for D2R; GLN and HIS for D3R; ASN and ARG for D4R; LYS and ARG for D5R). Regarding the HX8 domain of DR complexes, Arrs interact with TM7/HX8 through PHE, except for the ARR3-D5R complex, ASN, a nonpolar aliphatic residue (ALA for Arrs-D1R, Arrs-D4R, and Arrs-D5R; ILE for Arrs-D2R and Arrs-D3R), and an acid negatively charged residue (ASP for Arrs-D1-like receptors; GLU for Arrs-D2-like receptors). Moreover, Arrs interact with the ICL1 and ICL3 domains of D1-like receptors, with the TM7 domain of D1R, D2R, D3R, and D4R, with the TM3 domain of D2R, and with the TM5 domain of D4R. Analysing the residues involved in the interaction of Arrs with the respective domains, it was found that Arrs bind to SER and ALA residues of the ICL1 domain of D1R and D5R, respectively, and interact with PHE and ASN residues of the TM7 domain of D1R, D2R, D3R and D4R, for Arrs-D1R, Arrs-D2R, Arrs-D3R, and Arrs-D4R complexes. Concerning the ICL3 domain of D1-like receptors, Arrs bind to GLN residue of D1R and to LEU and GLU residues of D5R. In addition, Arrs interact with ALA of TM3 domain of D2R and with GLN residue of TM5 of D4R.

A careful analysis of the structural features of Arrs with each DR subtype was performed. Regarding the Arrs-D1R complexes, it was observed that Arrs interact with ICL1, ICL2, ICL3, and TM7/HX8 of DR complexes. More specifically, Arrs bind to SER residue of ICL1 and to GLN residue of ICL3 of D1R. Common interaction patterns were detected by analysing the interaction of Arrs with ICL2 and with TM7/HX8. In fact, Arrs interact with PRO, PHE. GLU, and ARG residues of ICL2 and with PHE, ASN, ALA, and ASP residues of TM7/HX8 of D1R. Considering the Arrs-D2R complexes, Arrs interact with TM3, ICL2, and TM7/HX8. More specifically, Arrs bind to ALA residue of TM3 domain, with MET, PRO, MET, ASN, and THR residues of ICL2, and with PHE, ASN, ILE, and GLU residues of TM7/HX8 of D2R. Regarding the Arrs-D3R interaction, Arrs interact with ICL2 and TM7/HX8 domains of D3R. More specifically, Arrs bind to MET, PRO, VAL, GLN, and HIS residues of ICL2, and PHE, ASN, ILE, and GLU residues of TM7/HX8 of D3R. The ICL2, TM5 and TM7/HX8 motifs of D4R are involved in the interaction with Arrs. By analysing the amino acid residues of the interacting structural motifs on the interface, it was observed that Arrs interact with PRO, LEU, ASN, and ARG residues of ICL2 domain, with a TRP residue of TM5 domain, and with PHE, ASN, ALA, and GLU residues of TM7/HX8 domain of D4R. The ICL1, ICL2, TM5, ICL3, and TM7/HX8 domains of D5R are involved in the interaction with Arrs. More specifically, Arrs bind to ALA of ICL1 domain, PRO, PHE, LYS, and ARG residues of ICL2 domain, and ASN, ALA, and ASP of TM7/HX8 domain of D5R.

Analysing the interaction of each Arr subtype with the five DR subtypes, allowed us to observe that ARR2 interacts with ICL1, TM3, ICL2, and TM7/HX8 of DR complexes, and with TM5 of D1-like receptors and D4R, and ARR3 interacts with ICL2 and TM7/HX8 of DR complexes, and with ICL1 and ICL3 of D1-like receptors.

Concerning the ARR2-DR complexes, it was observed that ARR2 interacts with THR residue of ICL1 of D2-like receptors. Additionally, ARR2 binds to ARG and a nonpolar aliphatic residue (ILE for D1-like receptors; ALA or VAL for D2-like receptors) of TM3 domain. Moreover, ARR2 interacts with PHE, ASN, a nonpolar aliphatic residue (ALA for D1-like receptors; ILE for D2-like receptors), and an acid negatively charged residue (ASP for D1-like receptors; GLU for D2-like receptors) of TM7/HX8 domain of DR complexes. Regarding the ICL2 domain, ARR2 interacts with PRO, a nonpolar aromatic residue for D1-like receptors (PHE) or a nonpolar aliphatic residue for D2-like receptors (MET for D2R, VAL for D3R, and LEU for D4R), and two hydrophilic residues (GLU and ARG for D1R; ASN and THR for D2R; GLN and HIS for D3R; ASN and ARG for D4R; LYS and ARG for D5R) of DR complexes.

Analysing the ARR3-DR complexes, it was found that ARR3 interacts with PRO, a nonpolar aromatic residue for D1-like receptors (PHE) or a nonpolar aliphatic residue (MET for D2R, VAL for D3R, and LEU for D4R), and two hydrophilic residues (GLU and ARG for D1R; ASN and THR for D2R; GLN and HIS for D3R; ASN and ARG for D4R; LYS and ARG for D5R) of ICL2 domain of DR complexes. Additionally, ARR3 binds to PHE, ASN, a nonpolar aliphatic residue (ALA for D1-like receptors and D4R; ILE for D2R and D3R), and a negatively charged residue (ASP for D1-like receptors and GLU for D2-like receptors) of TM7/HX8 of DR complexes. Moreover, ARR3 interacts with SER and ALA of ICL1 domain of D1R and D5R, respectively. In ICL3 of D1-like receptors, Arrs binds to GLN of D1R and to LEU and GLU of D5R. Similarly, to Arrs-DXR complexes, the interaction of the distinct G-protein isoforms with DR subtypes exhibits analogous structural features.

Regarding the Gi1-DR complexes, it was found that Gi1 interacts with TM3, ICL2, ICL3, TM6, and TM7/HX8 domains of D1R, D2R, and D3R. More specifically, the interaction of Gi1 with TM3 domain of DR complexes involves ALA and a nonpolar aliphatic residue (ILE for D1R and VAL for D2R and D3R). Additionaly, Gi1 interacts with LYS, a nonpolar aliphatic residue (VAL and LEU for D1R, and ALA for D2R and D3R), and THR of TM6 of DR complexes. The interaction of Gi1 with ICL2 involves PRO, a nonpolar aromatic residue for D1R (PHE) or a nonpolar aliphatic residue (MET for D2R; VAL for D3R), and two hydrophilic residues (GLU and ARG for D1R; ASN and THR for D2R; GLN and HIS for D3R). Considering the TM7/HX8 domains, Gi1 binds to PHE, ASN, and a nonpolar aliphatic residue (ALA for D1R; ILE for D2R), except for the Gi1-D3R complex. In addition, Gi1 interacts with LEU of TM5 of D2R and D3R. In ICL3 of DR complexes, a large amount of amino acid residues in the interface of Gi1-DR complexes are involved. Although ARG seems to be the common amino acid residue across all Gi1-DR complex interfaces in ICL3 domain, there is no clear residue pattern of interaction in this region.

Concerning the interface of Gi2-DR complexes, the Gi2 can interact with TM3, ICL2, TM4, TM5, ICL3, TM6, and TM7 of D1R, D2R, and D3R. More specifically, Gi2 interacts with nonpolar aliphatic residues (ALA and VAL for D2R; ALA, VAL, and ALA for D3R) of TM3 domain, with ALA or LEU residues of TM5 of D1R or D2R/D3R, respectively. Regarding TM7/HX8, it was found that Gi2 interacts with ASN residue present in the three DR complexes. In addition, Gi2 binds to PHE of D2R and D3R and to a nonpolar aliphatic residue of D1R (ALA) and D2R (ILE). Concerning the interaction of Gi2 with TM6 domain of DR complexes, it was detected a similar amino acid pattern of interaction when Gi2 binds to D2R and D3R, in which ALA and MET are the common residues. In addition, Gi2 interacts with LEU, VAL, and THR of TM6 of D1R. In ICL3 of DR complexes, a large amount of residues in the interface of Gi2-DR complexes are involved. Although ARG residue seems to be the common amino acid residue present in the interface of Gi2-DR complexes in ICL3 domain, there is no clear residue pattern of interaction in this region.

By analysing the Gi3-DR complexes, it was found that Gi3 binds to TM3, ICL2, TM5, ICL3, TM6 and TM7/HX8 of D2R and D3R. More specifically, Gi3 interacts with ALA and VAL of TM3 and with LEU of TM5 of DR complexes. The interaction of Gi3 with TM6 and TM7/HX8 involves the LYS, ALA, and MET residues, and PHE, ALA, and ILE residues, respectively. Moreover, Gi3 binds to PHE, ASN, MET, a nonpolar aliphatic residue (MET for D2R; VAL for D3R), and two hydrophilic residues (ASN and THR for D2R; GLN and HIS for D3R) of ICL2 of DR complexes. Although there is a possible involvement of basic cationic residues in the interaction of Gi3 with ICL3 of D2R (LYS) and D3R (ARG), there is no common residue pattern of interaction in this region.

Concerning the Go-DR complexes, it was observed that Go binds to TM3, ICL2, TM5, ICL3, TM6 and TM7/HX8 of D1R and D2R. More specifically, Go interacts with ALA and a nonpolar aliphatic residue (ILE for D1R; VAL for D2R) of TM3, with PRO, a nonpolar aromatic residue for D1R (PHE) or a nonpolar aliphatic residue for D2R (MET), and two hydrophilic residues (GLU and ARG for D1R; ASN and THR for D2R). Additionally, Go interacts with LYS and THR residues of TM6, and with ASN residue of TM7/HX8 of D1R and D2R. Although there is no evident residue pattern of interaction of Go with TM5 and ICL3, there is a possible involvement of nonpolar aliphatic residues and the ARG residue on the interaction with the TM5 and ICL3 motifs, respectively.

Upon analysis of the Gz-DR complexes, it was found that Gz interacts with TM3, ICL2, TM5, ICL3, TM6 and TM7/HX8 of D2R, D3R, D4R, and D5R. Regarding the TM3 motif, Gz binds to ALA and VAL of D2-like receptors and ILE of D5R. Additionally, Gz interacts with PRO, a nonpolar aliphatic residue (MET for D2R; VAL for D3R; LEU for D4R), and two hydrophilic residues (ASN and THR for D2R; GLN and HIS for D3R; ARG and ASN for D4R) of ICL2 or only with PRO for D5R. Moreover, distinct interaction patterns were found in the interaction of Gz with TM5: Gz interacts with TYR and LEU of D2R and D3R, to TRP and GLU of D4R, and with ALA of D5R. Additionally, Gz binds to PHE and ASN of TM7/HX8 of D2-like receptors, and to ASN of TM7/HX8 of D5R. Although there are many residues and a possible involvement of basic positively charged residues (LYS and ARG) in the interaction of Gz proteins with ICL3 of DR complexes, there is no evident residue pattern of interaction in this region.

Regarding the Gs(sh)-DR complexes, there is an involvement of TM3, ICL2, ICL3, TM6, and TM7/HX8 in the interaction of Gs(sh) with D1R, D3R, and D5R. More specfically, the interaction of Gs(sh) to TM3 is commanded by nonpolar aliphatic residues: ALA and ILE for D1R, ARG, ALA, VAL, and VAL for D3R, and ILE for D5R. In addition, Gs(sh) interacts with PRO, a nonpolar aromatic residue for D1-like receptors (PHE) or a nonpolar aliphatic residue for D3R (VAL), and two hydrophilic residues (GLU and ARG for D1R; GLN and HIS for D3R; LYS and ARG for D5R) of ICL2 of DR complexes. Moreover, the interaction of Gs(sh) with TM6 involves LYS and VAL for D1-like receptors, and ALA and MET for D3R. Regarding TM7/HX8, the Gs(sh) protein interacts with ASN residue. Additionally, the Gs(sh) proteins binds to PHE of TM7/HX8 of D1R and D3R. Although there are many residues and a possible involvement of ARG in the interaction of Gs(sh) proteins with the ICL3 domain of DR complexes, there is no evident residue pattern of interaction in this region.

Concerning the Gs(lo)-DR complexes, it was found that Gs(lo) interacts with TM3, ICL2, ICL3, TM6 and TM7/HX8 of D1R, D3R, and D5R. More specifically, the interaction of Gs(lo) with TM3 is commanded by nonpolar aliphatic residues: ALA and ILE for D1R, ARG, ALA, VAL, and VAL for D3R, and ILE of D5R. In addition, Gs(lo) interacts with PRO, a nonpolar aromatic residue for D1-like receptors (PHE) or a nonpolar aliphatic residue (VAL) for D3R, and two hydrophilic residues (GLU and ARG for D1R; GLN and HIS for D3R; LYS and ARG for D5R), at ICL2. The interaction of Gs(lo) with TM6 involves LYS, VAL, and THR for D1-like receptors, and MET and VAL for D3R. Regarding the TM7/HX8 domain, Gs(lo) interacts with PHE, ASN, and a nonpolar aliphatic residue (ALA for D1-like receptors; ILE for D3R) of D1R and D3R. Although LEU and ALA are the common residues found in the interface of Gs(lo)-DR complexes in ICL3, there is no evident residue pattern of interaction in this region.

The interaction of G-proteins with each DR subtype was carefully analysed and several common structural features were uncovered. Regarding the interaction of D1R with their cognate G-proteins, it was found that G-proteins interact with TM3, ICL2, ICL3, TM6, and TM7/HX8 of D1R. More specifically, all cognate G-proteins bind to ALA and ILE residues, except for Gi2, which only interacts with ILE residue of TM3 of D1R. In addition, G-proteins interact with PRO, PHE, GLU, and ARG of ICL2 of D1R, apart from Gi2, which only interacts with PRO, PHE, and ARG residues. Moreover, G-proteins bind to LYS, VAL, and THR of TM6, and to ASN and ALA of TM7/HX8 of D1R, except for Go, which only binds to ALA. Although there is no evident residue pattern of interaction concerning the ICL3 of D1R and their cognate G-proteins, a few recurring residues, particularly ALA and ARG, are observed in the interface of G-protein-D1R complexes.

Through analysis of the interaction of G-proteins with D2R, it was found that G-proteins interact with TM3, ICL2, TM5, ICL3, TM6 and TM7/HX8. More specifically, G-proteins interact with ALA, VAL, and ALA of TM3 of D2R, except for Gi3, which only interacts with ALA and VAL. In addition, G-proteins bind to MET, PRO, MET, ASN, THR, and ARG of ICL2, to LEU of TM5, and to ARG, VAL, and LYS of ICL3 of D2R, with the exception of Gi3, which does not interact with ARG and VAL. Moreover, G-proteins bind to ALA, MET, and LEU residues of TM6, apart from Gi3, in which no interaction with LEU was observed, and to PHE, ASN, and ILE of TM7/HX8 of D2R, apart from Go, which only binds to PHE and ASN.

Regarding the D3R-G-protein complexes, there is an involvement of TM3, ICL2, TM5, ICL3, TM6 and TM7/HX8 in the D3R-G-proteins interaction. More specifically, G-proteins interact with ALA and VAL of TM3, in addition Gs proteins (Gs(lo) and Gs(sh)) interact with ARG, ALA, VAL, and VAL of TM3 of D3R. Moreover, G-proteins bind to PRO, VAL, GLN, and HIS residues of ICL2, to LEU of TM5, and to PHE and ASN of TM7/HX8 of D1R, except for Gq, which only interacts with PHE. Regarding ICL3, many residues can be observed in the interface of D3R-G-proteins and the presence of many basic positively charged amino acid residues, in particular ARG, with the exception of GsL-D3R complex, in which no interaction with ARG was detected. No common residue pattern of interaction was uncovered in the ICL3. In TM6 of D3R, the interaction of G-proteins seems to be commanded by nonpolar aliphatic residues by ALA and MET, apart from GsL-D3R complex, which does not interact with ALA.

Concerning the D4R-G-protein complexes, it was found that G-proteins interact with TM3, ICL2, TM5, ICL3, TM6 and TM7/HX8. More specifically, the interaction of G-proteins with D4R involves the ALA and VAL of TM3, the PRO, LEU, ASN and ARG of ICL2, and TRP and GLU of TM5 of D4R. Additionally, G-proteins bind to LYS, LEU, HIS, GLY, ARG, ALA, GLN and ARG of ICL3 domain, to GLU, LYS, ALA, MET, VAL and LEU of TM6, and to PHE, ASN and ALA of TM7/HX8 of D4R. Moreover, Gz and Gt2 proteins interact with GLY of TM4 of D4R.

Regarding the interactions of D5R-G-protein complexes, the structural domains TM3, ICL2, ICL3, TM6 and TM7/HX8 are involved in the interaction of G-proteins with the receptor. More specifically, the G-proteins interact with ILE residue of TM3, Gs proteins (Gs(lo) and Gs(sh)) interact with PRO, PHE, LYS and ARG residues of ICL2, and Gz only binds to PRO of ICL2 of D5R. In addition, G-proteins interact with LEU, GLU, ARG, ALA, ALA and LYS residues of ICL3, with LYS and VAL of TM6, and with ASN of TM7/HX8 of D5R. Moreover, Gz proteins interact with ALA of TM5 of D5R.

4.1.4D Root-Mean-Square Deviation

RMSD measurements were conducted in order to assess the most relevant changes between the templates and the modelled GPCR structures. These RMSD measurements target the substructures, in order to understand which ones are more highly affected by GPCR involvement in complex-related interactions. The results can be found in the web-server at the 'STRUCTURAL RMSD' subtab of the 'DR COMPLEX STRUCTURE' tab.

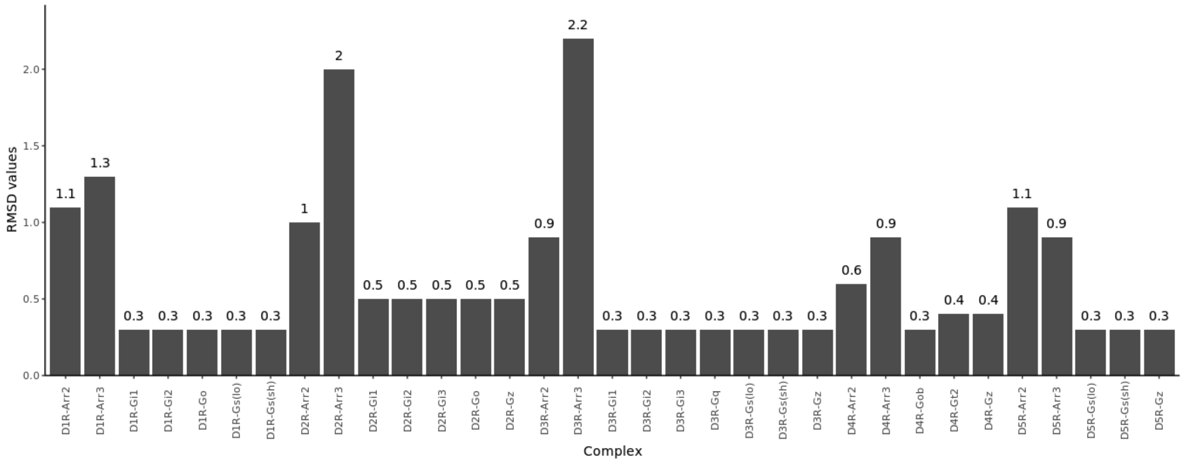


Figure 13: ICL2 RMSD in comparison with the templates, for all the complexes

Figure 13, above, highlights clearly that arrestins induce larger changes in ICL2 structure than G-proteins, for which the RMSD values are not as relevant. Furthermore, particular substructures of some complexes show higher mobility. For instance, D4R, at ICL3, seems to be a lot more affected by G-protein coupling than the remaining complexes. Additionally, at ICL1, G-proteins seem to have a greater affect in the GPCR structure than in any other system.

4.2 Hot-Spot prediction

The two datasets considered were subjected to all the six pre-processing approaches previously described and exposed in table 9: PCA, PCA Down, PCA Up, Scaled, Scaled Down and Scaled Up. We have used a 10 folded 10-cross-validation, using 70% of the instances for training and the remaining 30% for testing. Apart from the simpler ML models enumerated before, some more should be mentioned, since they are highly recurrent in the upcoming results: ORFsvm, PDA, svmPoly, PLR and bagEarth. 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. 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 themselves. PLR is a regular logistic regression with both L1 and L2 regularization. Finally, the algorithms indicated as bagEarth are bagging algorithms that make use of MARS to perform regression analysis.

4.2.1 Hot-Spot prediction from the original dataset – without coevolution

Dataset selection and treatment as well as performance estimation are still major challenges in the application of ML to this field. To propose a general methodology, it is necessary to compare the performance of various algorithms and different data extraction techniques. Some classifiers (linear discriminant analysis or generalized linear models) come from statistics, others come from data mining (tree-based), and some are connectionist approaches (such as neural networks). All can behave differently when applied to different datasets. So, identifying the best classifier for a given problem is crucial, as the No-Free-Lunch Theorem from Wolpert [233] states: “*The best classifier may not be the same for all the datasets”.* In this work, structure- and sequence-based features were combined to evaluate 51 classifiers and compare their performance on six differently pre-processed datasets. These classifiers were subjected to hierarchical clustering and grouped into 5 different clusters. We have compared the algorithms’ performance in each cluster and chosen the best of each for a global comparison. Within Cluster I, the top performance methods were either based on neuronal networks (avNNet) or on random forests (rf, RRFglobal). While avNNet, a simple shallow neural network, and rf, a forest composed of decision trees, are somewhat simple methods, RRFglobal is a regularized version of a basic random forest, capable of selecting the best feature subset with higher accuracy. Within Cluster II, the best methods were either bagging (bagEarth and bagEarthSVM), support vector machines-based (ORFsvm) or additive logistic regression models (ada). Bagging (bootstrap aggregating) generates several training subsets out of the original training set and performs a majority vote of all models. ORFsvm uses oblique decision trees which can split the feature space obliquely instead of using solely axis-parallel feature space splitting enabling a finer tuning of the model, which s explain their success. Ada uses boosting, creating an ensemble of logistic regression models, and therefore a stronger classification predictor. For Cluster III, the best results were achieved for regression models (glmboost and plr). Even though both are based on regression models, the key aspects of each is quite different as glmboost uses boosting to create an ensemble of generalized linear models, while plr uses L2 penalized regression models. L2 penalization is usually successful thanks to its ability to prevent overfitting by minimizing regression coefficients. Cluster IV was composed solely of SVM approaches. The most successful was svmPoly, which uses polynomial kernels of the original variables to construct a SVM, enabling it to act as a non-linear model. The other SVM, which was the best only in the PCA pre-sampling condition (with far worst F1-score, however), combines cost regularization that enables control over the smoothness of the fitted function, and a radial basis function that represents the input space as the distance between each vector. Cluster V features only discriminant analysis models (rda, amdai, pda and stepLDA) able to perform combinations of features for classification. Rda uses regularization to determine the best linear combination of features and fine tune their coefficients while amdai is essentially a regular discriminant predictor with slight alterations that render it capable of adapting to new classes in the testing set. Pda is a parametric discriminant classifier, which assumes a probability distribution for the population and stepLDA is a linear discriminant analysis featuring stepwise feature selection.

The original dataset, without coevolutionary features, had 533 instances, of which 125 were HS, meaning around 24% of the instances were HS. The clustering of the various ML algorithms by their common characteristics allowed us to combine their results into a ML ensemble that uses rf, svmPoly and pda. Our predictor outperforms the currently available methods in the literature with an AUROC of 0.91, sensitivity of 0.98 and specificity of 0.94 on the test set. Up-sampling of the minor class was quite effective as it allowed us to work with a balanced dataset without losing any information on the major class. This novel approach for HS prediction can now be freely applied by researchers through the SpotOn webserver (http://milou.science.uu.nl/services/SPOTON/). Table 11 shows the individual best algorithms' scores attained for the different experiments and the ensemble models for which they were used. For an overall comparison, only the individual methods will be considered.

Table 11: Best test results for the original dataset, achieved with the Scaled Up pre-processing

Pre-processing	Metric	Algorithms					Ensemble models	
Scaled Up		Cluster I	Cluster II	Cluster III	Cluster IV	Cluster V	Full Regression	rf + svmPoly + pda
		C5,0	pda	plr	rf	svmPoly		
	AUROC	0.83	0.84	0.85	0.83	0.83	0.91	0.91
	Accuracy	0.91	0.88	0.85	0.90	0.90	0.95	0.95
	Sensitivity	0.68	0.76	0.84	0.71	0.68	0.98	0.98
	Specificity	0.98	0.91	0.85	0.96	0.97	0.85	0.85
	PPV	0.90	0.73	0.64	0.84	0.87	0.95	0.95
	NPV	0.91	0.93	0.95	0.91	0.91	0.94	0.94
	FDR	0.32	0.24	0.16	0.29	0.32	0.02	0.02
	F1-score	0.78	0.74	0.73	0.77	0.76	0.97	0.97

4.2.2 Hot-Spot prediction with datasets with coevolutionary features

After taking into consideration the results for the original dataset, the dataset “Allrows”, “Fold”, “Both” and “Complex” were subjected to the same computational approach. The best results per cluster are fully displayed in tables 7 to 9 (regarding the training results) and the tables 10 to 12 (regarding the test results), in the annexes of this thesis. “Both” dataset results are not displayed as a large majority of the runs failed, likely due to the very small number of observations (55 instances). The top performance was attained for the “Fold” dataset as clearly stressed out by comparison with both “Allrows” and “Complex” datasets. The “Fold” dataset, with 65 HS (around 25% of the 264 instances), yielded considerably higher AUROC values, up to 0.97, on the Scale Up pre-processing, considering the training instances, for which the correspondent test values yielded a AUROC of 0.90. Although the training values might be higher for some methods in “Allrows” and “Complex” datasets, the corresponding test results drop drastically, as can be seen on tables 12 and 13.

However, the data must be taken into consideration by its particular characteristics, and, regarding HS detection, it is to note that, not only in the dataset used, but most likely also in any empirical protein experiment, the amount of HS will be low, usually by a large margin, in comparison to the amount of NS. This is a problem by itself, many ML models, given a much larger number of instances of one class than the others, tend to have misleading higher AUROC values. These models learn from experience and, with highly discrepantly populated classes, they do not have an amount of information adequate for their main purpose. Although the Up or Down pre-processing is useful to deal with the unbalanced data, it is still limited to the information provided. Other methods such as SMOTE [234] could be used here and may be tested in the future. As AUROC is not able to capture the purpose of specifically predicting HS, we also gave attention to Sensitivity (TPR), as it is a metric that fits these requirements. However, the overall picture should not be overlooked, simply aiming at high TPR can also lead to high FPR. So, at the end our goal was to maximize AUROC, as well as TPR.

Table 12: Metrics for the best training results

Dataset	"Fold"	"Fold"	"Allrows"	"Allrows"	"Complex"	"Complex"
Pre-processing	PCA Up	Scaled Up	PCA Up	Scaled Up	PCA Up	Scaled Up
CLUSTER	IV	IV	IV	IV	IV	IV
METHOD	ORFlog	ORFsvm	ORFlog	ORFsvm	ORFlog	ORFsvm
AUROC	0.97	0.97	0.98	0.97	0.98	0.97
Accuracy	1.00	1.00	1.00	1.00	1.00	0.99
Sensitivity	1.00	1.00	1.00	1.00	1.00	1.00
Specificity	1.00	1.00	1.00	1.00	1.00	0.99
PPV	1.00	1.00	1.00	1.00	1.00	0.99
NPV	1.00	1.00	1.00	1.00	1.00	1.00
FDR	0.00	0.00	0.00	0.00	0.00	0.01
F1-score	1.00	1.00	1.00	1.00	1.00	0.99

Table 13: Metrics for the test results corresponding to the best training results

Dataset	"Fold"	"Fold"	"Allrows"	"Allrows"	"Complex"	"Complex"
Pre-processing	PCA Up	Scaled Up	PCA Up	Scaled Up	PCA Up	Scaled Up
CLUSTER	IV	IV	IV	IV	IV	IV
METHOD	ORFlog	ORFsvm	ORFlog	ORFsvm	ORFlog	ORFsvm
AUROC	0.82	0.90	0.70	0.72	0.71	0.66
Accuracy	0.82	0.88	0.77	0.78	0.80	0.76
Sensitivity	0.43	0.62	0.36	0.43	0.50	0.50
Specificity	0.96	0.98	0.91	0.91	0.89	0.83
PPV	0.82	0.93	0.60	0.62	0.56	0.45
NPV	0.82	0.88	0.80	0.82	0.86	0.86
FDR	0.80	0.07	0.56	0.38	0.44	0.55
F1-score	0.20	0.74	0.56	0.51	0.53	0.48

Tables 12 and 13 show, respectively, the best two results for the training runs for "Fold", "Allrows" and "Complex" and the corresponding best test results. On a first notice, should be pointed out that the up-sampling pre-processing are the best training results, for all the datasets considered, which stresses out the importance of the size of the dataset, since up sampling allows the model to consider more instances than the other pre-processing approaches. Additionally, another overall point to stress is the dominance of cluster IV, which makes sense, since it is constituted by ORF approaches, that employ ensemble methods, which therefore tends to maximize the use of several models.

Although the previous statements are straight forward in agreement with what was expected, both the value of having datasets with higher samplings and the value of ensemble models, the fact that the "Fold" dataset outperforms the "Allrows" dataset, even though it has only around half of the instances of "Allrows", strikes as unusual. The "Allrows" performance, although not bad on the overall picture, is unexpected, since this dataset had the same number of observations as the SpotON dataset with a few additional features: Consurf, CoeViz and InterEV. By performing worse than most of the five best methods assessed on the original dataset, it is suggested that the new features, included, caused the decrease in performance.

The two examples regarding the “Fold” dataset listed in tables 12 and 13 showed relevantly better results in most metrics. This dataset was composed of 265 observations, and therefore was bigger than other datasets tested, with the exception of “Allrows”. ORFsvm displayed the higher AUROC, and performed well in almost every metric. In the end, the best test scores, which are listed in table 14, were achieved on “Fold” dataset.

Table 14: Best achieved scores by cluster at “Fold” dataset

Pre-processing	Metrics	Algorithms					Pre-processing	
CLUSTER		I	II	III	IV	V	Scaled Up	IV
PCA		bagEarth	lda	glmboost	ORFlog	svmLinear		ORFsvm
	AUROC	0.86	0.82	0.72	0.89	0.81		0.90
	Accuracy	0.86	0.83	0.76	0.87	0.78		0.88
	TPR	0.57	0.52	0.19	0.57	0.24		0.62
	TNR	0.96	0.95	0.96	0.98	0.98		0.98
	PPV	0.86	0.79	0.67	0.92	0.83		0.93
	NPV	0.86	0.84	0.76	0.86	0.78		0.88
	FDR	0.14	0.21	0.33	0.08	0.17		0.07
	F1-score	0.69	0.63	0.30	0.71	0.37		0.74

Overall, the size of the sampling seems to have played a large role on the final outcoming. Nevertheless, this does not explain why the “Allrows” dataset performs worse than the “Fold” dataset, or, more importantly, than the original dataset without coevolution features. From this, we have to point out that the inclusion of these evolutionary-related information seems to be decreasing the performance of the methods. The explanation can be just the simple fact that the amount of genomic data used to calculated them was not enough to retrieve the overall mechanistic picture of the evolution at the binding interfaces.

4.2.3 Hot-Spot prediction with selected coevolutionary features

The problem pointed at the end of the previous subsection suggests that CoeViz, InterEV and/or Consurf might be causing the drop of performance, since they are the only difference between "Allrows" and the original dataset. It seems that they might be lowering the overall performance of the models, suggesting that these features might be intervening as confounding variables. It seems that the EVFold features improve the overall method performance. For this reason, a new dataset was built, removing CoeViz and InterEV from the "Fold" dataset. Without changing the pre-processing approach, we run ML for this dataset ("Fold*"). Its testing results were then compared with the previous "Fold" dataset, since it had the best performing models and the best scores for the original dataset. The full test results are displayed in table 13 in the annexes section.

Table 15 demonstrates that the overall AUROC for all methods increased in the new "Fold*" dataset. We observed that the overall sensitivity is greatly reduced. The cluster with the best AUROC was chosen, for both the "Fold", "Fold*" datasets and the original dataset. The results are displayed in figures 14 and 15. The best AUROC for the dataset after removing the features was achieved with the scaled pre-processing ('Fold* Scaled', in the same Figures). The best AUROC for the dataset before the removal of the features occurred with PCA pre-processing ('Fold PCA'). Finally, the best AUROC for the original dataset was achieved with the scaled up pre-processing ('Original Scaled Up'). The later dataset displays the best AUROC results (Figure 14), whereas the sensitivity is greatly reduced in comparison.

Table 15: Comparison of the scores attained before and after removing Coeviz and InterEV from "Fold" dataset

	Fold Dataset with CoeViz and InterEV ("Fold")		Fold Dataset without CoeViz and InterEV ("Fold*")	
	AUROC	Sensitivity	AUROC	Sensitivity
Average of all clusters and all pre-processing	0.76	0.57	0.78	0.50
Best of all clusters and all pre-processing	ORFsvm, Cluster IV, Scaled Up	ORFpls, Cluster IV, Scaled Down	Glmboost, Cluster III, Scaled	svmRadial, Cluster IV, Scaled Up
Value	0.90	0.86	0.92	0.84
Best pre-processing	PCA	Scaled Down	Scaled	PCA Down
Best average cluster value	0.82	0.73	0.88	0.71

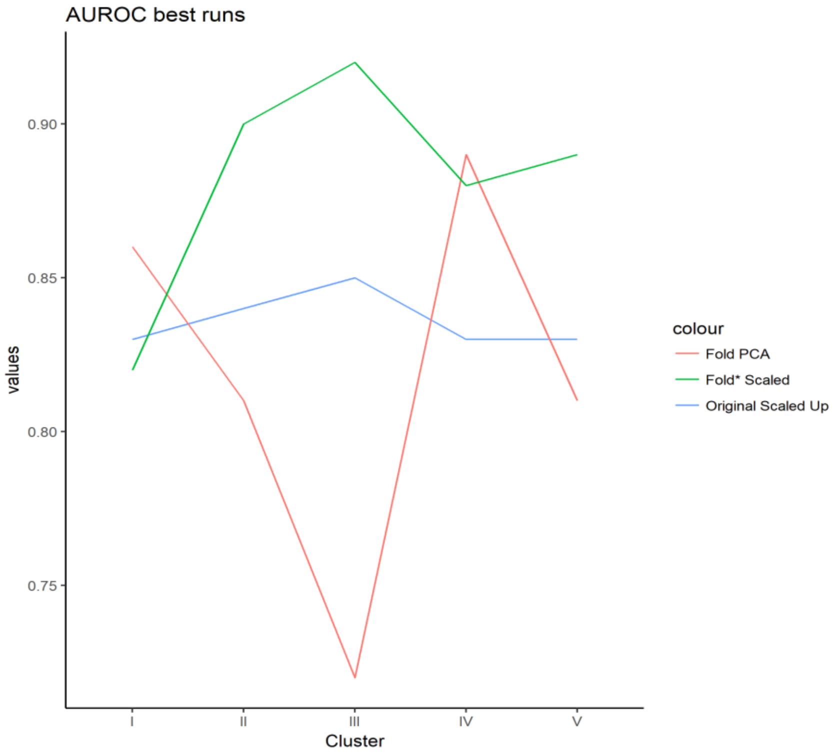


Figure 14: Best AUROC for each of the experimental conditions

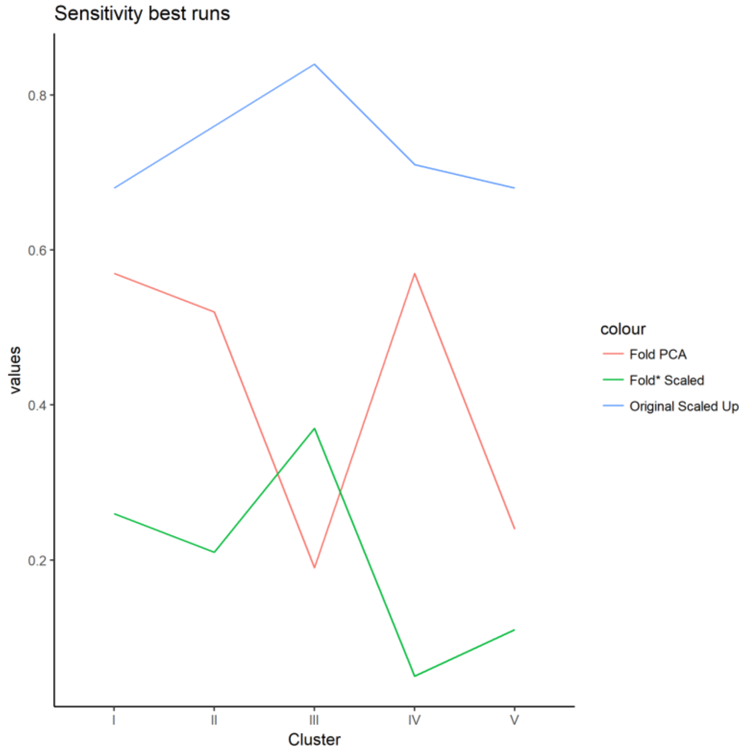


Figure 15: Sensitivity correspondent to the best AUROC for each of the attempts

The increase of AUROC displayed in the later attempt (in which they were absent), particularly in comparison with the first coevolution attempt, suggests that, as hypothesized, CoeViz and InterEV can be confounding variables. This, on the other hand, does not mean that all coevolutionary features are confounding variables, since EVFold seems to improve the AUROC. The fact that, in comparison, "Allrows" performed worse than the original, even though the only difference were the CoeViz and InterEV features, supports this statement. The improvement with "Fold*" without CoeViz and InterEV, additionally, supports the statement that the lack of instances is critical on sensitivity scores, particularly when considering that the dataset was significantly smaller, with about half of the instances and HS. A further argument that supports the possibility of CoeViz and InterEV being confounding features is the fact that, in coevolutionary runs, the bagging algorithms performed particularly well. Since bagging algorithms work by performing bootstrap aggregating, they generate different conformations of the same dataset that order the instances differently, and select different features. The selection of features is not necessarily based on correlation, but rather simply blindly tries different approaches for the same model. For this reason, since confounding variables are independent and bagging algorithms can indiscriminately leave features out, the better than usual performance of these algorithms can be due to the fact that the best performing cases were leaving some of the confounding variables aside.

5. CONCLUSIONS AND FUTURE WORK

This section will be split again in the two interconnected themes of the thesis. Regarding the GPCRs complexes, with the aim to understand and identify structural and interfacial patterns, it was noticed that the GPCRs and G-proteins present physicochemical complementarity, especially regarding residue content. Whereas most groups' percentage stays unchanged between the two interfaces of the complexes, acid negatively charged and basic positively charged residue groups show more differences, suggesting the residues belonging to this groups are some of the most important for coupling. Glutamate's high percentage at GPCR's interface suggest its role as a major contributor to complex formation. However, the same pattern was not found in D4R-G-protein complexes. Dopamine receptors, when in complex with proteins G_{i1}, G_{i2}, G_{i3} have abnormally high acid negatively charged residue percentages, suggesting these residues are important, not only at stablishing interface between G-proteins and dopamine receptors, but also at discriminating between the different G-proteins. In terms of evolutionary information, the residues at the interface were consistently more conserved than the remaining, which reports on the importance of these for the establishment of meaningful structural and functional motifs. Surface area is a good indicator of the substructures that are most involved in complex interactions. We saw that ICL1 was more related to GPCRs' interaction with arrestins than with G-proteins, particularly when the complexes involve both arrestins and D2-like receptors. Both ICL2 and ICL3 were heavily involved in interactions with both arrestins and G-proteins. All the arrestins and, in particular, complexes involving D5R showed high surface areas associated to the interaction with TM7/HX8. HB/SB content was heavily concentrated at the ICLs, as can be seen on the Heatmap displayed in Figure 11, and should be noted that the few HB/SB classified as 'Other' are present on the residues at TMs very close to the ICLs, which supports the statement that ICLs play a major role in GPCRs' interaction with G proteins and arrestins. ICL1 and HX8 have low amounts of HB/SB, although still significative. ICL3 is the substructure most heavily populated with HB/SB. G-protein and D4R complexes exhibit higher than most of others amounts of HB/SB. More than particular interacting residues, as can be seen on the circular plots, residue coupling motifs can be identified as recurring at GPCRs, when interacting with the different partners, while other motifs display a more selective role, only appearing for specific G protein types or the arrestins. This is identified for the arrestins, particular sets of G-proteins (Gs(sh) and Gs(lo); Gi1, Gi2 and Gi3; Go and Gob, Gz) and particular dopamine receptors, since, particularly between D1-like and D2-like, the differences in specific motifs arise, not only regarding the partner, but also the GPCR itself. In particular, D4R seems to adopt slightly different patterns, suggesting stronger interactions (as already did regarding HB/SB) and patterns were identified for D1-like and D2-like dopamine receptors. The patterns might not exactly concern particular amino acids, but rather physicochemical repeated motifs, such as occurs at TM7/HX8, in which three or four residues appear insistently in all complexes, however not being exactly the same residues, they exhibit similar physicochemical profiles that translates into a well-defined physicochemical pattern on this region, for all the complexes. This region seems to be a competition site for complex formation due to the fact that it appears in all the complexes assessed. ICL2 and ICL3 motifs also seem to emerge as competition promoting motifs, although more specific, since they are not the same for all the complexes but rather different specific residues or physicochemical patterns that repeatedly appear at the interaction with specific G proteins groups (Gs(sh) and Gs(lo); Gi1, Gi2 and Gi3; Go and Gob, Gz) or arrestins, laying out the importance of these patterns on the formation of the specific complexes. The main conclusions of this work will be shared with experimental partners that will experimentally test the possibility of changing the coupling of the dopamine receptors towards different partners by performing simple amino acid mutations at the interface.

Regarding the second part of the thesis, the detection of HS employing ML with and without coevolutionary features, the main aim was to surpass the previously best models, available in the literature. Here, we were able to establish a new method and a user-friendly web-server: SpotON. Due to the high relevance of coevolution in literature as a promising tool to better characterize protein-based interfaces, we have also checked if the incorporation of this kind of information would increase the performance of the attained method. The performance of the various ML algorithms in the dataset with coevolutionary features constituted of the same observations of the original dataset ("Allrows") performed was poor. This seems to indicate that, if the in-build feature selection at the different algorithms was performed correctly, the added variables, CoeViz, InterEV and ConSurf were all independent features that were leading to a decrease in performance. This suggests that, although independent, they have no causal relation to the class variable, thus not contributing in the performance of the model, but rather introducing noise in the dataset that cannot be excluded by common feature selection methods, usually based on variable independency. If these variables were indeed independent but still yield clearly worse results, they are most likely confounding variables. Considering the results of the datasets with coevolutionary features, and the hypothesis of confounding variables being present, there is another difficulty added: not all the datasets have the same amount of observations as the original dataset, due to unavailability from the external servers. In fact, the best performing dataset was the "Fold" dataset, even better than the "Allrows" dataset (the dataset with the original features and CoeViz, Consurf and InterEV), even tough it was constituted by only about one third of the samples of the original. This suggests that the EVFold features are contributing to increase the performance of the models. The "Fold" dataset, as was shown, yielded, for its best pre-processing – Scaled Up– overall better AUROC than that of the original dataset.

However, the sensitivity scores, relevant due to the final aim of predicting HS, were lower. Finally, was performed the same run, for the “Fold” dataset, upon removal of both CoeViz and InterEV (Fold*). Once again, the AUROC increased, whereas the sensitivity decreased. Overall, the two issues at hand, possibility of confounding variables and small sampling need to be tackled to properly understand the potential of coevolutionary features on HS prediction. Thus, future work will pass by looking in more details for the literature, maybe using text-mining algorithms to increase the biological sampling. On the same time, it would be important that the amount of protein sequences necessary in all coevolution methods would increase. However, this fact depends on experimental sequence and it is out of our control as computational bioinformatics.

6. REFERENCES

1. Berg JM, T.J., Stryer L, *Protein Structure and Function*, W.H. Freeman, Editor. 2002: New York.
2. Alberts B, J.A., Lewis J, et al, *Protein Function*, in *Molecular Biology of the Cell*, G. Science, Editor. 2002: New York.
3. Yan, C., et al., *Characterization of Protein–Protein Interfaces.* The protein journal, 2008. **27**(1): p. 59-70.
4. Bendell, C.J., et al., *Transient protein-protein interface prediction: datasets, features, algorithms, and the RAD-T predictor.* BMC Bioinformatics, 2014. **15**: p. 82-82.
5. Jones, S. and J.M. Thornton, *Principles of protein-protein interactions.* Proc Natl Acad Sci U S A, 1996. **93**(1): p. 13-20.
6. Jones, S. and J.M. Thornton, *Analysis of protein-protein interaction sites using surface patches.* J Mol Biol, 1997. **272**(1): p. 121-32.
7. Morrow, J.K. and S. Zhang, *Computational Prediction of Hot Spot Residues.* Current pharmaceutical design, 2012. **18**(9): p. 1255-1265.
8. Thorn, K.S. and A.A. Bogan, *ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions.* Bioinformatics, 2001. **17**(3): p. 284-5.
9. Xia, J., et al., *Predicting hot spots in protein interfaces based on protrusion index, pseudo hydrophobicity and electron-ion interaction pseudopotential features.* Oncotarget, 2016. **7**(14): p. 18065-18075.
10. Tuncbag, N., O. Keskin, and A. Gursoy, *HotPoint: hot spot prediction server for protein interfaces.* Nucleic Acids Research, 2010. **38**(Web Server issue): p. W402-W406.
11. Morrow, J.K. and S. Zhang, *Computational prediction of protein hot spot residues.* Curr Pharm Des, 2012. **18**(9): p. 1255-65.
12. Chen, P., et al., *Accurate prediction of hot spot residues through physicochemical characteristics of amino acid sequences.* Proteins, 2013. **81**(8): p. 1351-62.
13. Moreira, I.S., et al., *SpotOn: High Accuracy Identification of Protein-Protein Interface Hot-Spots.* Scientific Reports, 2017. **7**: p. 8007.
14. Almeida, J.G., et al., *Membrane proteins structures: A review on computational modeling tools.* Biochim Biophys Acta, 2017. **1859**(10): p. 2021-2039.
15. MÜLler, D.J., N.A.N. Wu, and K. Palczewski, *Vertebrate Membrane Proteins: Structure, Function, and Insights from Biophysical Approaches.* Pharmacological reviews, 2008. **60**(1): p. 43-78.
16. Pucci, B., M. Kasten, and A. Giordano, *Cell Cycle and Apoptosis.* Neoplasia (New York, N.Y.), 2000. **2**(4): p. 291-299.
17. Gromiha, M.M. and Y.Y. Ou, *Bioinformatics approaches for functional annotation of membrane proteins.* Brief Bioinform, 2014. **15**(2): p. 155-68.
18. Wallin, E. and G. von Heijne, *Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms.* Protein Science, 1998. **7**(4): p. 1029-1038.
19. Yildirim, M.A., et al., *Drug-target network.* Nature Biotechnology, 2007. **25**(10): p. 1119-1126.
20. Boucher, J., A. Kleinridders, and C.R. Kahn, *Insulin Receptor Signaling in Normal and Insulin-Resistant States.* Cold Spring Harbor Perspectives in Biology, 2014. **6**(1): p. a009191.
21. Lovinger, D.M., *Communication networks in the brain: neurons, receptors, neurotransmitters, and alcohol.* Alcohol Res Health, 2008. **31**(3): p. 196-214.
22. Seaton, B.A. and M.F. Roberts, *Peripheral Membrane Proteins*, in *Biological Membranes: A Molecular Perspective from Computation and Experiment*, K.M. Merz and B. Roux, Editors. 1996, Birkhäuser Boston: Boston, MA. p. 355-403.
23. Whited, A.M. and A. Johs, *The interactions of peripheral membrane proteins with biological membranes.* Chem Phys Lipids, 2015. **192**: p. 51-9.
24. Monje-Galvan, V. and J.B. Klauda, *Peripheral membrane proteins: Tying the knot between experiment and computation.* Biochim Biophys Acta, 2016. **1858**(7 Pt B): p. 1584-93.

25. Lenaz, G., *Lipid fluidity and membrane protein dynamics.* Biosci Rep, 1987. **7**(11): p. 823-37.
26. Tice, C.M. and Y.J. Zheng, *Non-canonical modulators of nuclear receptors.* Bioorg Med Chem Lett, 2016. **26**(17): p. 4157-64.
27. Alberts B, J.A., Lewis J, et al., *Membrane Proteins*, in *Molecular Biology of the Cell*, G. Science, Editor. 2002: New York.
28. Knorre, D.G., N.V. Kudryashova, and T.S. Godovikova, *Chemical and functional aspects of posttranslational modification of proteins.* Acta Naturae, 2009. **1**(3): p. 29-51.
29. White, S.H. and W.C. Wimley, *Membrane protein folding and stability: physical principles.* Annu Rev Biophys Biomol Struct, 1999. **28**: p. 319-65.
30. Schulz, G.E., *Transmembrane beta-barrel proteins.* Adv Protein Chem, 2003. **63**: p. 47-70.
31. Peyronnet, R., et al., *Mechanosensitive channels: feeling tension in a world under pressure.* Front Plant Sci, 2014. **5**.
32. Haswell, E.S., R. Phillips, and D.C. Rees, *Mechanosensitive channels: what can they do and how do they do it?* Structure, 2011. **19**(10): p. 1356-69.
33. Linton, K.J., *Structure and function of ABC transporters.* Physiology (Bethesda), 2007. **22**: p. 122-30.
34. Fredriksson, R., et al., *The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints.* Mol Pharmacol, 2003. **63**(6): p. 1256-72.
35. Kobilka, B.K., *G Protein Coupled Receptor Structure and Activation.* Biochimica et biophysica acta, 2007. **1768**(4): p. 794-807.
36. Arango-Lievano, M., et al., *A GIPC1-Palmitate Switch Modulates Dopamine Drd3 Receptor Trafficking and Signaling.* 2016. **36**(6): p. 1019-31.
37. Sensoy, O. and H. Weinstein, *A mechanistic role of Helix 8 in GPCRs: Computational modeling of the dopamine D2 receptor interaction with the GIPC1–PDZ-domain.* Biochimica et Biophysica Acta (BBA) - Biomembranes, 2015. **1848**(4): p. 976-983.
38. Kobilka, B.K., *G protein coupled receptor structure and activation.* Biochim Biophys Acta, 2007. **1768**(4): p. 794-807.
39. Ding, X., X. Zhao, and A. Watts, *G-protein-coupled receptor structure, ligand binding and activation as studied by solid-state NMR spectroscopy.* Biochem J, 2013. **450**(3): p. 443-57.
40. WL, D., *The PyMOL molecular graphics system.* . 2002.
41. Rasmussen, S.G., et al., *Crystal structure of the beta2 adrenergic receptor-Gs protein complex.* Nature, 2011. **477**(7366): p. 549-55.
42. Park, J.H., et al., *Opsin, a structural model for olfactory receptors?* Angew Chem Int Ed Engl, 2013. **52**(42): p. 11021-4.
43. Heng, B.C., D. Aubel, and M. Fussenegger, *An overview of the diverse roles of G-protein coupled receptors (GPCRs) in the pathophysiology of various human diseases.* Biotechnol Adv, 2013. **31**(8): p. 1676-94.
44. Isberg, V., et al., *GPCRdb: an information system for G protein-coupled receptors.* Nucleic Acids Research, 2016. **44**(Database issue): p. D356-D364.
45. Becker, O.M., et al., *Modeling the 3D structure of GPCRs: advances and application to drug discovery.* Curr Opin Drug Discov Devel, 2003. **6**(3): p. 353-61.
46. Tsvetanova, N.G., R. Irannejad, and M. von Zastrow, *G Protein-coupled Receptor (GPCR) Signaling via Heterotrimeric G Proteins from Endosomes.* J Biol Chem, 2015. **290**(11): p. 6689-96.
47. Johnston, C.A. and D.P. Siderovski, *Receptor-Mediated Activation of Heterotrimeric G-Proteins: Current Structural Insights.* Molecular Pharmacology, 2007. **72**(2): p. 219-230.
48. Patel, T.B., *Single Transmembrane Spanning Heterotrimeric G Protein-Coupled Receptors and Their Signaling Cascades.* Pharmacological Reviews, 2004. **56**(3): p. 371-385.
49. Maurice, P., et al., *GPCR-interacting proteins, major players of GPCR function.* Adv Pharmacol, 2011. **62**: p. 349-80.

50. Smith, J.S. and S. Rajagopal, *The beta-Arrestins: Multifunctional Regulators of G Protein-coupled Receptors.* J Biol Chem, 2016. **291**(17): p. 8969-77.

51. Dromey, J.R. and K.D. Pfleger, *G protein coupled receptors as drug targets: the role of beta-arrestins.* Endocr Metab Immune Disord Drug Targets, 2008. **8**(1): p. 51-61.

52. Tuteja, N., *Signaling through G protein coupled receptors.* Plant Signaling & Behavior, 2009. **4**(10): p. 942-947.

53. Sengupta, D. and A. Chattopadhyay, *Molecular dynamics simulations of GPCR–cholesterol interaction: An emerging paradigm.* Biochimica et Biophysica Acta (BBA) - Biomembranes, 2015. **1848**(9): p. 1775-1782.

54. Scarselli, M. and J.G. Donaldson, *Constitutive Internalization of G Protein-coupled Receptors and G.* J Biol Chem, 2009. **284**(6): p. 3577-85.

55. Missale, C., et al., *Dopamine receptors: from structure to function.* Physiol Rev, 1998. **78**(1): p. 189-225.

56. Jaber, M., et al., *Dopamine receptors and brain function.* Neuropharmacology, 1996. **35**(11): p. 1503-19.

57. Beaulieu, J.M. and R.R. Gainetdinov, *The physiology, signaling, and pharmacology of dopamine receptors.* Pharmacol Rev, 2011. **63**(1): p. 182-217.

58. Moreira, I.S., et al., *Structural Basis of Dopamine Receptor Activation*, in *The Dopamine Receptors*, K.A. Neve, Editor. 2010, Humana Press: Totowa, NJ. p. 47-73.

59. Arnam, E.B.V., et al., *Investigation of Dopamine Receptor Structure and Function by Structure Prediction and Unnatural Amino Acid Mutagenesis.* Biophysical Journal. **102**(3): p. 247a.

60. Salmas, R.E., et al., *Modeling and protein engineering studies of active and inactive states of human dopamine D2 receptor (D2R) and investigation of drug/receptor interactions.* Mol Divers, 2015. **19**(2): p. 321-32.

61. Durdagi, S., et al., *Binding Interactions of Dopamine and Apomorphine in D2High and D2Low States of Human Dopamine D2 Receptor Using Computational and Experimental Techniques.* ACS Chem Neurosci, 2016. **7**(2): p. 185-95.

62. Lee, A.G., *How lipids affect the activities of integral membrane proteins.* Biochimica et Biophysica Acta (BBA) - Biomembranes, 2004. **1666**(1–2): p. 62-87.

63. Rawlings, A.E., *Membrane proteins: always an insoluble problem?* Biochem Soc Trans, 2016. **44**(3): p. 790-5.

64. Moraes, I., et al., *Membrane protein structure determination - the next generation.* Biochim Biophys Acta, 2014. **1838**(1 Pt A): p. 78-87.

65. Alonso, M.A. and J. Millán, *The role of lipid rafts in signalling and membrane trafficking in T lymphocytes.* Journal of Cell Science, 2001. **114**(22): p. 3957-3965.

66. Mao, H.-B., et al., *Effects of glycerol and high temperatures on structure and function of phycobilisomes in Synechocystis sp. PCC 6803.* FEBS Letters, 2003. **553**(1–2): p. 68-72.

67. Brown, D.A. and E. London, *Functions of lipid rafts in biological membranes.* Annu Rev Cell Dev Biol, 1998. **14**: p. 111-36.

68. 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, 1997. **94**(21): p. 11375-11380.

69. Zhang, Y.P., et al., *Interaction of a peptide model of a hydrophobic transmembrane alpha-helical segment of a membrane protein with phosphatidylcholine bilayers: differential scanning calorimetric and FTIR spectroscopic studies.* Biochemistry, 1992. **31**(46): p. 11579-88.

70. Grau-Campistany, A., et al., *Hydrophobic mismatch demonstrated for membranolytic peptides, and their use as molecular rulers to measure bilayer thickness in native cells.* Sci Rep, 2015. **5**: p. 9388.

71. Lebowitz, J., M.S. Lewis, and P. Schuck, *Modern analytical ultracentrifugation in protein science: A tutorial review.* Protein Sci, 2002. **11**(9): p. 2067-79.

72. Weiss, T.M., et al., *Hydrophobic Mismatch between Helices and Lipid Bilayers.* Biophys J, 2003. **84**(1): p. 379-85.

73. Webb, R.J., et al., *Hydrophobic mismatch and the incorporation of peptides into lipid bilayers: a possible mechanism for retention in the Golgi.* Biochemistry, 1998. **37**(2): p. 673-9.

74. Engelman, D.M., et al., *Membrane protein folding: beyond the two stage model.* FEBS Letters, 2003. **555**(1): p. 122-125.

75. Doerr, A., *Membrane protein structures.* Nat Meth, 2009. **6**(1): p. 35-35.

76. Moraes, I., et al., *Membrane protein structure determination — The next generation.* Biochimica et Biophysica Acta (BBA) - Biomembranes, 2014. **1838**(1, Part A): p. 78-87.

77. Grouleff, J., et al., *The influence of cholesterol on membrane protein structure, function, and dynamics studied by molecular dynamics simulations.* Biochimica et Biophysica Acta (BBA) - Biomembranes, 2015. **1848**(9): p. 1783-1795.

78. Grouleff, J., et al., *The influence of cholesterol on membrane protein structure, function, and dynamics studied by molecular dynamics simulations.* Biochim Biophys Acta, 2015. **1848**(9): p. 1783-95.

79. Hong, H., Y.C. Chang, and J.U. Bowie, *Measuring transmembrane helix interaction strengths in lipid bilayers using steric trapping.* Methods Mol Biol, 2013. **1063**: p. 37-56.

80. Hopf, Thomas A., et al., *Three-Dimensional Structures of Membrane Proteins from Genomic Sequencing.* Cell. **149**(7): p. 1607-1621.

81. Kall, L., A. Krogh, and E.L. Sonnhammer, *A combined transmembrane topology and signal peptide prediction method.* J Mol Biol, 2004. **338**(5): p. 1027-36.

82. Khadria, A. and A. Senes, *Measurement of transmembrane peptide interactions in liposomes using Forster resonance energy transfer (FRET).* Methods Mol Biol, 2013. **1063**: p. 19-36.

83. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Res, 2000. **28**(1): p. 235-42.

84. Carpenter, E.P., et al., *Overcoming the challenges of membrane protein crystallography.* Current Opinion in Structural Biology, 2008. **18**(5): p. 581-586.

85. Cherezov, V., et al., *A robotic system for crystallizing membrane and soluble proteins in lipidic mesophases.* Acta Crystallographica Section D, 2004. **60**: p. 1795-1807.

86. Lund, S., et al., *Detergent structure and associated lipid as determinants in the stabilization of solubilized Ca2+-ATPase from sarcoplasmic reticulum.* J Biol Chem, 1989. **264**(9): p. 4907-15.

87. Postis, V.L., et al., *A high-throughput assay of membrane protein stability.* Mol Membr Biol, 2008. **25**(8): p. 617-24.

88. Gluck, J.M., et al., *Integral membrane proteins in nanodiscs can be studied by solution NMR spectroscopy.* J Am Chem Soc, 2009. **131**(34): p. 12060-1.

89. Caffrey, M., *A comprehensive review of the lipid cubic phase or in meso method for crystallizing membrane and soluble proteins and complexes.* Acta Crystallogr F Struct Biol Commun, 2015. **71**(Pt 1): p. 3-18.

90. Cherezov, V. and M. Caffrey, *Membrane protein crystallization in lipidic mesophases. A mechanism study using X-ray microdiffraction.* Faraday Discuss, 2007. **136**: p. 195-212; discussion 213-29.

91. Hunte, C., et al., *Structure at 2.3 Å resolution of the cytochrome bc1 complex from the yeast Saccharomyces cerevisiae co-crystallized with an antibody Fv fragment.* Structure, 2000. **8**(6): p. 669-684.

92. Liang, B. and L. Tamm, *NMR as a tool to investigate the structure, dynamics and function of membrane proteins.* Nat Struct Mol Biol, 2016. **23**(6): p. 468-474.

93. Oxenoid, K. and J.J. Chou, *A functional NMR for membrane proteins: dynamics, ligand binding, and allosteric modulation.* Protein science : a publication of the Protein Society, 2016. **25**(5): p. 959-73.

94. Hong, M., Y. Zhang, and F. Hu, *Membrane Protein Structure and Dynamics from NMR Spectroscopy.* Annual review of physical chemistry, 2012. **63**: p. 1-24.

95. Murray, D.T., N. Das, and T.A. Cross, *Solid State NMR Strategy for Characterizing Native Membrane Protein Structures.* Accounts of Chemical Research, 2013. **46**(9): p. 2172-2181.

96. Shahid, S.A., et al., *Membrane-protein structure determination by solid-state NMR spectroscopy of microcrystals.* Nat Meth, 2012. **9**(12): p. 1212-1217.

97. Watts, A., et al., *Membrane Protein Structure Determination Using Solid-State NMR BT - Protein NMR Techniques*, A.K. Downing, Editor. 2004, Humana Press: Totowa, NJ. p. 403-473.

98. Wang, S., et al., *Paramagnetic relaxation enhancement reveals oligomerization interface of a membrane protein.* J Am Chem Soc, 2012. **134**(41): p. 16995-8.

99. Ganguly, S., B.E. Weiner, and J. Meiler, *Membrane Protein Structure Determination using Paramagnetic Tags.* Structure, 2011. **19**(4): p. 441-3.

100. Kaplan, M., et al., *Nuclear magnetic resonance (NMR) applied to membrane-protein complexes.* Q Rev Biophys, 2016. **49**: p. e15.

101. Milne, J.L., et al., *Cryo-electron microscopy--a primer for the non-microscopist.* FEBS J, 2013. **280**(1): p. 28-45.

102. Sliwoski, G., et al., *Computational Methods in Drug Discovery.* Pharmacological Reviews, 2014. **66**(1): p. 334-395.

103. Lee, A.G., *How lipids affect the activities of integral membrane proteins.* Biochimica et biophysica acta, 2004. **1666**(1-2): p. 62-87.

104. Moult, J., et al., *Critical assessment of methods of protein structure prediction (CASP) — round x.* Proteins: Structure, Function, and Bioinformatics, 2014. **82**: p. 1-6.

105. Ebejer, J.P., et al., *Memoir: template-based structure prediction for membrane proteins.* Nucleic Acids Res, 2013. **41**(Web Server issue): p. W379-83.

106. Kelm, S., J. Shi, and C.M. Deane, *MEDELLER: homology-based coordinate generation for membrane proteins.* Bioinformatics, 2010. **26**(22): p. 2833-40.

107. Webb, B. and A. Sali, *Protein structure modeling with MODELLER.* Methods Mol Biol, 2014. **1137**: p. 1-15.

108. Kozma, D. and G.E. Tusnady, *TMFoldWeb: a web server for predicting transmembrane protein fold class.* Biol Direct, 2015. **10**: p. 54.

109. Kozma, D. and G.E. Tusnady, *TMFoldRec: a statistical potential-based transmembrane protein fold recognition tool.* BMC Bioinformatics, 2015. **16**: p. 201.

110. Yarov-Yarovoy, V., J. Schonbrun, and D. Baker, *Multipass membrane protein structure prediction using Rosetta.* Proteins, 2006. **62**(4): p. 1010-25.

111. Teixeira, P.L., et al., *Membrane protein contact and structure prediction using co-evolution in conjunction with machine learning.* PLoS One, 2017. **12**(5): p. e0177866.

112. Baştanlar, Y. and M. Özuysal, *Introduction to Machine Learning*, in *miRNomics: MicroRNA Biology and Computational Analysis*, M. Yousef and J. Allmer, Editors. 2014, Humana Press: Totowa, NJ. p. 105-128.

113. Lavecchia, A., *Machine-learning approaches in drug discovery: methods and applications.* Drug Discovery Today, 2015. **20**(3): p. 318-331.

114. McGuffin, L.J., K. Bryson, and D.T. Jones, *The PSIPRED protein structure prediction server.* Bioinformatics (Oxford, England), 2000. **16**(4): p. 404-405.

115. Cid, H., et al., *Prediction of secondary structure of proteins by means of hydrophobicity profiles.* FEBS Letters, 1982. **150**(1): p. 247-254.

116. Koehler Leman, J., M.B. Ulmschneider, and J.J. Gray, *Computational modeling of membrane proteins.* Proteins, 2015. **83**(1): p. 1-24.

117. Hessa, T., et al., *Recognition of transmembrane helices by the endoplasmic reticulum translocon.* Nature, 2005. **433**(7024): p. 377-381.

118. von Heijne, G., *Membrane-protein topology.* Nat Rev Mol Cell Biol, 2006. **7**(12): p. 909-18.

119. Marks, D.S., et al., *Protein 3D Structure Computed from Evolutionary Sequence Variation.* PLoS One, 2011. **6**(12).

120. Baker, F.N. and A. Porollo, *CoeViz: a web-based tool for coevolution analysis of protein residues.* BMC Bioinformatics, 2016. **17**(1): p. 119.

121. Andreani, J., G. Faure, and R. Guerois, *InterEvScore: a novel coarse-grained interface scoring function using a multi-body statistical potential coupled to evolution.* Bioinformatics, 2013. **29**(14): p. 1742-9.

122. Viklund, H. and A. Elofsson, *OCTOPUS: improving topology prediction by two-track ANN-based preference scores and an extended topological grammar.* Bioinformatics (Oxford, England), 2008. **24**(15): p. 1662-1668.

123. Eddy, S.R., *What is a hidden Markov model?* Nat Biotech, 2004. **22**(10): p. 1315-1316.

124. Jones, D.T., W.R. Taylor, and J.M. Thornton, *A Model Recognition Approach to the Prediction of All-Helical Membrane Protein Structure and Topology.* Biochemistry, 1994. **33**(10): p. 3038-3049.

125. Jones, D.T., *Improving the accuracy of transmembrane protein topology prediction using evolutionary information.* Bioinformatics (Oxford, England), 2007. **23**(5): p. 538-544.

126. Nugent, T. and D.T. Jones, *Transmembrane protein topology prediction using support vector machines.* BMC Bioinformatics, 2009. **10**: p. 159-159.

127. Hayat, S., et al., *Inclusion of dyad-repeat pattern improves topology prediction of transmembrane beta-barrel proteins.* Bioinformatics (Oxford, England), 2016. **32**(10): p. 1571-1573.

128. Feinauer, C., et al., *Inter-Protein Sequence Co-Evolution Predicts Known Physical Interactions in Bacterial Ribosomes and the Trp Operon.* PLoS One, 2016. **11**(2): p. e0149166.

129. Jones, D.T., et al., *MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins.* Bioinformatics, 2015. **31**(7): p. 999-1006.

130. Rost, B. and C. Sander, *Combining evolutionary information and neural networks to predict protein secondary structure.* Proteins: Structure, Function, and Bioinformatics, 1994. **19**(1): p. 55-72.

131. Capra, J.A., et al., *Predicting Protein Ligand Binding Sites by Combining Evolutionary Sequence Conservation and 3D Structure.* PLoS Comput Biol, 2009. **5**(12): p. e1000585-e1000585.

132. Capra, J.A. and M. Singh, *Predicting functionally important residues from sequence conservation.* Bioinformatics (Oxford, England), 2007. **23**(15): p. 1875-82.

133. Ng, J., et al., *Evolutionary conservation and predicted structure of the Drosophila extra sex combs repressor protein.* Molecular and Cellular Biology, 1997. **17**(11): p. 6663-6672.

134. Ashkenazy, H., et al., *ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules.* Nucleic Acids Res, 2016. **44**(Web Server issue): p. W344-50.

135. Buslje, C.M., et al., *Correction for phylogeny, small number of observations and data redundancy improves the identification of coevolving amino acid pairs using mutual information.* Bioinformatics, 2009. **25**(9): p. 1125-31.

136. Ciancetta, A., et al., *Advances in Computational Techniques to Study GPCR-Ligand Recognition.* Trends Pharmacol Sci, 2015. **36**(12): p. 878-90.

137. De Juan, D., F. Pazos, and A. Valencia, *Emerging methods in protein co-evolution.* Nat Rev Genet, 2013. **14**.

138. Dehzangi, A., et al., *Proposing a highly accurate protein structural class predictor using segmentation-based features.* BMC Genomics, 2014. **15**(Suppl 1).

139. Dekker, J.P., et al., *A perturbation-based method for calculating explicit likelihood of evolutionary co-variance in multiple sequence alignments.* Bioinformatics, 2004. **20**.

140. Figliuzzi, M., et al., *Coevolutionary Landscape Inference and the Context-Dependence of Mutations in Beta-Lactamase TEM-1.* Mol Biol Evol, 2016. **33**.

141. Fodor, A.A. and R.W. Aldrich, *On evolutionary conservation of thermodynamic coupling in proteins.* J Biol Chem, 2004. **279**.

142. Yip, K.Y., et al., *An integrated system for studying residue coevolution in proteins.* Bioinformatics, 2008. **24**.

143. Yu, J., et al., *InterEvDock: a docking server to predict the structure of protein-protein interactions using evolutionary information.* Nucleic Acids Res, 2016. **44**(W1): p. W542-9.

144. Xiang, Z., *Advances in Homology Protein Structure Modeling.* Current protein & peptide science, 2006. **7**(3): p. 217-227.

145. Nugent, T., *De novo membrane protein structure prediction.* Methods Mol Biol, 2015. **1215**: p. 331-50.

146. Sompornpisut, P., B. Roux, and E. Perozo, *Structural Refinement of Membrane Proteins by Restrained Molecular Dynamics and Solvent Accessibility Data.* Biophysical Journal, 2008. **95**(11): p. 5349-5361.

147. Hogeweg, P. and B. Hesper, *The alignment of sets of sequences and the construction of phyletic trees: an integrated method.* J Mol Evol, 1984. **20**(2): p. 175-86.

148. Sievers, F., et al., *Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega.* Molecular Systems Biology, 2011. **7**: p. 539-539.

149. Mount, D.W., *Using hidden Markov models to align multiple sequences.* Cold Spring Harb Protoc, 2009. **2009**(7): p. pdb.top41.

150. de Juan, D., F. Pazos, and A. Valencia, *Emerging methods in protein co-evolution.* Nat Rev Genet, 2013. **14**(4): p. 249-261.

151. Jones, D.T., et al., *PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments.* Bioinformatics, 2012. **28**(2): p. 184-90.

152. Morcos, F., et al., *Direct-coupling analysis of residue coevolution captures native contacts across many protein families.* Proceedings of the National Academy of Sciences, 2011. **108**(49): p. E1293-E1301.

153. Marks, D.S., et al., *Protein 3D structure computed from evolutionary sequence variation.* PLoS One, 2011. **6**(12): p. e28766.

154. Angermueller, C., et al., *Deep learning for computational biology.* Molecular Systems Biology, 2016. **12**(7): p. 878.

155. Vidyasagar, M., *Machine learning methods in the computational biology of cancer.* Proceedings. Mathematical, Physical, and Engineering Sciences / The Royal Society, 2014. **470**(2167): p. 20140081.

156. Wei, L. and Q. Zou, *Recent Progress in Machine Learning-Based Methods for Protein Fold Recognition.* International Journal of Molecular Sciences, 2016. **17**(12): p. 2118.

157. Larranaga, P., et al., *Machine learning in bioinformatics.* Brief Bioinform, 2006. **7**(1): p. 86-112.

158. Chou, K.C., *Some remarks on protein attribute prediction and pseudo amino acid composition.* J Theor Biol, 2011. **273**(1): p. 236-47.

159. Chen, W., et al., *PAI: Predicting adenosine to inosine editing sites by using pseudo nucleotide compositions.* Sci Rep, 2016. **6**: p. 35123.

160. Feng, P., et al., *Identifying RNA 5-methylcytosine sites via pseudo nucleotide compositions.* Mol Biosyst, 2016. **12**(11): p. 3307-3311.

161. Chen, W., et al., *RAMPred: identifying the N(1)-methyladenosine sites in eukaryotic transcriptomes.* Sci Rep, 2016. **6**: p. 31080.

162. Wu, Z., *A Review of Statistical Methods for Preprocessing Oligonucleotide Microarrays.* Statistical methods in medical research, 2009. **18**(6): p. 533-541.

163. Smolinska, A., et al., *Current breathomics--a review on data pre-processing techniques and machine learning in metabolomics breath analysis.* J Breath Res, 2014. **8**(2): p. 027105.

164. Hawkins, D.M., *The Problem of Overfitting.* Journal of Chemical Information and Computer Sciences, 2004. **44**(1): p. 1-12.

165. Erickson, B.J., et al., *Machine Learning for Medical Imaging.* Radiographics : a review publication of the Radiological Society of North America, Inc, 2017. **37**(2): p. 505-515.

166. Hira, Z.M. and D.F. Gillies, *A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data.* Advances in Bioinformatics, 2015. **2015**: p. 198363.
167. Estevez, P.A., et al., *Normalized mutual information feature selection.* IEEE Trans Neural Netw, 2009. **20**(2): p. 189-201.
168. Guo, Q., et al., *Feature selection in principal component analysis of analytical data.* Chemometrics and Intelligent Laboratory Systems, 2002. **61**(1): p. 123-132.
169. Krstajic, D., et al., *Cross-validation pitfalls when selecting and assessing regression and classification models.* Journal of Cheminformatics, 2014. **6**(1): p. 10.
170. Geisser, S., *The Predictive Sample Reuse Method with Applications.* Journal of the American Statistical Association, 1975. **70**(350): p. 320-328.
171. van den Berg, R.A., et al., *Centering, scaling, and transformations: improving the biological information content of metabolomics data.* BMC Genomics, 2006. **7**: p. 142-142.
172. Cao, X.H., I. Stojkovic, and Z. Obradovic, *A robust data scaling algorithm to improve classification accuracies in biomedical data.* BMC Bioinformatics, 2016. **17**(1): p. 359.
173. Breiman, L., *Bagging Predictors.* Machine Learning, 1996. **24**: p. 123-140.
174. Henderson, A.R., *The bootstrap: a technique for data-driven statistics. Using computer-intensive analyses to explore experimental data.* Clin Chim Acta, 2005. **359**(1-2): p. 1-26.
175. Datta, S., V. Pihur, and S. Datta, *An adaptive optimal ensemble classifier via bagging and rank aggregation with applications to high dimensional data.* BMC Bioinformatics, 2010. **11**: p. 427-427.
176. Natekin, A. and A. Knoll, *Gradient boosting machines, a tutorial.* Frontiers in Neurorobotics, 2013. **7**: p. 21.
177. Blagus, R. and L. Lusa, *Boosting for high-dimensional two-class prediction.* BMC Bioinformatics, 2015. **16**: p. 300.
178. Arsov, N., et al., *Generating highly accurate prediction hypotheses through collaborative ensemble learning.* Scientific Reports, 2017. **7**: p. 44649.
179. Zhang, Z., *Naïve Bayes classification in R.* Annals of Translational Medicine, 2016. **4**(12): p. 241.
180. Zhang, Z., *Introduction to machine learning: k-nearest neighbors.* Annals of Translational Medicine, 2016. **4**(11): p. 218.
181. Noble, W.S., *What is a support vector machine?* Nat Biotechnol, 2006. **24**(12): p. 1565-7.
182. Krogh, A., *What are artificial neural networks?* Nat Biotech, 2008. **26**(2): p. 195-197.
183. Svetnik, V., et al., *Random forest: a classification and regression tool for compound classification and QSAR modeling.* J Chem Inf Comput Sci, 2003. **43**(6): p. 1947-58.
184. Sperandei, S., *Understanding logistic regression analysis.* Biochemia Medica, 2014. **24**(1): p. 12-18.
185. Robin, X., et al., *pROC: an open-source package for R and S+ to analyze and compare ROC curves.* BMC Bioinformatics, 2011. **12**: p. 77-77.
186. Skelly, A.C., J.R. Dettori, and E.D. Brodt, *Assessing bias: the importance of considering confounding.* Evidence-Based Spine-Care Journal, 2012. **3**(1): p. 9-12.
187. Andrade, C., *Confounding.* Indian Journal of Psychiatry, 2007. **49**(2): p. 129-131.
188. Webb, B. and A. Sali, *Comparative Protein Structure Modeling Using MODELLER.* Curr Protoc Bioinformatics, 2014. **47**: p. 5 6 1-32.
189. van Zundert, G.C.P., et al., *The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes.* Journal of Molecular Biology, 2016. **428**(4): p. 720-725.
190. Grant, B.J., et al., *Bio3d: an R package for the comparative analysis of protein structures.* Bioinformatics, 2006. **22**(21): p. 2695-6.
191. Negi, S.S., et al., *InterProSurf: a web server for predicting interacting sites on protein surfaces.* Bioinformatics, 2007. **23**(24): p. 3397-9.
192. Vangone, A., et al., *COCOMAPS: a web application to analyze and visualize contacts at the interface of biomolecular complexes.* Bioinformatics, 2011. **27**(20): p. 2915-6.

193. Ashkenazy, H., et al., *ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules.* Nucleic Acids Res, 2016. **44**(W1): p. W344-50.

194. Armon, A., D. Graur, and N. Ben-Tal, *ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information.* J Mol Biol, 2001. **307**.

195. Kurkcuoglu, O., et al., *Focused Functional Dynamics of Supramolecules by Use of a Mixed-Resolution Elastic Network Model.* Biophysical Journal, 2009. **97**(4): p. 1178-1187.

196. Marti-Renom, M.A., et al., *Comparative protein structure modeling of genes and genomes.* Annu Rev Biophys Biomol Struct, 2000. **29**: p. 291-325.

197. Rasmussen, S.G.F., et al., *Crystal Structure of the β(2)Adrenergic Receptor-Gs protein complex.* Nature, 2011. **477**(7366): p. 549-555.

198. Sievers, F. and D.G. Higgins, *Clustal Omega, accurate alignment of very large numbers of sequences.* Methods Mol Biol, 2014. **1079**: p. 105-16.

199. Kang, Y., et al., *Crystal structure of rhodopsin bound to arrestin by femtosecond X-ray laser.* Nature, 2015. **523**(7562): p. 561-567.

200. van Zundert, G.C., et al., *The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes.* J Mol Biol, 2016. **428**(4): p. 720-5.

201. Ray, N., et al., *Intersurf: dynamic interface between proteins.* J Mol Graph Model, 2005. **23**(4): p. 347-54.

202. Humphrey, W., A. Dalke, and K. Schulten, *VMD: visual molecular dynamics.* J Mol Graph, 1996. **14**(1): p. 33-8, 27-8.

203. Rasmussen, S.G.F., *Crystal Structure of the β(2)Adrenergic Receptor-Gs protein complex.* 2011. **477**(7366): p. 549-55.

204. Kang, Y., et al., *Crystal structure of rhodopsin bound to arrestin by femtosecond X-ray laser.* Nature, 2015. **523**(7562): p. 561-7.

205. Ballesteros, J.A. and H. Weinstein, *[19] Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in G protein-coupled receptors.* Methods in neurosciences, 1995. **25**: p. 366-428.

206. Cock, P.J.A., et al., *Biopython: freely available Python tools for computational molecular biology and bioinformatics.* Bioinformatics, 2009. **25**(11): p. 1422-1423.

207. Gu, Z., et al., *circlize Implements and enhances circular visualization in R.* Bioinformatics, 2014. **30**(19): p. 2811-2.

208. Goldenberg, O., et al., *The ConSurf-DB: pre-calculated evolutionary conservation profiles of protein structures.* Nucleic Acids Res, 2009. **37**(Database issue): p. D323-7.

209. Wang, G. and R.L. Dunbrack, Jr., *PISCES: a protein sequence culling server.* Bioinformatics, 2003. **19**(12): p. 1589-91.

210. Katoh, K., et al., *MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform.* Nucleic Acids Research, 2002. **30**(14): p. 3059-3066.

211. Boutet, E., et al., *UniProtKB/Swiss-Prot, the Manually Annotated Section of the UniProt KnowledgeBase: How to Use the Entry View.* Methods Mol Biol, 2016. **1374**: p. 23-54.

212. Pupko, T., et al., *Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues.* Bioinformatics, 2002. **18 Suppl 1**: p. S71-7.

213. Morcos, F., et al., *Direct-coupling analysis of residue coevolution captures native contacts across many protein families.* Proc Natl Acad Sci U S A, 2011. **108**(49): p. E1293-301.

214. Wolfinger, R. and M. Oconnell, *Generalized Linear Mixed Models - a Pseudo-Likelihood Approach.* Journal of Statistical Computation and Simulation, 1993. **48**(3-4): p. 233-243.

215. Wickam, H., *ggplot2: Elegant Graphics for Data Analysis.* 2009, New York: Springer-Verlag.

216. Inc., R. *Easy web applications in R.* 2013; Available from: http://www.rstudio.com/shiny/.

217. Moreira, I.S., P.A. Fernandes, and M.J. Ramos, *Hot spots--a review of the protein-protein interface determinant amino-acid residues.* Proteins, 2007. **68**(4): p. 803-12.

218. Fischer, T.B., et al., *The binding interface database (BID): a compilation of amino acid hot spots in protein interfaces.* Bioinformatics, 2003. **19**(11): p. 1453-4.

219. Moal, I.H. and J. Fernandez-Recio, *SKEMPI: a Structural Kinetic and Energetic database of Mutant Protein Interactions and its use in empirical models.* Bioinformatics, 2012. **28**(20): p. 2600-7.

220. Kumar, M.D. and M.M. Gromiha, *PINT: Protein-protein Interactions Thermodynamic Database.* Nucleic Acids Res, 2006. **34**(Database issue): p. D195-8.

221. Miller, S., et al., *The accessible surface area and stability of oligomeric proteins.* Nature, 1987. **328**(6133): p. 834-6.

222. Humphrey, W., A. Dalke, and K. Schulten, *VMD: Visual molecular dynamics.* Journal of Molecular Graphics, 1996. **14**(1): p. 33-38.

223. Camacho, C., et al., *BLAST+: architecture and applications.* BMC Bioinformatics, 2009. **10**: p. 421-421.

224. Xiao, N., et al., *protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences.* Bioinformatics, 2015. **31**(11): p. 1857-9.

225. Kawashima, S., H. Ogata, and M. Kanehisa, *AAindex: Amino Acid Index Database.* Nucleic Acids Res, 1999. **27**(1): p. 368-9.

226. Ashkenazy, H., et al., *ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules.* Nucleic Acids Research, 2016. **44**(Web Server issue): p. W344-W350.

227. Kajan, L., et al., *FreeContact: fast and free software for protein contact prediction from residue co-evolution.* BMC Bioinformatics, 2014. **15**: p. 85.

228. Hopf, T.A., et al., *Sequence co-evolution gives 3D contacts and structures of protein complexes.* eLife, 2014. **3**: p. e03430.

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

230. Faure, G., J. Andreani, and R. Guerois, *InterEvol database: exploring the structure and evolution of protein complex interfaces.* Nucleic Acids Res, 2012. **40**(Database issue): p. D847-56.

231. Mayrose, I., et al., *Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior.* Mol Biol Evol, 2004. **21**(9): p. 1781-91.

232. Kuhn, M., *caret: Classification and regression training.* Astrophysics Source Code Library, 2015. **1**: p. 05003.

233. Meester, R., *Simulation of biological evolution and the NFL theorems.* Biol Philos, 2009. **24**(4): p. 461-72.

234. Wang, Q., et al., *A Novel Ensemble Method for Imbalanced Data Learning: Bagging of Extrapolation-SMOTE SVM.* Computational Intelligence and Neuroscience, 2017. **2017**: p. 1827016.

7. ANNEXES

Table 1: D1-5R and ARR2 complexes' interface summary

Complex			D1R-ARR2	D2R-ARR2	D3R-ARR2	D4R-ARR2	D5R-ARR2
POLAR area/energy			11562.14	11629.42	11443.62	11586.21	12226.51
APOLAR area/energy			21608.76	22003.05	21862.27	21354.04	22920.89
TOTAL area/energy			33170.90	33632.47	33305.89	32940.25	35147.40
Number of surface atoms			3128.00	3185.00	3149.00	3154.00	3279.00
Number of buried atoms			2151.00	1981.00	1979.00	1916.00	2145.00
Protein	DR		D1R	D2R	D3R	D4R	D5R
Number	total interface aa		46.00	41.00	42.00	38.00	40.00
Number of amino acids in the interface (%)	Nonpolar Aliphatic	Glycine	0.00	0.00	7.14	7.89	0.00
		Alanine	6.52	7.32	7.14	5.26	7.50
		Valine	2.17	2.44	7.14	7.89	5.00
		Leucine	0.00	7.32	7.14	5.26	2.50
		Isoleucine	6.52	2.44	0.00	2.63	5.00
		Methionine	4.35	4.88	2.38	2.63	5.00
	Polar Uncharged	Serine	13.04	12.2	2.38	5.26	12.50
		Threonine	2.17	12.20	11.90	10.53	2.50
		Cysteine	0.00	0.00	0.00	0.00	0.00
		Proline	4.35	2.44	9.52	7.89	2.50
		Asparagine	2.17	7.32	4.76	0.00	5.00
		Glutamine	6.52	7.32	7.14	2.63	10.00
	Basic Positively Charged	Lysine	13.04	9.76	7.14	5.26	7.50
		Arginine	15.22	12.20	7.14	26.32	17.50
		Histidine	0.00	0.00	4.76	0.00	0.00
	Acidic Negatively Charged	Aspartate	2.17	0.00	0.00	0.00	2.50
		Glutamate	4.34	2.44	4.76	2.63	2.50
	Nonpolar Aromatic	Phenylalanine	13.04	2.44	4.76	5.26	7.50
		Tyrosine	2.17	7.32	4.76	2.63	2.50
		Tryptophan	2.17	0.00	0.00	0.00	2.50
	Aliphatic	Total	19.57	24.39	30.95	31.58	250
	Polar	Total	28.26	41.46	35.71	26.32	32.50

	Basic	Total	28.26	21.95	19.05	31.58	25.00
	Acidic	Total	6.52	2.44	4.76	2.63	5.00
	Aromatic	Total	17.39	9.76	9.52	7.89	12.50
Protein	ARRX		ARR2	ARR2	ARR2	ARR2	ARR2
Number	total interface aa		42.00	37.00	41.00	38.00	43.00
Number of interfacial amino acids (%)	Nonpolar Aliphatic	Glycine	7.14	5.41	9.76	5.26	9.30
		Alanine	9.52	8.11	9.76	7.89	6.98
		Valine	9.52	13.51	9.76	7.89	11.63
		Leucine	14.29	13.51	12.20	15.79	11.63
		Isoleucine	9.52	8.11	7.32	7.89	4.65
		Methionine	0.00	0.00	0.00	0.00	0.00
	Polar Uncharged	Serine	0.00	0.00	0.00	0.00	0.00
		Threonine	4.76	8.11	4.88	5.26	6.98
		Cysteine	7.14	5.41	4.88	5.26	4.65
		Proline	7.14	2.70	7.32	5.26	4.65
		Asparagine	4.76	5.41	4.88	5.26	4.65
		Glutamine	0.00	0.00	2.44	2.63	2.33
	Basic Positively Charged	Lysine	4.76	5.41	4.88	5.26	4.65
		Arginine	4.76	5.41	4.88	5.26	6.98
		Histidine	2.38	0.00	0.00	0.00	2.33
	Acidic Negatively Charged	Aspartate	7.14	8.11	7.32	7.89	6.98
		Glutamate	2.38	2.70	2.44	2.63	2.33
	Nonpolar Aromatic	Phenylalanine	4.76	5.41	4.88	5.26	4.65
		Tyrosine	0.00	2.70	2.44	5.26	4.65
		Tryptophan	0.00	0.00	0.00	0.00	0.00
	Aliphatic	Total	50.00	48.65	48.78	44.74	44.19
	Polar	Total	23.81	21.62	24.39	23.68	23.26
	Basic	Total	11.91	10.81	9.76	10.53	13.95
	Acidic	Total	9.52	10.81	9.76	10.53	9.30
	Aromatic	Total	4.76	8.11	7.32	10.53	9.30
	HB/SB	Count	16.00	9.00	16.00	10.00	17.00
	Consurf	Average	7.21	6.95	6.98	7.21	7.11
	EVFold	Average	2.78	2.71	2.47	2.41	2.44
	ICL1	Intersurf	71.35	171.09	101.26	137.34	71.73

		CocoMaps	182.03	212.33	145.52	231.24	195.14
		SB/HB	0.00	0.00	1.00	0.00	0.00
	ICL2	Intersurf	586.71	505.22	477.70	434.44	750.74
		CocoMaps	613.43	475.58	499.70	383.89	659.98
		SB/HB	5.00	1.00	3.00	3.00	11.00
	ICL3	Intersurf	738.99	595.54	598.68	627.86	403.08
		CocoMaps	893.33	667.67	924.77	665.79	497.02
		SB/HB	10.00	7.00	8.00	7.00	3.00
	HX8	Intersurf	104.04	109.02	157.95	86.23	88.99
		CocoMaps	191.50	129.84	183.47	106.21	193.62
		SB/HB	1.00	0.00	2.00	0.00	1.00

Table 2: D1-5R and ARR3 complexes' interface summary

Complex			D1R-ARR3	D2R-ARR3	D3R-ARR3	D4R-ARR3	D5R-ARR3
POLAR area/energy			11245.21	11385.71	11249.77	11198.09	12111.13
APOLAR area/energy			21772.86	22487.65	nan	22014.85	22943.53
TOTAL area/energy			33018.07	33873.35	nan	33212.95	35054.67
Number of surface atoms			3118.00	3209.00	3116.00	3165.00	3286.00
Number of buried atoms			2099.00	1904.00	1950.00	1843.00	2076.00
Protein	DR		D1R	D2R	D3R	D4R	D5R
Number	total interface aa		41.00	41.00	42.00	36.00	44.00
Number of interfacial amino acids (%)	Nonpolar Aliphatic	Glycine	0.00	0.00	7.14	8.33	0.00
		Alanine	9.76	7.32	9.52	8.33	11.36
		Valine	2.44	2.44	7.14	5.56	4.55
		Leucine	0.00	7.32	7.14	2.78	2.27
		Isoleucine	7.32	2.44	0.00	0.00	4.55
		Methionine	4.88	4.88	2.38	2.78	4.55
	Polar Uncharged	Serine	12.20	9.76	2.38	5.56	11.36
		Threonine	4.88	12.20	14.29	8.33	2.27
		Cysteine	0.00	0.00	0.00	0.00	0.00
		Proline	4.88	2.44	9.52	8.33	2.27
		Asparagine	2.44	7.32	4.76	5.56	4.55
		Glutamine	4.88	4.88	9.52	2.78	9.09
	Basic Positively Charged	Lysine	12.20	12.20	7.14	2.78	6.82
		Arginine	12.20	12.20	4.76	27.78	15.91
		Histidine	0.00	0.00	4.76	0.00	0.00
	Acidic Negatively Charged	Aspartate	2.44	0.00	0.00	0.00	2.27
		Glutamate	4.88	4.88	2.38	2.78	2.27
	Nonpolar Aromatic	Phenylalanine	9.76	2.44	2.38	5.56	11.36
		Tyrosine	2.44	7.32	4.76	2.78	2.27
		Tryptophan	2.44	0.00	0.00	0.00	2.27
	Aliphatic	Total	24.39	24.39	33.33	27.78	27.27
	Polar	Total	29.27	36.59	40.48	30.56	29.55
	Basic	Total	24.39	24.39	16.67	30.56	22.73

	Acidic	Total	7.32	4.88	2.38	2.78	4.55
	Aromatic	Total	14.63	9.76	7.14	8.33	15.91
Protein	ARRX		ARR3	ARR3	ARR3	ARR3	ARR3
Number	Total interface aa		40.00	39.00	41.00	35.00	40.00
Number of amino acids in the interface (%)	Nonpolar Aliphatic	Glycine	7.50	5.13	9.76	5.71	10.00
		Alanine	10.00	10.26	9.76	8.57	7.50
		Valine	7.50	7.69	7.32	8.57	7.50
		Leucine	12.50	12.82	12.20	17.14	12.50
		Isoleucine	5.00	2.56	4.88	2.86	2.50
		Methionine	0.00	0.00	0.00	0.00	0.00
	Polar Uncharged	Serine	7.50	7.69	7.32	8.57	7.50
		Threonine	7.50	7.69	4.88	5.71	7.50
		Cysteine	5.00	5.13	4.88	5.71	5.00
		Proline	7.50	2.56	7.32	5.71	5.00
		Asparagine	0.00	2.56	2.44	0.00	0.00
		Glutamine	0.00	2.56	2.44	2.86	2.50
	Basic Positively Charged	Lysine	7.50	7.69	2.44	2.86	5.00
		Arginine	5.00	5.13	4.88	5.71	5.00
		Histidine	0.00	0.00	0.00	0.00	2.50
	Acidic Negatively Charged	Aspartate	7.50	7.69	7.32	8.57	7.50
		Glutamate	2.50	5.13	4.88	2.86	2.50
	Nonpolar Aromatic	Phenylalanine	5.00	5.13	4.88	5.71	5.00
		Tyrosine	2.50	2.56	2.44	2.86	5.00
		Tryptophan	0.00	0.00	0.00	0.00	0.00
	Aliphatic	Total	42.50	38.46	43.90	42.86	40.00
	Polar	Total	27.50	28.21	29.27	28.57	27.50
	Basic	Total	12.50	12.82	7.32	8.57	12.50
	Acidic	Total	10.00	12.82	12.20	11.43	10.00
	Aromatic	Total	7.50	7.69	7.32	8.57	10.00
	HB/SB	Count	20.00	10.00	16.00	12.00	16.00
	Consurf	Average	7.08	6.74	7.10	7.34	6.78
	EVFold	Average	2.18	2.28	2.38	2.42	2.46
	ICL1	Intersurf	63.28	122.36	162.31	105.73	55.16
		CocoMaps	169.37	180.40	285.18	139.80	166.86

		SB/HB	0.00	1.00	1.00	0.00	1.00
	ICL2	Intersurf	580.58	496.06	545.42	442.08	734.26
		CocoMaps	626.95	496.33	575.53	390.76	703.19
		SB/HB	7.00	3.00	6.00	2.00	7.00
	ICL3	Intersurf	759.74	553.67	600.16	551.16	373.53
		CocoMaps	829.64	689.65	839.10	560.88	424.29
		SB/HB	10.00	6.00	7.00	9.00	4.00
	HX8	Intersurf	95.54	106.02	162.71	87.17	114.55
		CocoMaps	122.31	103.71	169.9	105.42	220.88
		SB/HB	0.00	0.00	1.00	0.00	0.00

Table 3: D1R and G proteins complexes' interface summary

Complex			D1R-Gi1	D1R-Gi2	D1R-Go	D1R-Gslo	D1R-Gssh
POLAR area/energy			12686.10	12534.40	11457.90	14010.30	13592.00
APOLAR area/energy			23975.50	23420.30	23876.90	25353.90	24455.70
TOTAL area/energy			36661.60	35954.70	35334.80	39364.10	38047.70
Number of surface atoms			3249.00	3241.00	3129.00	3521.00	3430.00
Number of buried atoms			2054.00	2068.00	2149.00	2164.00	2157.00
Protein	DR		D1R	D1R	D1R	D1R	D1R
Number	total interface aa		22.00	31.00	28.00	29.00	36.00
Number of amino acids in the interface	Nonpolar Aliphatic	Glycine	4.50	3.20	7.10	0.00	0.00
		Alanine	0.00	6.50	3.60	6.90	5.60
		Valine	0.00	0.00	3.60	3.40	2.80
		Leucine	13.60	9.70	10.70	10.30	8.30
		Isoleucine	9.10	6.50	10.70	3.40	2.80
		Methionine	0.00	0.00	0.00	0.00	0.00
	Polar Uncharged	Serine	0.00	3.20	0.00	0.00	0.00
		Threonine	4.50	3.20	7.10	6.90	8.30
		Cysteine	4.50	3.20	3.60	0.00	0.00
		Proline	0.00	0.00	7.10	6.90	11.10
		Asparagine	4.50	6.50	10.70	0.00	0.00
		Glutamine	4.50	3.20	3.60	13.80	11.10
	Basic Positively Charged	Lysine	13.60	16.10	3.60	3.40	5.60
		Arginine	4.50	6.50	7.10	10.30	11.10
		Histidine	0.00	0.00	0.00	6.90	5.60
	Acidic Negatively Charged	Aspartate	13.60	16.10	3.60	6.90	8.30
		Glutamate	18.20	12.90	10.70	6.90	5.60
	Nonpolar Aromatic	Phenylalanine	4.50	3.20	3.60	3.40	2.80
		Tyrosine	0.00	0.00	3.60	10.30	11.10
		Tryptophan	0.00	0.00	0.00	0.00	0.00
	Aliphatic	Total	27.30	25.80	35.70	24.10	19.40
	Polar	Total	22.20	25.40	41.10	35.60	41.60
	Basic	Total	18.20	22.60	10.70	20.70	22.20
	Acidic	Total	31.80	29.00	14.30	13.80	13.90
	Aromatic	Total	4.50	3.20	7.10	13.80	13.90

Protein	Ga		Gi1	Gi2	Go	Gslo	Gssh
Number	total interface aa		24.00	27.00	28.00	31.00	36.00
Number of amino acids in the interface (%)	Nonpolar Aliphatic	Glycine	0.00	0.00	0.00	0.00	0.00
		Alanine	16.70	22.20	21.40	3.20	19.40
		Valine	4.20	3.70	3.60	6.50	5.60
		Leucine	8.30	3.70	3.60	9.70	2.80
		Isoleucine	8.30	7.40	14.30	3.20	8.30
		Methionine	0.00	0.00	0.00	0.00	0.00
	Polar Uncharged	Serine	0.00	3.70	0.00	0.00	5.60
		Threonine	8.30	7.40	7.10	3.20	8.30
		Cysteine	0.00	0.00	0.00	0.00	0.00
		Proline	8.30	7.40	3.60	6.50	5.60
		Asparagine	4.20	3.70	3.60	0.00	2.80
		Glutamine	4.20	3.70	3.60	12.90	5.60
	Basic Positively Charged	Lysine	8.30	11.10	10.70	6.50	11.10
		Arginine	12.50	11.10	14.30	12.90	8.30
		Histidine	0.00	0.00	0.00	9.70	0.00
	Acidic Negatively Charged	Aspartate	0.00	0.00	0.00	9.70	0.00
		Glutamate	8.30	7.40	7.10	3.20	5.60
	Nonpolar Aromatic	Phenylalanine	8.30	7.40	7.10	3.20	5.60
		Tyrosine	0.00	0.00	0.00	9.70	5.60
		Tryptophan	0.00	0.00	0.00	0.00	0.00
	Aliphatic	Total	37.50	37.00	42.90	22.60	36.10
	Polar	Total	31.00	32.9	22.90	29.60	37.80
	Basic	Total	20.80	22.20	25.00	29.00	19.40
	Acidic	Total	8.30	7.40	7.10	12.90	5.60
	Aromatic	Total	8.30	7.40	7.10	12.90	11.10
	HB/SB	Count	13.00	11.00	17.00	15.00	11.00
	Consurf	Average	6.20	6.50	5.40	6.00	6.30
	EVFold	Average	3.10	3.30	2.80	3.20	3.60
	ICL1	Intersurf	0.00	0.00	0.00	32.20	35.70
		CocoMaps	17.80	0.00	0.00	57.60	57.90
		SB/HB	0.00	0.00	0.00	0.00	0.00
	ICL2	Intersurf	442.40	478.80	505.60	542.10	542.60

		CocoMaps	369.60	431.30	365.50	565.70	501.70
		SB/HB	7.00	7.00	7.00	3.00	1.00
	ICL3	Intersurf	250.90	321.68	460.65	425.52	693.92
		CocoMaps	266.49	351.78	501.10	522.78	727.82
		SB/HB	2.00	2.00	7.00	7.00	7.00
	HX8	Intersurf	17.10	22.3	0.00	36.80	0.00
		CocoMaps	26.8	25.21	0.00	34.20	68.40
		SB/HB	1.00	1.00	0.00	2.00	1.00

Table 4: D2R and G proteins complexes' interface summary

Complex			D2R-Gi1	D2R-Gi2	D2R-Gi3	D2R-Go	D2R-Gz
POLAR area/energy			12857.20	12503.30	12253.90	11407.50	12491.50
APOLAR area/energy			23144.00	22856.50	23261.90	22584.00	23438.80
TOTAL area/energy			36001.20	35359.80	35515.90	33991.40	35930.40
Number of surface atoms			3234.00	3198.00	3192.00	3059.00	3282.00
Number of buried atoms			1956.00	1998.00	2009.00	2106.00	1953.00
Protein		DR	D2R	D2R	D2R	D2R	D2R
Number	total interface aa		23.00	25.00	25.00	27.00	26.00
Number of amino acids in the interface (%)	Nonpolar Aliphatic	Glycine	4.30	8.00	4.00	7.40	3.80
		Alanine	4.30	8.00	0.00	7.40	3.80
		Valine	0.00	0.00	0.00	3.70	0.00
		Leucine	13.00	12.00	12.00	11.10	11.50
		Isoleucine	8.70	8.00	8.00	11.10	11.50
		Methionine	0.00	0.00	0.00	0.00	0.00
	Polar Uncharged	Serine	0.00	0.00	0.00	0.00	3.80
		Threonine	4.30	4.00	4.00	3.70	3.80
		Cysteine	4.30	4.00	4.00	3.70	0.00
		Proline	0.00	0.00	0.00	0.00	0.00
		Asparagine	4.30	4.00	4.00	7.40	3.80
		Glutamine	0.00	0.00	0.00	3.70	7.70
	Basic Positively Charged	Lysine	13.00	12.00	8.00	7.40	0.00
		Arginine	8.70	8.00	12.00	7.40	15.40
		Histidine	0.00	0.00	0.00	0.00	0.00
	Acidic Negatively Charged	Aspartate	13.00	12.00	8.00	3.70	3.80
		Glutamate	17.40	16.00	24.00	14.80	19.20
	Nonpolar Aromatic	Phenylalanine	4.30	4.00	4.00	3.70	3.80
		Tyrosine	0.00	0.00	8.00	3.70	7.70
		Tryptophan	0.00	0.00	0.00	0.00	0.00
	Aliphatic	Total	30.40	36.00	24.00	40.70	30.80
	Polar	Total	16.00	15.00	15.00	23.50	24.20
	Basic	Total	21.70	20.00	20.00	14.80	15.40
	Acidic	Total	30.40	28.00	32.00	18.50	23.10

	Aromatic	Total	4.30	4.00	12.00	7.40	11.50
Protein	Ga		Gi1	Gi2	Gi3	Go	Gz
Number	total interface aa		25.00	23.00	23.00	27.00	28.00
Number of amino acids in the interface (%)	Nonpolar Aliphatic	Glycine	0.00	0.00	0.00	0.00	0.00
		Alanine	4.00	4.30	0.00	7.40	7.10
		Valine	8.00	4.30	8.70	11.10	7.10
		Leucine	8.00	8.70	4.30	7.40	7.10
		Isoleucine	8.00	4.30	8.70	3.70	7.10
		Methionine	8.00	8.70	8.70	11.10	7.10
	Polar Uncharged	Serine	4.00	4.30	4.30	3.70	7.10
		Threonine	4.00	4.30	4.30	3.70	7.10
		Cysteine	0.00	0.00	0.00	0.00	0.00
		Proline	4.00	4.30	4.30	3.70	3.60
		Asparagine	8.00	8.70	8.70	3.70	7.10
		Glutamine	0.00	0.00	0.00	0.00	0.00
	Basic Positively Charged	Lysine	12.00	13.00	13.00	11.10	10.70
		Arginine	24.00	26.10	26.10	25.90	21.40
		Histidine	0.00	0.00	0.00	0.00	0.00
	Acidic Negatively Charged	Aspartate	0.00	0.00	0.00	0.00	0.00
		Glutamate	0.00	0.00	0.00	0.00	0.00
	Nonpolar Aromatic	Phenylalanine	4.00	4.30	4.30	3.70	3.60
		Tyrosine	4.00	4.30	4.30	3.70	3.60
		Tryptophan	0.00	0.00	0.00	0.00	0.00
	Aliphatic	Total	36.00	30.40	30.40	40.70	35.70
	Polar	Total	25.00	26.70	26.70	18.80	32.00
	Basic	Total	36.00	39.10	39.10	37.00	32.10
	Acidic	Total	0.00	0.00	0.00	0.00	0.00
	Aromatic	Total	8.00	8.70	8.70	7.40	7.10
HB/SB		Count	20.00	17.00	19.00	25.00	16.00
Consurf		Average	6.00	5.50	5.60	5.40	5.40
EVFold		Average	3.10	3.30	3.20	2.90	3.20
	ICL1	Intersurf	0.00	0.00	0.00	0.00	0.00
		CocoMaps	19.30	0.00	0.00	0.00	37.10
		SB/HB	0.00	0.00	0.00	0.00	0.00

	ICL2	Intersurf	431.92	397.40	516.60	539.20	436.00
		CocoMaps	370.80	360.80	429.00	476.60	409.30
		SB/HB	7.00	6.00	8.00	4.00	4.00
	ICL3	Intersurf	332.03	263.93	314.95	409.14	330.54
		CocoMaps	275.12	230.26	256.40	383.59	328.28
		SB/HB	11.00	6.00	7.00	15.00	12.00
	HX8	Intersurf	46.70	42.60	52.40	0.00	46.20
		CocoMaps	29.10	29.20	42.90	3.80	30.30
		SB/HB	0.00	1.00	1.00	0.00	0.00

Table 5: D3R and G proteins complexes' interface summary

Complex			D3R-Gi1	D3R-Gi2	D3R-Gi3	D3R-Gslo	D3R-Gssh	D3R-Gz	D3R-Gq
POLAR area/energy			12739.40	12329.50	12380.30	14234.40	13903.30	12639.00	13285.10
APOLAR area/energy			23761.30	23419.50	23685.20	25037.50	24333.20	23977.00	24725.50
TOTAL area/energy			36500.70	35749.10	36065.50	39271.90	38236.50	36616.00	38010.50
Number of surface atoms			3186.00	3186.00	3193.00	3410.00	3350.00	3288.00	3299.00
Number of buried atoms			1966.00	1972.00	1970.00	2124.00	2086.00	1909.00	1986.00
Protein		DR	D3R	D3R	D3R	D3R	D3R	D3R	D3R
Number	total interface aa		25.00	23.00	23.00	26.00.00	26.00	24.00	25.00
Number of amino acids in the interface (%)	Nonpolar Aliphatic	Glycine	4.00	4.30	4.30	7.70	0.00	4.20	0.00
		Alanine	0.00	4.30	0.00	3.80	3.80	4.20	0.00
		Valine	0.00	0.00	0.00	3.80	3.80	0.00	12.00
		Leucine	12.00	13.00	13.00	7.70	7.70	12.50	12.00
		Isoleucine	12.00	8.70	13.00	3.80	3.80	12.50	8.00
		Methionine	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Polar Uncharged	Serine	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		Threonine	4.00	4.30	4.30	3.80	7.70	4.20	0.00
		Cysteine	4.00	4.30	4.30	0.00	3.80	0.00	0.00
		Proline	0.00	0.00	0.00	3.80	3.80	0.00	4.00
		Asparagine	4.00	4.30	0.00	0.00	0.00	4.20	8.00
		Glutamine	0.00	0.00	0.00	15.40	15.40	8.30	8.00
	Basic Positively Charged	Lysine	12.00	13.00	8.70	0.00	7.70	0.00	12.00
		Arginine	8.00	8.70	8.70	11.50	7.70	16.70	16.00
		Histidine	0.00	0.00	0.00	7.70	7.70	0.00	0.00

	Acidic Negatively Charged	Aspartate	16.00	13.00	8.70	7.70	3.80	4.20	4.00
		Glutamate	16.00	17.40	21.70	7.70	7.70	16.70	4.00
	Nonpolar Aromatic	Phenylalanine	4.00	4.30	4.30	3.80	3.80	4.20	4.00
		Tyrosine	4.00	0.00	8.70	11.50	11.50	8.30	8.00
		Tryptophan	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Aliphatic	Total	28.00	30.40	30.40	26.90	19.20	33.30	32.00
	Polar	Total	15.00	16.00	10.70	29.10	38.80	20.70	25.00
	Basic	Total	20.00	21.70	17.40	19.20	23.10	16.70	28.00
	Acidic	Total	32.00	30.40	30.40	15.40	11.50	20.80	8.00
	Aromatic	Total	8.00	4.30	13.00	15.40	15.40	12.50	12.00
Protein	Ga		Gi1	Gi2	Gi3	Gslo	Gssh	Gz	Gq
Number	total interface aa		23.00	24.00	23.00	28.00	27.00	27.00	22.00
Number of amino acids in the interface (%)	Nonpolar Aliphatic	Glycine	4.30	4.20	4.30	0.00	3.70	3.70	0.00
		Alanine	8.70	4.20	8.70	7.10	11.10	3.70	9.10
		Valine	8.70	8.30	8.70	14.30	7.40	14.80	9.10
		Leucine	13.00	4.20	8.70	14.30	11.10	7.40	13.60
		Isoleucine	4.30	8.30	4.30	3.60	0.00	7.40	0.00
		Methionine	8.70	4.20	4.30	7.10	7.40	7.40	4.50
	Polar Uncharged	Serine	4.30	4.20	4.30	0.00	3.70	3.70	4.50
		Threonine	4.30	4.20	4.30	7.10	11.10	3.70	4.50
		Cysteine	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		Proline	4.30	4.20	4.30	3.60	3.70	3.70	4.50
		Asparagine	0.00	4.20	4.30	3.60	3.70	0.00	4.50
		Glutamine	4.30	8.30	8.70	7.10	7.40	7.40	9.10
	Basic Positively Charged	Lysine	8.70	4.20	4.30	3.60	3.70	7.40	9.10
		Arginine	13.00	25.00	17.40	14.30	14.80	18.50	18.20
		Histidine	8.70	8.30	8.70	3.60	3.70	7.40	4.50

	Acidic Negatively Charged	Aspartate	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		Glutamate	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Nonpolar Aromatic	Phenylalanine	4.30	4.20	4.30	3.60	3.70	3.70	4.50
		Tyrosine	0.00	0.00	0.00	7.10	3.70	0.00	0.00
		Tryptophan	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Aliphatic	Total	47.80	33.30	39.10	46.40	40.70	44.40	36.40
	Polar	Total	21.40	31.00	32.10	27.40	37.60	23.50	33.30
	Basic	Total	30.40	37.50	30.40	21.40	22.20	33.30	31.80
	Acidic	Total	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Aromatic	Total	4.30	4.20	4.30	10.70	7.40	3.70	4.50
HB/SB		Count	12.00	14.00	12.00	10.00	5.00	11.00	11.00
Consurf		Average	5.80	5.80	5.80	6.30	7.50	5.70	5.00
EVFold		Average	3.00	3.00	3.00	3.00	3.40	3.10	3.50
	ICL1	Intersurf	0.00	0.00	0.00	16.8	39.2	0.00	0.00
		CocoMaps	0.00	0.00	0.00	15.1	27.4	0.00	0.00
		SB/HB	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	ICL2	Intersurf	427.70	415.90	421.60	413.80	393.80	399.50	330.10
		CocoMaps	392.10	448.20	393.40	441.90	498.90	438.40	303.50
		SB/HB	1.00	5.00	3.00	0.00	1.00	1.00	2.00
	ICL3	Intersurf	428.77	398.69	374.81	482.98	510.56	403.43	392.97
		CocoMaps	431.17	354.74	358.27	421.01	427.47	398.86	393.54
		SB/HB	11.00	8.00	8.00	7.00	4.00	10.00	8.00
	HX8	Intersurf	26.80	24.00	30.40	16.40	0.00	35.80	0.00
		CocoMaps	19.30	23.50	28.90	88.00	5.80	23.20	5.70
		SB/HB	0.00	1.00	1.00	0.00	0.00	0.00	0.00

Table 6: D4-5R and G proteins complexes' interface summary

Complex			D4R-Gob	D4R-Gt2	D4R-Gz	D5R-Gslo	D5R-Gssh	D5R-Gz
POLAR area/energy			12640.70	12141.20	12231.00	14280.20	14284.70	11748.10
APOLAR area/energy			23696.90	23065.20	23581.00	26284.80	25549.00	24276.30
TOTAL area/energy			36337.60	35206.40	35812.00	40564.90	39833.60	36024.30
Number of surface atoms			3205.00	3097.00	3248.00	3657.00	3548.00	3264.00
Number of buried atoms			1871.00	1978.00	1891.00	2173.00	2184.00	2229.00
Protein		DR	D4R	D4R	D4R	D5R	D5R	D5R
Number	total interface aa		27.00	31.00	33.00	30.00	32.00	28.00
Number of amino acids in the interface (%)	Nonpolar Aliphatic	Glycine	7.40	3.20	9.00	3.30	0.00	3.60
		Alanine	3.70	3.20	12.00	3.30	3.10	3.60
		Valine	0.00	0.00	6.00	3.30	6.30	0.00
		Leucine	11.10	12.9	9.00	10.00	9.40	10.70
		Isoleucine	11.10	9.70	3.00	3.30	3.10	14.30
		Methionine	0.00	3.20	0.00	0.00	0.00	0.00
	Polar Uncharged	Serine	7.40	3.20	3.00	0.00	0.00	3.60
		Threonine	7.40	3.20	6.00	3.30	3.10	3.60
		Cysteine	3.70	3.20	0.00	0.00	3.10	0.00
		Proline	3.70	0.00	6.00	6.70	6.30	0.00
		Asparagine	7.40	9.70	6.00	0.00	0.00	3.60
		Glutamine	0.00	3.20	3.00	13.30	12.50	10.70
	Basic Positively Charged	Lysine	7.40	12.90	3.00	6.70	6.30	3.60
		Arginine	0.00	0.00	21.00	13.30	12.50	10.70
		Histidine	0.00	0.00	3.00	6.70	6.30	3.60
	Acidic Negatively Charged	Aspartate	11.10	12.90	0.00	10.00	9.40	3.60
		Glutamate	11.10	9.70	3.00	3.30	6.30	14.30

	Nonpolar Aromatic	Phenylalanine	3.70	3.20	3.00	3.30	3.10	3.60
		Tyrosine	3.70	6.50	0.00	10.00	9.40	7.10
		Tryptophan	0.00	0.00	3.00	0.00	0.00	0.00
	Aliphatic	Total	33.30	32.30	39.00	23.30	21.90	32.10
	Polar	Total	37.60	29.60	24.00	30.30	33.00	27.40
	Basic	Total	7.40	12.90	27.00	26.70	25.00	17.90
	Acidic	Total	22.20	22.60	3.00	13.30	15.60	17.90
	Aromatic	Total	7.40	9.70	6.00	13.30	12.50	10.70
Protein	Ga		Gob	Gt2	Gz	Gslo	Gssh	Gz
Number	total interface aa		24.00	32.00	31.00	31.00	34.00	30.00
Number of amino acids in the interface (%)	Nonpolar Aliphatic	Glycine	4.20	3.10	3.00	0.00	0.00	0.00
		Alanine	8.30	15.60	3.00	12.90	17.60	13.30
		Valine	8.30	6.30	0.00	3.20	2.90	3.30
		Leucine	12.50	12.50	10.00	3.20	2.90	3.30
		Isoleucine	0.00	0.00	13.00	6.50	8.80	13.30
		Methionine	0.00	0.00	0.00	0.00	2.90	3.30
	Polar Uncharged	Serine	4.20	0.00	3.00	3.20	2.90	0.00
		Threonine	0.00	6.30	6.00	9.70	8.80	6.70
		Cysteine	0.00	0.00	0.00	0.00	0.00	0.00
		Proline	4.20	3.10	0.00	3.20	2.90	3.30
		Asparagine	4.20	6.30	6.00	3.20	5.90	3.30
		Glutamine	8.30	3.10	13.00	6.50	5.90	6.70
	Basic Positively Charged	Lysine	4.20	3.10	0.00	12.90	11.80	13.30
		Arginine	25.00	21.90	13.00	19.40	14.70	16.70
		Histidine	4.20	3.10	0.00	0.00	0.00	0.00
	Acidic Negatively Charged	Aspartate	0.00	0.00	6.00	3.20	0.00	3.30
		Glutamate	4.20	6.30	13.00	3.20	2.90	3.30
	Nonpolar Aromatic	Phenylalanine	4.20	3.10	3.00	6.50	5.90	6.70
		Tyrosine	0.00	3.10	6.00	3.20	2.90	0.00

		Tryptophan	4.20	3.10	0.00	0.00	0.00	0.00
	Aliphatic	Total	33.30	37.50	29.00	25.80	35.30	36.70
	Polar	Total	25.80	24.80	29.00	33.80	35.50	26.00
	Basic	Total	33.30	28.10	13.00	32.30	26.50	30.00
	Acidic	Total	4.20	6.30	19.00	6.50	2.90	6.70
	Aromatic	Total	8.30	9.40	10.00	9.70	8.80	6.70
HB/SB		Count	18.00	24.00	24.00	16.00	17.00	19.00
Consurf		Average	5.60	5.40	5.10	6.50	7.20	5.90
EVFold		Average	2.80	2.80	3.40	3.40	3.40	3.20
	ICL1	Intersurf	0.00	0.00	11.43	0.00	33.90	0.00
		CocoMaps	0.00	0.00	11.43	52.80	88.10	0.00
		SB/HB	0.00	0.00	0.00	0.00	0.00	0.00
	ICL2	Intersurf	358.10	392.80	602.20	510.00	523.80	457.30
		CocoMaps	300.30	322.80	288.80	417.70	514.00	483.80
		SB/HB	5.00	8.00	5.00	3.00	2.00	7.00
	ICL3	Intersurf	473.79	530.83	602.21	533.34	511.53	540.40
		CocoMaps	456.61	474.22	532.10	456.31	453.13	452.17
		SB/HB	12.00	13.00	17.00	10.00	10.00	8.00
	HX8	Intersurf	0.00	50.20	34.20	113.40	92.40	83.70
		CocoMaps	8.50	57.70	44.30	167.00	100.70	132.90
		SB/HB	0.00	0.00	0.00	0.00	1.00	0.00

Table 7: Best training results attained by cluster for the “Fold” dataset

pre-processing	metrics	algorithms				
		Cluster I	Cluster II	Cluster III	Cluster IV	Cluster V
PCA		bagEarth	lda	glmboost	ORFlog	svmLinear
	AUROC	0.75	0.76	0.76	0.76	0.78
	Accuracy	0.92	0.84	0.82	1.00	0.82
	Sensitivity	0.78	0.61	0.41	1.00	0.41
	Specificity	0.97	0.93	0.97	1.00	0.98
	PPV	0.91	0.76	0.84	1.00	0.87
	NPV	0.92	0.86	0.81	1.00	0.82
	FDR	0.09	0.24	0.16	0.00	0.13
	F1-score	0.84	0.67	0.55	1.00	0.56
PCA_Up		bagEarth	qda	Glm	ORFlog	svmLinear
	AUROC	0.93	0.89	0.84	0.97	0.84
	Accuracy	0.94	0.89	0.89	1.00	0.86
	Sensitivity	0.96	0.99	0.93	1.00	0.93
	Specificity	0.91	0.79	0.86	1.00	0.79
	PPV	0.92	0.82	0.87	1.00	0.81
	NPV	0.96	0.99	0.92	1.00	0.92
	FDR	0.08	0.18	0.13	0.00	0.19
	F1-score	0.94	0.90	0.90	1.00	0.87
PCA_Down		bagEarth	lda	avNNet	ORFpls	svmPoly
	AUROC	0.70	0.71	0.68	0.73	0.72
	Accuracy	0.98	0.80	0.92	1.00	0.84
	Sensitivity	0.98	0.82	0.92	1.00	0.82
	Specificity	0.98	0.78	0.92	1.00	0.86
	PPV	0.98	0.79	0.92	1.00	0.86
	NPV	0.98	0.82	0.92	1.00	0.83
	FDR	0.02	0.21	0.08	0.00	0.14
	F1-score	0.98	0.81	0.92	1.00	0.84
Scaled		bagEarth	knn	glmboost	ORFsvm	svmPoly

	AUROC	0.77	0.66	0.73	0.76	0.74
	Accuracy	0.96	0.79	0.79	0.97	0.88
	Sensitivity	0.88	0.47	0.33	0.94	0.59
	Specificity	0.99	0.90	0.96	0.99	0.99
	PPV	0.96	0.65	0.74	0.96	0.94
	NPV	0.96	0.82	0.79	0.98	0.86
	FDR	0.04	0.35	0.26	0.04	0.06
	F1-score	0.92	0.55	0.46	0.95	0.72
Scaled_Up		bagEarth	multinom	avNNet	ORFsvm	svmLinear
	AUROC	0.94	0.83	0.83	0.97	0.86
	Accuracy	0.96	1.00	0.96	1.00	0.97
	Sensitivity	0.99	1.00	0.95	1.00	0.99
	Specificity	0.93	0.99	0.96	1.00	0.95
	PPV	0.94	0.99	0.96	1.00	0.95
	NPV	0.98	1.00	0.95	1.00	0.99
	FDR	0.06	0.01	0.04	0.00	0.05
	F1-score	0.96	1.00	0.96	1.00	0.97
Scaled_Down		bagEarth	knn	avNNet	ORFpls	svmRadial
	AUROC	0.72	0.67	0.68	0.72	0.69
	Accuracy	1.00	0.79	0.96	1.00	0.80
	Sensitivity	1.00	0.82	0.96	1.00	0.84
	Specificity	1.00	0.76	0.96	1.00	0.76
	PPV	1.00	0.78	0.96	1.00	0.78
	NPV	1.00	0.81	0.96	1.00	0.83
	FDR	0.00	0.22	0.04	0.00	0.22
	F1-score	1.00	0.80	0.96	1.00	0.81

Table 8: Best training results attained by cluster for the “Allrows” dataset

pre-processing	metrics	algorithms				
		Cluster I	Cluster II	Cluster III	Cluster IV	Cluster V
PCA		bagEarth	lda	gamboost	ORFlog	svmRadial
	AUROC	0.81	0.77	0.78	0.81	0.80
	Accuracy	0.87	0.82	0.81	0.99	0.83
	Sensitivity	0.63	0.56	0.400	0.98	0.49
	Specificity	0.96	0.91	0.96	1.00	0.96
	PPV	0.85	0.700	0.80	1.00	0.80
	NPV	0.88	0.85	0.81	0.99	0.83
	FDR	0.15	0.30	0.20	0.00	0.20
	F1-score	0.73	0.63	0.53	0.99	0.60
PCA_Up		C5.0Tree	knn	ctree	ORFlog	svmRadial
	AUROC	0.88	0.84	0.87	0.98	0.84
	Accuracy	0.99	0.83	0.88	1.00	0.81
	Sensitivity	1.00	0.92	0.90	1.00	0.85
	Specificity	0.97	0.74	0.86	1.00	0.77
	PPV	0.98	0.78	0.87	1.00	0.79
	NPV	1.00	0.90	0.90	1.00	0.83
	FDR	0.02	0.22	0.13	0.00	0.21
	F1-score	0.99	0.84	0.88	1.00	0.82
PCA_Down		ada	lda	avNNet	ORFpls	svmRadialCost
	AUROC	0.80	0.79	0.78	0.81	0.79
	Accuracy	0.92	0.83	0.89	0.99	0.83
	Sensitivity	0.91	0.800	0.90	0.99	0.81
	Specificity	0.93	0.85	0.88	0.98	0.84
	PPV	0.93	0.84	0.88	0.98	0.84
	NPV	0.91	0.81	0.90	0.99	0.82
	FDR	0.07	0.16	0.12	0.02	0.16
	F1-score	0.92	0.82	0.89	0.99	0.82
Scaled		bagEarth	lda	glmboost	ORFsvm	svmRadial

	AUROC	0.82	0.73	0.80	0.82	0.80
	Accuracy	0.88	0.90	0.84	0.99	0.82
	Sensitivity	0.69	0.77	0.55	0.97	0.46
	Specificity	0.95	0.95	0.95	1.00	0.95
	PPV	0.82	0.86	0.79	0.99	0.77
	NPV	0.89	0.92	0.85	0.99	0.83
	FDR	0.18	0.14	0.21	0.01	0.23
	F1-score	0.75	0.81	0.65	0.98	0.57
Scaled_Up		bagEarth	lda	avNNet	ORFsvm	svmLinear
	AUROC	0.91	0.87	0.87	0.97	0.89
	Accuracy	0.91	0.92	0.95	1.00	0.94
	Sensitivity	0.93	0.96	0.96	1.00	0.97
	Specificity	0.88	0.89	0.95	1.00	0.92
	PPV	0.89	0.9	0.95	1.00	0.92
	NPV	0.93	0.95	0.96	1.00	0.97
	FDR	0.11	0.1	0.05	0.00	0.08
	F1-score	0.91	0.93	0.95	1.00	0.95
Scaled_Down		bagEarth	knn	glmboost	ORFsvm	svmRadial
	AUROC	0.79	0.76	0.79	0.81	0.77
	Accuracy	0.94	0.80	0.85	0.99	0.82
	Sensitivity	0.94	0.77	0.87	0.99	0.81
	Specificity	0.93	0.83	0.83	0.99	0.82
	PPV	0.93	0.82	0.84	0.99	0.82
	NPV	0.94	0.79	0.87	0.99	0.81
	FDR	0.07	0.18	0.16	0.01	0.18
	F1-score	0.94	0.80	0.85	0.99	0.82

Table 9: Best training results attained by cluster for the “Complex” dataset

Pre-processing	Metrics	Algorithms				
		Cluster I	Cluster II	Cluster III	Cluster IV	Cluster V
PCA		bagEarth	knn	glmboost	ORFpls	svmRadialCost
	AUROC	0.82	0.79	0.77	0.84	0.87
	Accuracy	0.99	0.86	0.81	0.98	0.90
	Sensitivity	0.96	0.64	0.24	0.92	0.64
	Specificity	1.00	0.93	0.98	1.00	0.98
	PPV	1.00	0.73	0.75	1.00	0.89
	NPV	0.99	0.900	0.82	0.98	0.9
	FDR	0.00	0.27	0.25	0.00	0.11
	F1-score	0.98	0.68	0.36	0.96	0.74
PCA_Up		bagEarth	qda	avNNet	ORFlog	svmRadial
	AUROC	0.98	0.94	0.91	0.98	0.94
	Accuracy	0.99	0.98	0.99	1.00	0.91
	Sensitivity	1.00	1.00	1.00	1.00	0.94
	Specificity	0.98	0.95	0.98	1.00	0.88
	PPV	0.98	0.96	0.98	1.00	0.89
	NPV	1.00	1.00	1.00	1.00	0.94
	FDR	0.02	0.04	0.02	0.00	0.11
	F1-score	0.99	0.98	0.99	1.00	0.92
PCA_Down		ada	rda	avNNet	ORFpls	svmRadialCost
	AUROC	0.87	0.84	0.83	0.88	0.87
	Accuracy	0.86	0.90	1.00	1.00	0.90
	Sensitivity	0.88	0.96	1.00	1.00	0.96
	Specificity	0.84	0.84	1.00	1.00	0.84
	PPV	0.85	0.86	1.00	1.00	0.86
	NPV	0.88	0.95	1.00	1.00	0.95
	FDR	0.15	0.14	0.00	0.00	0.14
	F1-score	0.86	0.91	1.00	1.00	0.91
Scaled		bagEarth	knn	ctree	ORFsvm	svmRadialCost

	AUROC	0.81	0.77	0.8	0.84	0.85
	Accuracy	1.00	0.83	0.88	0.99	0.86
	Sensitivity	1.00	0.56	0.76	0.96	0.60
	Specificity	1.00	0.91	0.92	1.00	0.94
	PPV	1.00	0.64	0.73	1.00	0.75
	NPV	1.00	0.88	0.93	0.99	0.89
	FDR	0.00	0.36	0.27	0.00	0.25
	F1-score	1.00	0.60	0.75	0.98	0.67
Scaled_Up		bagEarth	knn	ctree	ORFsvm	svmRadial
	AUROC	0.98	0.91	0.93	0.97	0.94
	Accuracy	1.00	0.89	0.92	0.99	0.92
	Sensitivity	1.00	0.95	0.94	1.00	0.98
	Specificity	1.00	0.83	0.91	0.99	0.87
	PPV	1.00	0.85	0.91	0.99	0.88
	NPV	1.00	0.95	0.94	1.00	0.97
	FDR	0.00	0.15	0.09	0.01	0.12
	F1-score	1.00	0.90	0.93	0.99	0.93
Scaled_Down		bagEarth	knn	glmboost	ORFpls	svmRadial
	AUROC	0.82	0.78	0.83	0.87	0.86
	Accuracy	1.00	0.68	0.92	1.00	0.90
	Sensitivity	1.00	0.72	0.96	1.00	0.96
	Specificity	1.00	0.64	0.88	1.00	0.84
	PPV	1.00	0.67	0.89	1.00	0.86
	NPV	1.00	0.70	0.96	1.00	0.95
	FDR	0.00	0.33	0.11	0.00	0.14
	F1-score	1.00	0.69	0.92	1.00	0.91

Table 10: Best test results attained by cluster for the “Fold” dataset

pre-processing	metrics	algorithms				
		Cluster I	Cluster II	Cluster III	Cluster IV	Cluster V
PCA		bagEarth	lda	glmboost	ORFlog	svmLinear
	AUROC	0.86	0.81	0.72	0.89	0.81
	Accuracy	0.86	0.83	0.76	0.87	0.78
	Sensitivity	0.57	0.52	0.19	0.57	0.24
	Specificity	0.96	0.95	0.96	0.98	0.98
	PPV	0.86	0.79	0.67	0.92	0.83
	NPV	0.86	0.84	0.76	0.86	0.78
	FDR	0.14	0.21	0.33	0.08	0.17
	F1-score	0.69	0.63	0.30	0.71	0.37
PCA_Up		bagEarth	qda	glm	ORFlog	svmLinear
	AUROC	0.80	0.67	0.68	0.82	0.72
	Accuracy	0.83	0.73	0.74	0.82	0.77
	Sensitivity	0.57	0.57	0.62	0.43	0.71
	Specificity	0.93	0.79	0.79	0.96	0.79
	PPV	0.75	0.50	0.52	0.82	0.56
	NPV	0.85	0.83	0.85	0.82	0.88
	FDR	0.25	0.50	0.48	0.18	0.44
	F1-score	0.65	0.53	0.57	0.56	0.62
PCA_Down		bagEarth	lda	avNNet	ORFpls	svmPoly
	AUROC	0.71	0.73	0.74	0.77	0.77
	Accuracy	0.76	0.78	0.79	0.82	0.82
	Sensitivity	0.71	0.67	0.71	0.67	0.62
	Specificity	0.77	0.82	0.82	0.88	0.89
	PPV	0.54	0.58	0.60	0.67	0.68
	NPV	0.88	0.87	0.89	0.88	0.86
	FDR	0.46	0.42	0.40	0.33	0.32
	F1-score	0.61	0.62	0.65	0.67	0.65
Scaled		bagEarth	knn	glmboost	ORFsvm	svmPoly

	AUROC	0.77	0.64	0.72	0.83	0.82
	Accuracy	0.82	0.73	0.77	0.85	0.79
	Sensitivity	0.62	0.29	0.29	0.57	0.29
	Specificity	0.89	0.89	0.95	0.95	0.98
	PPV	0.68	0.50	0.67	0.80	0.86
	NPV	0.86	0.77	0.78	0.86	0.79
	FDR	0.32	0.50	0.33	0.20	0.14
	F1-score	0.65	0.36	0.40	0.67	0.43
Scaled_Up		bagEarth	multinom	avNNet	ORFsvm	svmLinear
	AUROC	0.71	0.68	0.71	0.90	0.70
	Accuracy	0.77	0.74	0.77	0.88	0.76
	Sensitivity	0.62	0.57	0.62	0.62	0.62
	Specificity	0.82	0.81	0.82	0.98	0.81
	PPV	0.57	0.52	0.57	0.93	0.54
	NPV	0.85	0.84	0.85	0.88	0.85
	FDR	0.43	0.48	0.43	0.07	0.46
	F1-score	0.59	0.55	0.59	0.74	0.58
Scaled_Down		bagEarth	knn	avNNet	ORFpls	svmRadial
	AUROC	0.79	0.69	0.76	0.76	0.70
	Accuracy	0.83	0.73	0.81	0.81	0.76
	Sensitivity	0.86	0.71	0.76	0.67	0.67
	Specificity	0.82	0.74	0.82	0.86	0.79
	PPV	0.64	0.50	0.62	0.64	0.54
	NPV	0.94	0.88	0.90	0.88	0.87
	FDR	0.36	0.50	0.38	0.36	0.46
	F1-score	0.73	0.59	0.68	0.65	0.60

Table 11: Best test results attained by cluster for the “Allrows” dataset

pre-processing	metrics	algorithms				
		Cluster I	Cluster II	Cluster III	Cluster IV	Cluster V
PCA		bagEarth	lda	gamboost	ORFlog	svmRadial
	AUROC	0.75	0.71	0.70	0.73	0.68
	Accuracy	0.80	0.77	0.76	0.77	0.75
	Sensitivity	0.45	0.38	0.26	0.29	0.31
	Specificity	0.92	0.91	0.94	0.95	0.91
	PPV	0.68	0.62	0.61	0.67	0.57
	NPV	0.82	0.80	0.78	0.79	0.79
	FDR	0.32	0.38	0.39	0.33	0.43
	F1-score	0.54	0.47	0.37	0.40	0.40
PCA_Up		C5.0Tree	knn	ctree	ORFlog	svmRadial
	AUROC	0.63	0.67	0.61	0.70	0.61
	Accuracy	0.70	0.68	0.69	0.77	0.66
	Sensitivity	0.52	0.79	0.48	0.36	0.60
	Specificity	0.77	0.64	0.76	0.91	0.68
	PPV	0.45	0.44	0.42	0.60	0.40
	NPV	0.82	0.89	0.80	0.80	0.82
	FDR	0.55	0.56	0.58	0.40	0.60
	F1-score	0.48	0.56	0.44	0.45	0.48
PCA_Down		ada	lda	avNNet	ORFpls	svmRadialCost
	AUROC	0.63	0.61	0.61	0.63	0.60
	Accuracy	0.65	0.64	0.64	0.67	0.63
	Sensitivity	0.71	0.64	0.67	0.62	0.62
	Specificity	0.62	0.64	0.62	0.69	0.63
	PPV	0.41	0.39	0.39	0.42	0.38
	NPV	0.86	0.83	0.84	0.84	0.82
	FDR	0.59	0.61	0.61	0.58	0.62
	F1-score	0.52	0.49	0.49	0.50	0.47
Scaled		bagEarth	lda	glmboost	ORFsvm	svmRadial

	AUROC	0.76	0.69	0.72	0.72	0.63
	Accuracy	0.81	0.76	0.77	0.78	0.73
	Sensitivity	0.45	0.5	0.33	0.38	0.29
	Specificity	0.93	0.85	0.93	0.92	0.89
	PPV	0.70	0.55	0.64	0.64	0.48
	NPV	0.83	0.83	0.80	0.81	0.78
	FDR	0.30	0.45	0.36	0.36	0.52
	F1-score	0.55	0.52	0.44	0.48	0.36
Scaled_Up		bagEarth	lda	avNNet	ORFsvm	svmLinear
	AUROC	0.76	0.65	0.72	0.72	0.72
	Accuracy	0.81	0.71	0.78	0.78	0.77
	Sensitivity	0.76	0.60	0.60	0.43	0.71
	Specificity	0.83	0.75	0.85	0.91	0.79
	PPV	0.62	0.46	0.58	0.62	0.55
	NPV	0.91	0.84	0.85	0.82	0.88
	FDR	0.38	0.54	0.42	0.38	0.45
	F1-score	0.68	0.52	0.59	0.51	0.62
Scaled_Down		bagEarth	knn	glmboost	ORFsvm	svmRadial
	AUROC	0.68	0.64	0.63	0.64	0.62
	Accuracy	0.70	0.65	0.65	0.64	0.63
	Sensitivity	0.79	0.74	0.69	0.76	0.69
	Specificity	0.67	0.62	0.64	0.59	0.61
	PPV	0.46	0.41	0.41	0.40	0.39
	NPV	0.90	0.87	0.85	0.87	0.85
	FDR	0.54	0.59	0.59	0.60	0.61
	F1-score	0.58	0.53	0.51	0.52	0.50

Table 12: Best test results attained by cluster for the “Complex” dataset

Pre-processing	Metrics	Algorithms				
		Cluster I	Cluster II	Cluster III	Cluster IV	Cluster V
PCA		bagEarth	knn	glmboost	ORFpls	svmRadialCost
	AUROC	0.68	0.71	0.74	0.71	0.71
	Accuracy	0.78	0.8	0.8	0.80	0.80
	Sensitivity	0.50	0.40	0.20	0.50	0.40
	Specificity	0.86	0.92	0.97	0.89	0.92
	PPV	0.50	0.57	0.67	0.56	0.57
	NPV	0.86	0.85	0.81	0.86	0.85
	FDR	0.50	0.43	0.33	0.44	0.43
	F1-score	0.50	0.47	0.31	0.53	0.47
PCA_Up		bagEarth	qda	avNNet	ORFlog	svmRadial
	AUROC	0.66	0.62	0.69	0.71	0.69
	Accuracy	0.76	0.74	0.78	0.80	0.78
	Sensitivity	0.50	0.40	0.60	0.50	0.60
	Specificity	0.83	0.83	0.83	0.89	0.83
	PPV	0.45	0.40	0.50	0.56	0.50
	NPV	0.86	0.83	0.88	0.86	0.88
	FDR	0.55	0.60	0.50	0.44	0.50
	F1-score	0.48	0.40	0.55	0.53	0.55
PCA_Down		ada	rda	avNNet	ORFpls	svmRadialCost
	AUROC	0.64	0.68	0.72	0.72	0.68
	Accuracy	0.70	0.76	0.80	0.80	0.76
	Sensitivity	0.70	0.70	0.70	0.70	0.70
	Specificity	0.69	0.78	0.83	0.83	0.78
	PPV	0.39	0.47	0.54	0.54	0.47
	NPV	0.89	0.90	0.91	0.91	0.90
	FDR	0.61	0.53	0.46	0.46	0.53
	F1-score	0.50	0.56	0.61	0.61	0.56
Scaled		bagEarth	knn	ctree	ORFsvm	svmRadialCost

	AUROC	0.71	0.71	0.74	0.68	0.71
	Accuracy	0.80	0.80	0.83	0.78	0.80
	Sensitivity	0.50	0.40	0.60	0.50	0.40
	Specificity	0.89	0.92	0.89	0.86	0.92
	PPV	0.56	0.57	0.60	0.50	0.57
	NPV	0.86	0.85	0.89	0.86	0.85
	FDR	0.44	0.43	0.40	0.50	0.43
	F1-score	0.53	0.47	0.60	0.50	0.47
Scaled_Up		bagEarth	knn	ctree	ORFsvm	svmRadial
	AUROC	0.71	0.70	0.74	0.66	0.66
	Accuracy	0.80	0.78	0.83	0.76	0.76
	Sensitivity	0.50	0.70	0.60	0.50	0.50
	Specificity	0.89	0.81	0.89	0.83	0.83
	PPV	0.56	0.50	0.60	0.45	0.45
	NPV	0.86	0.91	0.89	0.86	0.86
	FDR	0.44	0.50	0.40	0.55	0.55
	F1-score	0.53	0.58	0.60	0.48	0.48
Scaled_Down		bagEarth	knn	glmboost	ORFpls	svmRadial
	AUROC	0.69	0.55	0.67	0.67	0.67
	Accuracy	0.78	0.61	0.74	0.76	0.74
	Sensitivity	0.60	0.50	0.70	0.60	0.70
	Specificity	0.83	0.64	0.75	0.81	0.75
	PPV	0.50	0.28	0.44	0.46	0.44
	NPV	0.88	0.82	0.90	0.88	0.90
	FDR	0.50	0.72	0.56	0.54	0.56
	F1-score	0.55	0.36	0.54	0.52	0.54

Table 13: Best test results attained by cluster for the "Fold*" dataset (after removing CoeViz and InterEV features)

pre-processing	metrics	algorithms				
		Cluster I	Cluster II	Cluster III	Cluster IV	Cluster V
Scaled		rf	ORFpls	glmboost	svmRadialCost	stepLDA
	AUROC	0.82	0.90	0.92	0.88	0.89
	Accuracy	0.81	0.81	0.85	0.77	0.78
	Sensitivity	0.26	0.21	0.37	0.05	0.11
	Specificity	0.98	1.00	1.00	1.00	1.00
	PPV	0.83	1.00	1.00	1.00	1.00
	NPV	0.81	0.80	0.83	0.77	0.78
	FDR	0.17	0.00	0.00	0.00	0.00
	F1-score	0.74	0.79	0.63	0.95	0.89

8. Papers that resulted from 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) [14]* - *Equal contribution*	**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.
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) [13]*	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 ⍰D structure and sequence. The SpotOn web interface is freely available at: http://milou.science.uu.nl/services/SPOTON/.

*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)*	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.
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 Dopamine Receptor Family, 2017.	(In preparation)

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

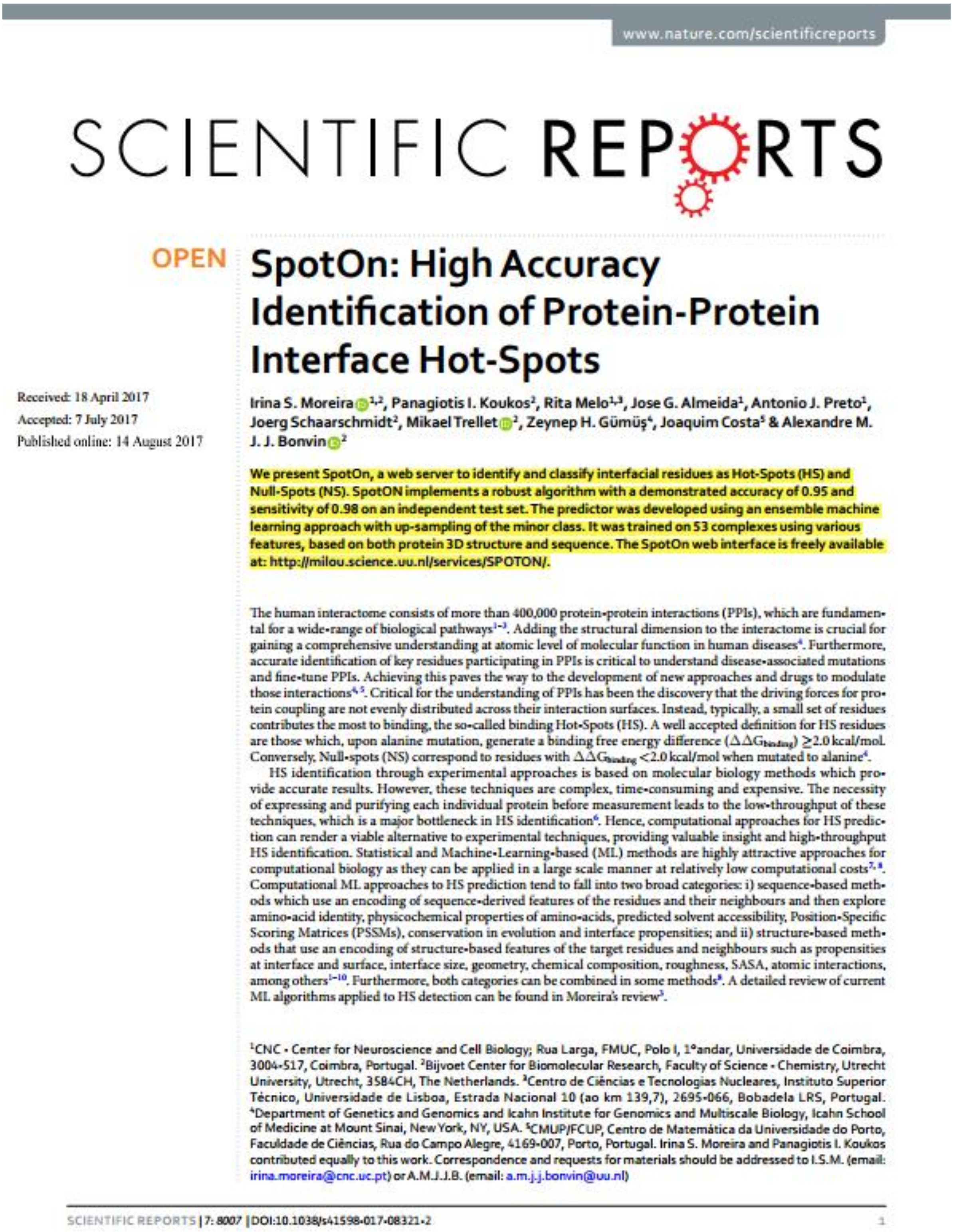
Agostinho Lemos[a], Rita Melo[b,c], Antonio J. Preto[b], Jose G. Almeida[b], Irina S. Moreira[*b,d], M. Natália D. S. Cordeiro*[a]

[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[3].

[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)