UNIVERSITÀ DEGLI STUDI DI MILANO

FACOLTÀ DI SCIENZE E TECNOLOGIE

Ph.D. course in Industrial Chemistry,
Chemistry Department

Computational studies of Protein-Protein and Protein-Antibody interactions

Implication for Molecular Design

Supervisors:
Prof. Laura Belvisi
Prof. Giorgio Colombo
Coordinator:
Prof. Dominique Roberto

Ph.D. Candidate:
Filippo Marchetti
Cycle XXXIII

Anno Accademico 2019-2020

Contents

Motivation

Proteins are the workhorses of cells: they carry out the most disparate tasks, from structural organization of cytoskeleton, to active and passive transport, to the catalysis of chemical reactions and the relay of information. Proteins do not work in isolation but in finely regulated networks, where they interact with other partners in a delicate interplay of formation and disruption of multicomponent assemblies.

The genomic revolution, the advancements in experimental analytical and structural approaches, and the development of proteomics and interactomics have made available a wealth of data to analyse. Chemistry occupies a center-stage position in this scenario, as chemical principles in chemistry underlie the determinants of interactions while the development of small molecule interactors of biomolecules may help understand the wiring of specific biological pathways. In parallel, the unprecedented increase in computer power in the last few years has had a dramatic impact in our capacity to investigate biological systems at the highest possible level of resolution. In this context, high performance computing has opened the possibility to investigate complex systems by simulating their dynamics and study of equilibrium and non-equilibrium settings in realistic settings. Molecular Dynamics (MD)simulations have emerged as one of the privileged methods to disentangle the intricacies of biochemical systems. For each atom of the system one can solve its Newton equations of motion, obtaining a trajectory in the phase space for the entire system, and study its behavior in equilibrium and non-equilibrium conditions. The constant rise in computational power gave the possibility to scientists to study larger and larger systems, while the advances in experimental techniques enhanced the possibility for direct comparisons between wet and *in silico* data at similar levels of resolution. Yet, despite the validity of Moore's Law (i.e., the exponential growth of the computing power due to transistors miniaturization), the timescale of the events that can be simulated has an upper limit of the millisecond with tailor-made computers which is not enough to study some biologically relevant phenomena, such as protein-protein interactions of the effects of allosteric ligands on the actual functionally-oriented motions of proteins. Starting from these considerations, in this thesis, I have set out to develop and validate novel methods to predict the location of potentially interacting surfaces on proteins and to predict the impact of small molecules on the activation vs. the inhibition of proteins' functional dynamic states. To this end, I have combined physico-chemical approaches to the study of protein dynamics and generate novel approaches that may overcome the current limitations of pure brute force MD simulations.

The work reported in this thesis is split in two parts, the first part is dedicated to the study of methods for the prediction of protein-protein binding while in the second part is adressed the study of the consequences of protein binding. In Chapter 1, I will give a general introduction to the problem of protein-protein interactions and to the approaches we have developed in this context. I will also introduce Machine Learning methods and Artificial Intelligence concepts, which we have applied to extract from normal MD simulations descriptors able to indicate whether a certain ligand will result in the activation or inhibition of protein functional states. In Chapter 2 there is a comparison of distinct approaches for the prediction of residues involved in protein interfaces. In Chapter 3 there is an application of prediction methods for the detection of epitopes in SARS-CoV-2 spike protein. In Chapter 4 there are presented two analysis of allosteric systems. One is the implementation of a learning classificator inside a docking protocol, in order to discriminate the activity of a ligand. The second is the study of allosteric signal in the dynamics of integrin αvβ6.

Chapter 1

Introduction

1.1 A general view of Protein-Protein Interactions

Interactions among proteins are essential for the maintenance of cellular activity and the modulation of protein complexes formation has an impact on signal transduction, correctness of folding, antibody recognition and many other biological processes. Therefore the comprehension of the physico-chemical traits underlying the formation of those complexes and their interactions will be crucial for the rational design of new drugs. It is important to underline that the same general physicochemical principles underpin the binding mechanisms of different types of biomolecules, from proteins to nucleic acids. Therefore, it is possible to classify the interactions without explicitly indicating the chemical nature of the ligands using different criteria. A possible classification criterion may reside in the size of the ligand, that is the distinction between small molecule or macromolecule binding, or a classification between “inactive” or “active” molecule, whereby the interacting particles can be considered as rigid or dynamics bodies, respectively. Further choices could be stability of the protein as a monomer and/or the dynamic profiles of the proteins and the ligands. Once the generality of the principles is pointed out, one can focus on general cases with a protein-protein and protein-peptide interactions and the models used to describe these interactions. Different general conceptual models have been developed to describe protein interactions, which I will briefly revise in the following.

Lock and Key

The first model for interactions in proteins (which covers both the cases of protein-protein and protein-ligand interactions) is “ Lock and Key ” (LK) mechanism proposed by Fisher in 1894[91]. This theory considers the interaction of proteins only on the basis of surface complementarity of the binding partners, that are supposed to remain rigidly preorganized in their unbound structure conformation upon the binding event.

Induced Fit

The idea of binding proteins viewed as rigid blocks, central in the “ Lock and Key ” model, was called into question, arising the hypothesis of alteration of the protein conformations upon binding. This hypothesis was corroborated by the observation of the experimental crystal structures of bound vs. unbound molecules, which highlighted changes in the backbone of the structures [23]. On this basis, the Induced Fit (IF) model was first proposed by Koshland, explaining protein binding as a series of conformational changes that are triggered progressively through molecular association[110]. It may be expressed as:

$$\mathrm{R} + \mathrm{L} \rightleftharpoons \mathrm{RL} \rightleftharpoons (\mathrm{RL})^* \tag{1.1}$$

where the receptor (R) and the ligand (L) first form the complex RL and subsequently a conformational change leading to the final, active state (RL)*[117].

A notable example of the characterization of induced fit in protein binding is represented by a study by Stella et al., who built a model using quantitative data[185]. The authors conducted time-resolved fluorescence analysis in order to study the kinetics of GST P1-1 dynamics. Fluorescence data results indicated that apo-GST has flexible regions and the protein adopts at least two families of conformations. With this data they could prove the relevance of structural fluctuation for the binding model: the process is divided in two steps. In the absence of substrate the region of the binding site has frequent (with timing faster than milliseconds) fluctuations between the different structures, the flexibility helps the partner to binds forming a weakly bound intermediate complex. Subsequently a much slower transition stabilize the complex in a final conformation. In this case the fluctuations in the unbound conditions help the substrate to reach an optimal docking pose.

Conformational Selection

Improvement in spectroscopy and nuclear magnetic resonance technique extend the ability of scientists to probe conformations of folded protein in states alternative to the native, ground state. Such conformations may be populated as a consequence of the thermal activation of specific dynamic modes. Structural investigations permitted to observe that these higher-energy conformers are compatible with the conformational variations observed in ligand binding events or in catalytic steps in enzymes[123][208]. Therefore, is not necessary to have binding-induced structural rearrangements and protein binding can be rationalized with the Conformational Selection (CS) model. The CS model was first proposed by Monod et al.[140] describing protein allostery and later adjusted for protein interactions by Kumar et al.[117]. Starting experimental observations were made by Zavodszky[218] and successively the theory strengthen when Frauenfelder published his paper on the energy landscape of protein[93]. This model depicts the unbound state of a protein as an ensemble of conformations, in which the eventual bound conformation pre-exists the binding event. During the binding event, the ligand selects its partner among the conformers in the ensemble, the final structures of the complexes depending on their binding energy. The conformational selection process may be formalized as follows:

$$R^{i}, R^{ii}, R^{iii}, ... + \mathrm{L} \rightleftharpoons (\mathrm{R}^{\mathrm{ii}}\mathrm{L})^{*} \tag{1.2}$$

where $R^{i}, R^{ii}\ldots$ are different conformers of protein receptor R; ligand L selects and binds in the activated complex $(R^{ii}L)^{*}$ with one of the conformers (R^{ii} in this case) without intermediate steps. The processes taken in account, conformational changes and binding/unbinding paths, need thermal activation and the crossing of a free-energy barrier. Usually processes of this kind spend less time in the transition (time required for barrier-crossing) with respect to the states between barrier-crossing events (dwell times)[208]. That is observed in single-molecule experiments where transition times between conformations of a protein happen below experimental resolution, and the conformational changes are detected as leaps between conformers[176].

1.1.1 Landscape and population shift in protein binding

The wider conceptual generalization for protein-protein and protein-ligand binding is represented by energy landscape model. This was initially applied to study folding mechanism with the introduction of folding funnels and subsequently it was extended to provide a general theory for protein binding[132]. Given that proteins display a moltitude of conformations in solution they could be defined in terms of statistical distributions. In particular the number of conformers depends on the flexibility of the molecule: for example, the more rigid is a protein, the smaller is the ensemble of structures. Moreover, the populations of conformers do not follow a plain distribution but some conformers have higher statistical weight, while for others the weight might be low. According to the funnel concept for folding, multiple conformations run downhill along the funnel by multiple routes and in proximity of the funnel bottom there are a range of different conformers[27][93][124]. The extent of the ruggedness of the bottom, defined by the depths of the wells and the height of the barriers, affects the size of the ensemble of conformations[132]. Similar characteristics are found in binding processes. Therefore, it is possible to extend the core idea of folding funnels into binding funnels. A binding funnel has a rugged bottom, like the folding one, that defines a conformational ensemble of protein complexes. In this case the ruggedness establishes in what manner the structures in the complex are allowed to bind, and what structures of the complexes can be populated[117].

This model could be extended to the so-called Dynamic energy landscape considering the hypothesis that the landscape of a funnel is not static but can change. The essence of the extension lies in the addition of environmental factors[197]. The environment can change due to merely physical factors, such as temperature, pressure, pH or ionic concentration, or, in alternative, the factor could be a binding partner. Studies by Sabelko et al. showed the possibility of environmental influence, observing a non exponential kinetics in the folding of some proteins caused by a modulation of the free energy surface following a raise in temperature from 2 to 88 C°[170]. The landscape is thus determined by specific environmental conditions; alterations of those conditions modify the wells' steepness, the bumps and barrier heights of the funnel. For example, if the funnel was to become gradually sharper and sheerer, this would imply that external factors make the protein more rigid. Dynamic changes in the landscape imply that the also the kinetic of transition between conformers are modified and, consequently, the size of the populations associated to the conformers. A population shift can be observed: structures that had small statistical weight become more represented in the ensemble and vice versa.

1.1.2 The impact of different binding models

For a better comprehension of the binding mechanism it is interesting to evaluate the impact of the different binding models on the studies of protein interactions. Indeed, some studies tried to characterize the association of flexible proteins and the impact of IF and CS[47]. Stein et al. carried out a systematic study developing a metric system in order to assess the frequency of rigid binding and the energetic cost of a conformational change[184]. The authors use both RMSD (root-mean-square deviation) differences and descriptors derived from normal mode analysis to find that most of the analysed proteins do not display notable structural rearrangement upon association, suggesting that LK model may represent a wide range of binding events. In cases with significant conformational changes, many kinetic signatures for IF and CS are found although for some proteins the metric cannot discriminate between the models. In addition, it also shown that structural changes in the backbone do not pay an high energetic price. The authors suggest that all three modality can coexist and that for certain families or domains one of the model prevails over the others. The process of structural changes is often coupled to ligand binding and that coupling could be complicated, making it difficult to distinguish between an IF or a CS mode. Progresses were made using data obtained in relaxation experiments, like single-molecule FRET (Förster resonance energy transfer) or NMR (nuclear magnetic resonance), where binding/unbinding occurrence and conformational changes were decoupled showing a time order[208]. In order to better represent this dual behaviour, Csermely et al. proposed an extension of the original conformational selection model to include both lock-and-key and induced fit, describing the binding as a mutually conditional step-wise recognition and encounter process[39]. In this context, a local version of CS models could also be developed, where some protein segments, such as hinge areas or zones with separate motion, are considered more significant than others in the binding event.

1.2 Computational methods

In this section I introduce two common computational methods that have been used to study the problem of recognition between two proteins or between a ligand and a protein, that are Docking and Molecular Dynamics. Next, I will give an overview of the methods used and developed in this thesis to tackle those problems.

1.2.1 Docking

Protein-ligand and protein-protein docking have been representing the common practice in the prediction of relative partner orientations in bound protein complexes. At its basics, molecular docking is based on modelling the interactions between compounds including electrostatic Coulomb-like potentials, van der Waals forces and hydrogen bonds. In order to quantify the goodness of the binding pose all the interaction terms are condensed in a docking score and various methods have been developed over the years for the docking score evaluation and the generation of the poses. In the example of elementary rigid-body systems the ligand position is searched by scanning a six-dimensional space, three dimension for rotations plus three for translations, for a correct fitting in the binding site[4]. This trivial example was improved introducing a molecular mechanics energy functions and information about the location of binding residues. During the last decades many different docking algorithms have been released for both academic and commercial use. In the case of protein-protein binding, the docking approach can be seen as a “ ab initio ” method for the generation of viable hypotheses on the relative placement of interaction partners and for the definition of potential complex structures. In general, most of those algorithms employ optimal shape complementarity as a core ratio in interactions prediction[180]. The most popular methods use fast Fourier transform correlation [118][111], that helps in the individuation of the regions with an optimized match between complementary protein surfaces, or geometric hashing methods for a fast comparison of geometric surface descriptors[135]. Another option in the generation of complexes is the simulation of protein-protein encounters with molecular dynamics (MD), Monte Carlo, Brownian motion or multi-start energy minimization[217]. Those approaches add to the binding partners a form of structural flexibility but are definitely slower than FFT or hashing methods. Conformational changes are also included with the addition in the energy minimization step of collective normal mode directions[216]. Finally, if the proteins are docked starting from completely unbound partners, a wide ensemble of candidates is created, and the selection of the best solutions among all the candidates could be subjected to the low accuracy of scoring functions. The process could be improved adding experimental data about the binding regions or residues known to be in contact in the complexes obtained, for example, with site-directed mutagenesis, cross-linking experiments, NMR studies, FRET characterizations etc. Some software explicitly incorporates this information to guide the search process in the definition of the most likely bound conformations. The most used examples include HADDOCK [48], ClusPro[214], PyDock[152] or ZDOCK[80].

1.2.2 Molecular Dynamics

Molecular Dynamics is a method for the simulation of a many-body system evolving under the laws of classical mechanics. The aim of such simulations is the sampling of physical plausible states for the computation of properties at equilibrium. Basically, a Molecular Dynamics task consists in the integration of Newton's equation of motion (see Formula 1.3), collecting a spatial trajectory of the atoms of the system.

$$\frac{d^2x_i}{dt^2} = \frac{F_x}{m_i} \tag{1.3}$$

meaning that a particle i of mass m_i moves subjected to a force F_x. The force can be derived from the potential function $V(\bar{r})$ that describe the essential interactions of the system:

$$F_i = -\frac{\partial V}{\partial r_i} \tag{1.4}$$

The potential function comprises the interactions between all the particles and can be viewed as a potential energy term in function of atomic positions $\bar{r} = r_1, r_2, ..., r_N$. A common expression V for a all-atom protein simulation is:

$$\begin{aligned} V(\bar{r}) = &\sum_{bond} \frac{1}{2}K_b(b-b_0)^2 + \sum_{angles} \frac{1}{2}K_\theta(\theta-\theta_0)^2 + \sum_{impdihed} \frac{1}{2}K_\phi(\phi-\phi_0)^2 \\ &+ \sum_{dihedral} K_\psi(1+cos(n\psi-\delta) + \sum_{(i,j)}(\frac{C_{12}(i,j)}{r_{ij}^{12}} - \frac{C_6(i,j)}{r_{ij}^6} + \frac{q_iq_j}{4\pi\epsilon_0\epsilon_r r_{ij}}) \end{aligned} \tag{1.5}$$

All bonded, angular and dihedrals are modelled with harmonic terms, while for long range interactions both Lennard-Jones and Coulombian potentials are used. The parameters are obtained from experimental data and by exploiting high quality *ab initio* calculations, then collected in a force field. During the years several types of force field were produced, and the choice of the appropriate set of parameters is restricted to the context in which it has been developed for. Different sets of force fields and integration algorithms were developed because the simulation of molecular systems with a huge variety of atoms and interactions, constraint and boundary conditions requires a refined treatment with sophisticated software. Commonly used software packages for the simulation of molecular systems were provided by both academia and development houses, such as: AMBER, NAMD, GROMACS, CHARMM and CHARMm[30][147][2][26].

1.3 How to predict the interaction regions on proteins and study the consequences of binding

In this thesis, the main questions are the prediction of the potential interaction sites on a protein, as well as the evaluation of the possible consequences of binding. Here, I will briefly review the main methodological approaches I applied and further developed to tackle these two fascinating problems.

1.3.1 Energy decomposition and MLCE method

As stated in the initial paragraph of this chapter, structure, dynamics and binding are strictly intertwined. In this context, the investigation of the determinants of protein structural 3D organization can aptly inform on the location of regions endowed with functional, partner recognition properties. To start putting this subject in perspective, the analysis of the impact of mutations on protein stability shows differences from highly susceptible sites to zone where the effect on stability is low[139], suggesting that the stabilization energy of a protein has an unbalanced distribution among the residues. Extensions of this context reach out to the rationalization of enzyme catalysis: the concept of Discrete Breathers, i.e. zones where the protein can store the energy received by excitation[172][157], has permitted to build a theoretical frameowork of biological behaviours such as the ability of enzymes to redirect the energy obtained after substrate docking to carry out chemical reactions[86]. It has computed that discrete breathers can release from 20 to 65% of the reserve of energy during structural rearrangements[39][158]. In this framework, our group developed a method to extrapolate the contribution of residues to the stability of a protein starting from the crystal structures of the native state. Overall, the method, called the Energy Decomposition Method (EDM) analyses the organization of pair-interaction energy[193]. It consists in the extrapolation of the main components of the matrix of interactions of all residues pairs, obtained on a single reference structure or averaged over an MD trajectory. For the energetic interactions only nonbonded terms are considered including Coulomb electrostatic and van der Waals potentials with the addition of a solvation related term using a MM-GBSA or MM-PBSA approach. The pair interaction term, for the couple of amino acids i and j, is thus:

$$E_{ij}^{nb} = E_{ij}^{elect} + E_{ij}^{VdW} + G_{ij}^{solv} \tag{1.6}$$

Those terms form a NxN interaction matrix, for a protein with N residues, which can be represented in terms of eigenvectors:

$$M_{ij} = \sum_{k=1}^{N} \lambda_k \omega_i^k \omega_j^k \tag{1.7}$$

where the eigenvalue is λ^k and ω_i^k are the components of the corresponding eigenvector. The total non-bonded energy can thus be defined as:

$$E^{nb} = \sum_{ij} M_{ij} \tag{1.8}$$

The eigenvectors are reordered in the summation according to the associated eigenvalue so that the first is the lowest (most negative), in this case $\lambda_1 \bar{\omega}^1 \bar{\omega}^{1\intercal}$ is the term with the highest contribution to the total energy. In the case of globular proteins with a well-defined structure, it was verified that the interaction matrix can be approximated by the matrix $\dot{M}$ [193][143]:

$$M_{ij} \approx \dot{M}_{ij} = \lambda_1 \omega_i^1 \omega_j^1 \tag{1.9}$$

reducing noise due to tenuous interactions in the identification of residues with a relevant contribution to stability. Therefore the total energy becomes:

$$E^{nb} = \sum_{ij} \lambda_1 \omega_i^1 \omega_j^1 \tag{1.10}$$

The physical meaning is that every couple of residues (i, j) interacts with an energy approximated to $\lambda_1 \omega_i^1 \omega_j^1$. λ_1 is a coupling parameter that defines the energy associated to the eigenvector $\bar{\omega}^1$. The components of $\bar{\omega}^1$ represent the contribution of each single residue to the global stabilization energy of the protein.

The energy decomposition was further extended for the study of multi-domain proteins[95] and interactive regions prediction[173][155]. In the latter case, the extended method is called MLCE (Matrix of the Local Coupling Energy). The rationale behind MLCE is that binding regions appear to have little role in participation in stabilization of the 3D structure, as they should be prone to support conformational changes and establish new interactions with a new partner at a minimal energetic cost. In the case of protein antigen-antibody interaction, moreover, interaction regions, a.k.a. epitopes, should be able to undergo mutations without impacting on the stability of the functional folded state[169]. For these reasons, our algorithm searches for localized networks with low intensity coupling with the remainder of the protein in the simplified interaction matrix. To this end, it is necessary to add topological information to the energetic one in order to unveil localized region with minimal internal coupling, which define potential interaction patches. That is obtained filtering the approximated interaction matrix with a contact map built with the scheme:

$$C_{ij} = \begin{cases} 1 & \text{if } r_{ij} \leq 6.5 \\ 0 & \text{if } r_{ij} > 6.5 \end{cases} \tag{1.11}$$

Where r_{ij} is the distance from residue i and j and the cut-off is 6.5 Å. The final Matrix of the Local Coupling Energy is then computed with:

$$L_{ij} = C_{ij} \cdot \dot{M}_{ij} \tag{1.12}$$

In this matrix the interactions between residues L_{ij} is then filtered to extract only the coupling with the weakest interactions. Such regions represent putative interaction areas (or epitopes).

1.3.2 Evolutionary Trace

Evolutionary studies have highlighted that proteins may preserve their folds even in the face of low sequence conservation. Indeed, it was aptly demonstrated that a group of homologous proteins with sequence similarity as low as 20% could display a maximum RMSD variation of 2.4 Å[35]. At the same time, residues in active sites tend to have a lower mutation rate, meaning a push towards activity preservation from evolution[224]. Those observations suggested scientist to look at the most conserved residues for the prediction of functionally relevant regions. One requirement for this task is the definition of a figure of merit to report on the actual conservation of a residue in a family of proteins. For the measurement of amino acid conservation, it is possible to make use of Shannon's entropy[178], defined as:

$$S(x) = \sum_a p(x, a) ln(p(x, a)) \tag{1.13}$$

where for a residue x the value $p(x,a)$ is the probability to have amino acid a in position x. The probability to find a certain amino acid in a position could be approximated with occurrence frequency, which can be computed from Multiple Sequence Alignments (MSA) of different proteins homologous to the one under study: in this picture every column of a MSA corresponds to the amino acids that are detected for a certain position in the protein fold and the occurrence frequency of every amino acid is obtainable by simple counting. The highly conserved sites are the ones with a lower entropy: in fact the lowest value is observed with a single amino acid at a certain position (no variability, occurrence frequency of 1). The value then raises while the variance is increased. Entropy gives a significant evaluation of the variability but it does not contemplate how the amino acids are distributed with respect to the similarity of the sequences and if the divergence is driven by evolution. For example, we can consider two residues (positions in the alignment) with the same frequency for an amino acid. In one case, the presence of the amino acid is shared only among sequences with high similarity while in the other case the amino acid is widespread regardless of the degree of similarity among sequences. In the first case, it is fair to hypothesize that there is an evolutionary pressure for the preservation of that particular amino acid, at least in a subgroup of protein, whereas in the second the occurrence of the same amino acid can be considered accidental. In the first case, the amino acid conservation may be a reverberation of a functional need.

In order to overcome this problem, the group of Dr Lichtarge (at Baylor College of Medicine) developed a technique that combines the computation of Shannon's entropy with phylogenetic information[138]. The evolutionary trace method consists in the building a phylogenetic tree from the sequences in the MSA using an UPGMA algorithm, then evaluate the entropy in every subgroup defined with increasing distance cutoff from leaf to the root. This approach permits to control if the frequency detected for a certain amino acid is due to a conservation maintained in the subgroups or is just an accidental occurrence. The general expression of the real value evolutionary trace is:

$$EVT(x) = 1 + \sum_{n=1}^{N} w_{node}(n) \sum_{g} w_{group}(g) S(g,x) \tag{1.14}$$

where $S(g,x) = -\sum_a f(g,x,a) ln(f(g,x,a))$ is the entropy of the single subgroup g, $w_{node}(n)$ is the weight assigned to the node n and $w_{group}(g)$ is a further weight assigned to the subgroup g. The weights can be used if it is necessary to emphasize more some groups than others. One example may be the need to favour sequence similar to the protein in study: in this case, it is possible to set a group weight dependent on the distance from the reference protein. For an unbiased scoring all the sequences in the alignment could be considered equally and the weights become: $w_{node} = 1/N$ and $w_{group} = 1$. The MLCE and evolutionary trace algorithms will constitute the core approaches to our studies of protein interactions, entailing the general case of protein-protein complex formation and the specific case of the prediction of regions where interactions with antibodies can be established.

1.4 Small molecule modulation of dynamic states

One part of this thesis will also be dedicated to predict the effects of ligand binding (in particular small molecules) in inducing specific dynamic states in proteins. In the general framework of the conformational selection model, modifying the dynamic states of a protein will result in a perturbation of its interaction surfaces and a modulation of its interaction spectra. The question is then whether we are able to predict the onset of specific states of a protein, activated or inhibited for instance, from the study of its interactions with small molecule ligands and internal dynamics.

Learning algorithms

The field of artificial intelligence is made of algorithms that do not require a specific programming for the task which they are assigned to but rather they aim learn how to complete a certain job through a process of trials and errors. For a formal point of view we can refer to the definition: " A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P , improves with experience E. "(see `http://www.deeplearningbook.org/`)[109] Usually the aim of learning a feature is to identify the variables that describe experimental data. Sometimes, however, those variables are hidden in the dataset or even not measurable. In these cases, machine learning methods can have the power to recognize defined patterns in the data. Indeed, they are used in a wide range of tasks that span from face recognition to text mining. In the structural biology field, the interest in machine learning is justified by the improvement in techniques and procedures that have increased the quantity of biological data that is machine readable, helping in the development of new approaches of artificial intelligence for biomedical studies and prospecting an impact in clinical pharmacology[221]. In particular, the field of machine learning is adequate for systems characterized by a large dimensionality and that is the case of chemical and physical spaces. Therefore their role in drug discovery is being actively investigated. In fact, it is possible to find applications of those tools in virtual screening [115][66], in molecular dynamics for the generation of efficient collective variables for metadynamcs[187] or for the prediction of protein activity[116]. Machine Learning tools can be divided in two main groups: unsupervised and supervised learning. Unsupervised learning means that the feature to learn is not known in the data while in supervised learning the data are labelled with the feature that has to be learned. In our case, only supervised methodologies will be discussed. The base approach for the dataset management is to split the database in two parts: one will be dedicated to the effective training and the other is needed for testing the trained model. The insertion of a testing step of the method is important because the training implies only the optimization of the system on the available dataset but the model should be able to repeat the task on new data, otherwise it will be useless. Overfitting occurs when a trained system has good performance on the training test but fails in prediction of new data.

Several supervised methods were developed. Here, we will use three common models that are: logistic regression, support vector machine and random forest.

Logistic Regression

The logistic regression, or logit regression, is a statistical method used to classify objects. Basically the method consist in evaluating the probability of an instance to be associated to the condition used for the classification, the single probability is computed with a logistic function. The logistic function is a sigmoid curve defined on the real space with values ranging between 0 and 1, formally expressed as:

$$S(t) = \frac{1}{(1 - e^{-t})} \tag{1.15}$$

The function t is assumed to be linear and in the case of a single variable model become $t = b_0 + b_1 x$. When more variables are considered t is the generalized with $t = \bar{b}^{\intercal}\bar{x} = b_0 + \sum_{i=1}^{N} b_i x_i$, for a system of N variables $\bar{b}$ is the vector of coefficients and $\bar{x}$ is the vector of variables. The learning is reached computing the error obtained on a train dataset, the error represents the divergence of the function outcome from the real value of the data. The process is repeated with an optimization algorithm to find the best $\bar{b}$ parameters that minimize the error.

In order to reduce overfitting a constraint is added to the error function (E) in the minimization and the total function to be minimized become:

$$min_{b,C}\left(||\bar{b}||^2 + C \cdot E(y(b), y_{obs})\right) \tag{1.16}$$

where $y(b)$ is the predicted outcome and y_{obs} is the expected outcome. The minimization runs between the modulation of the factor C, that gives the impact of the error function, and the constraint $||\bar{b}||^2$ that represent the magnitude of the parameter vector. The constraint works as a penalty giving a disadvantage when the parameter vector becomes too large and this is necessary for overfitting reduction because if the parameters are limited the system has less option for adapt to the training cases, in other terms the model variance is diminished meaning that difference in the input variables results in low difference in the predictions. Therefore, it helps in balancing the model between training accuracy and extrapolation ability.

Support Vector Machine

Support Vector Machines (SVM) are a class of supervised learning methodologies suitable for classification. Essentially, this algorithm separate instances in a dataset representing each data as a point in a N-dimensional space, i.e. each data is defined by N variables, building an $(N - 1)$ dimensional hyperplane. The optimal plane is found reducing the error according to data labels and maximizing the margin, the minimum distance from the plane and the separated groups, an higher margin involves a more robust model. A hyperplane is individuated by the equation $\bar{w}^{\intercal}\bar{x}–b = 0$, where the $\bar{w}$ vector is the normal to the plane and the value $\frac{b}{|\bar{w}|}$ resemble how much the plane is shifted from the origin. Therefore the components of the vector $\bar{w}$ and the value b are the parameters we are interested for.

The optimal parameters are obtained through a minimization of the loss function defined as:

$$\frac{1}{N}\sum_{i=1}^{N} max[0, 1 - y_i(\bar{w}^\intercal \bar{x}_i - b)] + a||\bar{w}||^2 \tag{1.17}$$

The term $y_i(\bar{w}^\intercal \bar{x}_i – b) > 1$ implies that the point i lie in the correct side of the separation plane without falling in the margin, otherwise the loss function will be penalized. Just like the Logistic Regression model there is a pay-off between the correctness of the classification and the magnitude of the parameters vector here modulated by the value of a.

Random Forest

It is an algorithm for classification consisting in building an ensemble of decision trees during training and predicting the classes via consensus with the idea that a forest will unify the efforts increasing the performance respect to a tree. A decision tree is a set of decision rules organized with an hierarchical tree and the training is necessary to find the optimum order of rules that let a correct classification of the data. In practice when an instance with N variables has to be classified a tree is generated with a descending processes from the root to the leaves, at every node the following step is decided with a classification rule on one of the variables. The optimal tree is built starting from the root and for every node is chosen a decision rule on the variable that gives the lowest error in classification. There are several ways to compute the error and to define the decision rules.

A shortcoming of Decision Trees is the tendency to overfitting to the training data[99] and Random Forest can obviate this defect. With Random Forest a group of different trees is generated, each tree is determined with a stochastic reduction of the set of features of the dataset in order to reduce the correlation among the trees. Once an ensemble of N tree is produced the instance x is assigned to the class that gives the best results in the forest, an example of the assignment procedure can be built considering the set of trees T, $v_j(x)$ the leaf node where x is putted in T_j (j = 1,2,. . . ,N) and c (c=1,2,. . . ,C) a certain class. The probability that x is associated to c is:

$$P(c|v_j(x)) = \frac{P(c, v_j(x))}{\sum_l^C P(c_l, v_j(x))} \tag{1.18}$$

The instance x is then associated to the class c that maximize the value of:

$$g_c(x) = \sum_{j=1}^{N} P(c|v_j(x)) \tag{1.19}$$

with this discriminant function g is possible to maintain the training accuracy while the results are averaged[105].

1.4.1 Goodness of prediction assessment

Once the models to use are set is important to measure if the system is working properly and giving significant predictions. Firstly, the results can be summarised counting the number of true positive (TP), true negative (TN), false positive (FP) and false negative (FN). It is difficult to make an esteem of the predictor's ability just considering those numbers alone, therefore is common practise to use a metric. During the years various metrics were proposed, each one with his peculiarity. The first metrics to use for obtaining general information about the predictor's behaviour are *accuracy* (ACC) and *error* (ERR), ACC is the sum of correct cases respect the whole number of cases while the ERR is the complementary to one:

$$ACC = \frac{TP + TN}{FP + FN + TP + TN} \tag{1.20}$$

$$ERR = \frac{FP + FN}{FP + FN + TP + TN} = 1 - ACC \tag{1.21}$$

Sometimes this measures are too general for the specific scope of our predictor, for example we can be interested on the precision of the postive cases predicted by our method or what is the probability to have false positive. For this reason other metrics are employed for give more insight on the predictor's ability. For getting information about TP is possible to compute the fractions of TP respect to all the positive predictions and it is called *precision* (PRE) but also the number of TP respect the the real positive cases could be interesting, i.e the *recall* or *true positive rate* (TPR). Otherwise the fraction of the retrieved cases respect all the negative ones, called *fall-out* or *false positive rate* (FPR), is a possibility with its complementary the *specificity* or *true negative rate* (TNR):

$$PRE = \frac{TP}{FP + TP} \tag{1.22}$$

$$TPR = \frac{TP}{FN + TP} \tag{1.23}$$

$$FPR = \frac{FP}{FP + TN} \tag{1.24}$$

$$TNR = \frac{TN}{FP + TN} = 1 - FPR \tag{1.25}$$

It does not exist a perfect metric that is always good, but it has to be evaluated case by case. For example we may be interested that the real positive cases were predicted with a good confidence and this mean to optimize the recall parameter, but in this way we have to be aware that the number of FP could be elevated. In order to try a balance between different measure other combinations were proposed, such as the *F1 score*:

$$F1 = 2 \cdot \frac{PRExTPR}{PRE + TPR} \tag{1.26}$$

Finally, another common way to control the performances for both TP and FP is the use of the ROC (*Receive Operating Characteristic*) curve. This method consist in building the curve of TPR vs FPR that were computed changing the cut-off on the selection parameter of the predictor. The diagonal correspond to the performances of a random predictor and if the curve goes above the diagonal it means that there is an actual prediction (the TPR is increasing faster than the FPR).

Part I

Protein-Protein interfaces and Antigen-Antibody recognition

Chapter 2

Trade-off between evolution and energy in PPIs

Biological pathways rely on a series of sophisticated networks of proteins that interact with each other. Usually those interactions depend on the characteristics of the protein surfaces that facilitate recognition and binding to the partner [13][31]. Modification of these patterns may provoke the insurgence of diseases. The study of the physicochemical traits driving the formation of protein-protein interfaces at a theoretical level gives the opportunity to isolate the segments responsible for the interactions increasing our knowledge of the relationship among sequence, structure and protein activity. Those information will have practical reverberations: in fact the reconstruction in atomistic detail of the patterns that guide surface binding could be useful for the development of new drugs that interfere with PPIs as target. Despite the fact that the performances in drug discovery are getting better, there is a variety of interesting PPIs that are considered difficult to target and only a small amount of compounds, directed to protein interfaces, arrived to drug stage in the trials[13][88]. For this reason a better knowledge of the surface features of protein interfaces may be helpful for a guided design of new proteins and small-molecules able to interfere with PPI formation.

Analysis and prediction of the regions responsible of the interaction among proteins has been conducted in different studies, from theory to experiments. Residue substitution tests on the surfaces highlight that just a limited amount of residues can strongly impact on the interface[14][36]. The accumulation of interactomic information, from sequence to structural data, support the development of computational data-driven analysis [146][112]. In this context, the improvement of the performances of techniques for interface prediction, using various criteria[46] and properties, is noteworthy. Generally speaking deVries et al.[46] describe three groups of features to use for prediction and two methods for combine them. The features are separated with respect to the level of information, some variables are linked to the kind of amino acids in the sequence (like statistical propensity evaluations), in other cases the information is extrapolated from the residue conservation among the protein family, and the third group of methods is related to three dimensional data of the protein structure. Those variables can be combined with two different approaches: patch-based method, where the values are averaged on a subsample of residues, or an approach based on the single residue value, whereby the variables are computed for every residue and then the most relevant ones are selected. In this approach, the sequence or spatial contiguity is not considered.

In the past years different methods that take advantage of diverse paradigms were implemented. Andreani et al.[10] reviewed a series of studies for the evaluation of the parameters that drive the binding zone predictions, varying from the use of mixture of variables that may regulates the interaction, like electrostatic properties or hydrophobicity, to the quantification of the evolutionary pressure on the surface, that showed a potential for improving structural based methods. Such concepts are extended in a review by Lua et al.[131] with the discussion on the improvement of docking algorithms using residue evolutionary conservation in support to binding affinity computation and the dovetail analysis on the surfaces of physicochemical and geometric properties. Advances in interfaces and contact prediction are obtained also with the employment of coevolution-based approaches[153][188]: an example is the work of Ovchinnikov et al.[150] where coevolutionary data extracted from multiple sequence alignment are used to compute a pairing score that resemble the coupling strength of evolution between two residues and then testing the system on 28 protein complexes. Coevolutionary data have also been used to build a prediction score with the combination of fragment docking and direct coupling analysis. The obtained scores supported the selection of multiple binding sites according to their druggability. Importantly, the technique was tested with proteins related to well-studied diseases[15]. Another flourishing field is the application of artificial intelligence[195]. Moreira et al.[141] developed a method for prediction of interaction hot spot building a score made by 881 features, with a mixture of sequence and structural terms, trained on a dataset composed by over 500 residues collected from 53 non-redundant proteins, achieving good accuracy. Others learning algorithms for hot spot detection were reported by Keskin et al.[113] suggesting that surface accessibility gave the most consistent impact. Remaining in structural based approaches it is interesting to cite the work by Kuttner et al.[121][120]. The authors study the dynamics of the backbone measuring residue stability assuming that interface recognition is guided by complementarity of "*stability patches*", finding those areas in the neighbourhood of the binding zone centre. Furthermore, predictive information can be retrieved from interatomic databases making advantage of the conservation of protein-protein relationship among species. Through network analysis, it is possible to obtain interologs, interactions of the orthologous genes of other organisms, for the prediction of the interactions of the studied protein. The server performs with a specificity from 72% to 98% while the sensitivity remains below 59%[94].

In this chapter we tackle the question whether the analysis of energetic terms and evolutionary profiles of residues on the surface can be harnessed to identify the regions more likely to be involved in interactions. A necessary observation is that the aim is not to estimate the binding energy between two interacting proteins. It has been shown that a protein complex may be stabilized with the help of suitable forces among the protein involved [200]. A considerable role is played by the effect of solvation terms as it has been shown that the residues in the interface are the less prone to be penalized by desolvation [90]. Finally, the dynamic rearrangement of the sidechains of interfacial residues could be clearly highlighted with elastic network methods[219].

A further step is to compare the chemical and physical traits with evolutionary pressure in the framework of the consideration of the ability of one protein to bind other proteins.

In this context, we formulated the hypothesis that sequence features are linked to the preservation of residues relevant for protein functionalities and energetic features are connected to the stabilization of the molecular structure: protein interfaces, essential for the correct activity of the protein inside the cell, have been subjected to evolutionary pressure to choose and maintain those physical and chemical configuration that ensures operative interactions. Starting from this concept, the quantification of the conservation of the amino acid type on the protein surface may disclose the residue preserved by evolutionary constraints for maintaining their activity, in the case of PPIs the interaction with a partner. In parallel, the evaluation of intraprotein energetics, that includes the interactions among residues of the same protein, in combination with structural characterization may reveal the presence of coordinated networks of residues apt to interact with a partner. The concrete interaction among two molecules can effectively take place if viable adjustments cause a bound state free energy smaller than the unbound state one. The observations that residues can be arranged in sectors could be seen as an indicator of energetic and evolutionary traits[125][130][98]: the division of the protein in subgroups could be functional. In fact while one sector works for the maintenance of structure stability, others groups can form an interface in a protein complex. The assumption is that two approaches, that are grounded on distinct assumptions, may complete each other and produce a convenient comprehensive picture, recapitulating the evolutionary and physicochemical requirements for interface definition. Here, we therefore set out to propose a comparison and a combination of two previously developed approaches, namely the Matrix of Local Coupling Energy (MLCE) for the energetic term and the Evolutionary Trace method to investigate the evolutionary preservation of the residues.

2.1 Computational implementation

Data set construction

Firstly a wide data set is gathered choosing various monomeric proteins involved in the establishment of molecular complexes. Through experiments are determined the conformations of the binding monomers taken alone and the corresponding complex. In this way, we have knowledge about the "*best guess*" for the real interface region for all the isolated partners, that information can be used as a reference for the performance testing of those algorithms we are going to test. The structures are collected from the 5th edition of the "docking benchmark" constructed by Vreven et al.[202]. The original set consists of 230 unbound monomers and for every couple there is the respective complex. The choice of the candidates is aimed at obtaining a coverage of small-medium monodomain proteins, which vary in length from 29 (1LU0, trypsin inhibitor) to 597 (1R42,ACE2) residues. Moreover, items that are involved in 3 or more interactions, cases of multiple protein family, man-made or composite proteins and protein without evolutionary annotations are discarded from the selection, decreasing the bulk of the original set to 84 complexes and 163 monomers. In addition to this, the data set employed by Scarabelli et al.[173] is integrated for a better coverage of cases with antigens, ultimately resulting in a set of 103 complexes and 183 monomers.

In the released docking database there is a subdivision of the PDBs according to their rigidity in binding: rigid-body, medium and difficult. Rigid-body refers to the proteins that display minor modifications when the monomer undergoes binding, whereas medium and difficult groups are the structures that manifest an appreciable reorganization, defined by the intensity of the variation: small for medium and large for difficult. The measure used by Vreven et al. to discriminate among the groups while evaluating protein-protein docking results is I-RMSD. This measure is obtained by overlapping the unbound monomer on the complex, then computing the RMSD of the Cα atoms of the sites that have at least one atom in the radius of 10 Å from any atom of the partner. Complexes where the I-RMSD larger than 2.2 Å are labelled as difficult, those having an I-RMSD lower than 1.5 Å and a ratio between non-native and native contacts that is below 0.4 are marked as rigid-body. All the other cases are labelled as medium. In the final database we include 107 rigid-body, 31 medium and 25 difficult systems, the full list is referred in appendix B.

Surface and interface definition

Usually the demarcation between inner (core) and surface residues could be ambiguous. For our purpose the boundaries are chosen with the use of solvent accessible surface area (SASA), considering a residue on the surface if the total contribution to SASA of all the atoms is equal or bigger than 25 Å^2. This definition has been applied for every analys, and the utilisation on the monomers taken alone. For the extraction of the interface residues from the crystal complexes the SASA data are employed as well. Operatively, SASA for the entire complex and the SASA of the monomers separately are computed; every residue in which the total SASA differs more than 1 Å^2 is assigned to the interface. In case we are dealing with antibodies (Abs) is necessary to apply an expedient calculation. In fact Abs are composed by a heavy and a light chain but both EVT and MLCE analysis are conducted on monomers; therefore, only the heavy chain is considered and when calculating the SASA the residues on the inter-chain interface are omitted.

For the effective interface evaluation-prediction the workflow is split in two steps: first the residue are ranked according to a score then top-ranking sites are grouped in connected patches.

Scores definition

Once surface residues and binding zones are determined, it is necessary to define the methods and scores for the energetically and evolutionary analysis. The internal energies are obtained with the support of MLCE technique, that evaluates a reduction to eigenvectors of a NxN matrix of non-bonded potential terms, where N is the number of residues. The nonbonded terms include van der Waals, electrostatics interactions and solvent effects. The eigenvalue decomposition emphasizes the regions where energetic pairing among residues are more robust or more breakable: the surface segments the are weekly coupled with the rest of the protein are the ones more prone to establish an interaction without interfering with the global stability. Summarising, we can say that the supposed interfaces consist of frustrated energetic couplings. The energetic couplings are computed using an MM-GBSA approach with the amber14 software, the force field employed is ff14SB. The obtained interaction matrix is approximated by the first eigenvalue, as expressed in the equation 1.9, and therefore the per-residue contribution is determined by the first eigenvector, respect to the original MLCE the approximated matrix is not filtered by a contact map but is only used for the computation of the scores. As previously mentioned the quantification of the contribution of evolution is computed with the Evolutionary Trace method, an improved version of the information entropy. Through the scanning of an ensemble of sequences belonged to the same protein family, the information entropy is computed in a hierarchical tree architecture: a lower entropy value reflects an higher conservation for the selected site, therefore an higher propensity to occupy a role in the protein activity. The calculation are made with the Mammoth server of the Lichtarge group (`http://lichtargelab.org/software/ETserver`) using the real value Evolutionary Trace (rvET) defined by the formula 1.14 with an uniform weight distribution, thus the equation is reduced to:

$$s(k) = 1 + \sum_{n=1}^{N-1} \left[\frac{1}{n} \sum_{g=1}^{n} p^g(\pi, k) ln \left(p^g(\pi, k) \right) \right] \tag{2.1}$$

where the score s for the site k is obtained computing the occurrence frequency of an amino acid of the π type in every instance of the subtree g. The values are computed with the default setting of the server: the reference sequences are obtained from the database (`/mammoth/blast/data/customuniref90`) derived from BLAST with a e-value limit of 0.5 and then aligned using muscle algorithm[53], the phylogenetic tree is obtained with the UPGMA approach using the blosum62 matrix as a base[182][101].

In this work three different scores are computed for every residue, the definitions for a site i are:

- **MLCE score** is the module of the i-th component of the first eigenvector normalized by the maximum value in the vector.

- **EVT score** is the real value obtained with the formula 2.1 for the i-th site divided by the maximum value of the whole protein.

- **EVT + MLCE** is direct sum of EVT and MLCE score.

Various feasible scores were tested, like the use of logarithm, however no significant improvement was detected.

Building potential interaction patches

The top-ranking sites alone individuate only hot spots, but we are interested in networks of good-scoring connected residues. Therefore the following step is the construction of surface patches bringing together the best scoring sites. To this aim, the proteins are represented as a graph, where the best-scoring residues are the nodes and the contacts connecting them are the edges. A contact between two spots is placed if the distance of the Cα atoms is below 9 Å. After the graph is built, it is possible to extract subsections of the graph that are completely connected, i.e. the subgroup of nodes where all the nodes are joined through a path. These subgraphs are called connected components. An easy way to identify the nodes that take part in a connected component is the deep first search (DFS) algorithm[196], described in table2.1.

Step	Description
1	Make a list of every node (best-scoring residues) and associate a label to each node
2	Reset the labels for every node and set a general variable clusterID to zero
3	Select the next node in the list
4	If a selected node has label equal to 0 assign the current value of clusterID to the label variable, continue to step 4. Elsewhere return to step 3, because the current node was assigned to a cluster before.
5	Perform a cycle to every neighbour (other nodes connected to the current one) and call DFS recursively.
6	When all the neighbours are visited the DFS recursion stops. Add 1 to clusterID an go back to step 3.

Table 2.1: Schematic example of the DFS algorithm used.

The residues are ranked according to their score, only the cases with a score smaller than a cut-off are selected. In fact, a low value corresponds to low energetic coupling (for MLCE score), high evolutionary conservation (for EVT score) and both the features for EVT + MLCE score. Since the ranges of the values can vary from case to case, we preferred to use a system-dependent cut-off on the scores and a general cut-off on patches dimension. The cut-off value is determined starting from 0.1 for EVT score and 0.2 for EVT + MLCE score and building the relative patches; at this point if the biggest patch has less than 10 residue the procedure is repeated increasing the limiting values by 0.1 for EVT and 0.2 for EVT + MLCE, as far as a patch with 10 or more spots is discovered. Energetic scores show a different distribution with respect to EVT; for this reason the cut-offs of MLCE scores are modified in a logarithmic fashion: $[10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 5 \cdot 10^{-3}, 10^{-2}, 5 \cdot 10^{-2}, 10^{-1}, 5 \cdot 10^{-1}, 1]$. It is necessary to stress that the experimental information of the interfacial residues are used only for comparison with our prediction, since the measure are evaluated on isolated monomers without complex information. Furthermore, as a clarification, I want to point out that both the approaches used do not expect to compute the binding affinities or other pairing terms: in fact, there are several features involved in binding affinity estimation, like protein or water conformational changes, that are not considered in our methods.

2.2 Results

The practical assumption that is behind the analysis is that sites that lie in the binding region are more preserved by evolution (in order to maintain a proper functionality) or have a low participation in the reinforcement of protein energetic stability. To be more precise, it is plausible to imagine that a residue with a fundamental role for the function of the binding site needs to be preserved; on the other hand, a substantial variance in amino acid type, for sites in surface zones that are not designated for protein-protein interaction, is expected. This different behaviour should be highlighted by a distinct entropic measure for interfacial residues respect to non-interacting ones. In parallel, a focus on the fold energetics should determine the presence of chemical patterns for the binding region, analysing the strength of amino acid pair interactions and their involvement in the global structure stability: as it is often necessary that the interface conforms with the partner surface, it needs the possibility to undergo conformational adjustments. Thus, residues forming the binding region should not be intensely coupled to other segments of the fold.

The analysis on sequence-based and structure-based data has the goal to determine sets of residues that display a distinct arrangement with respect to the rest of the molecule. First of all, it is necessary to inspect energetic and entropic data to verify if is possible to statistically discriminate binding zones from the rest of the protein surface. For this purpose, the interfaces individuated in the experimental complexes are used as a "*ground truth*" for the real binding regions; then the residues are split in two groups: one with interfacial residues and the other with the rest of the surface. With the use of the Kolmogorov-Smirnov (KS) statistical test[3], the differences between the profiles of the real binding residues (sample) and the remaining surface ones (reference distribution).

At this point, a study of evolutionary differences of interaction sites respect to a reference distribution is set out. The hot spots for functional relevance are determined through the evaluation of entropic information in an evolutionary tree built from multiple sequence alignments. The evolutionary profile are determined with the real value Evolutionary Trace equation 2.1 [138]. For the energetic characterization of PPI binding regions, all the energetic coupling between residues are computed in all the protein structures in the dataset using the previously mentioned MLCE method. The first eigenvector is used as energetic profile without filtering the approximated interaction matrix. The components of the eigenvector are assigned to the relative residue and the set split in two groups according to the principle described before.

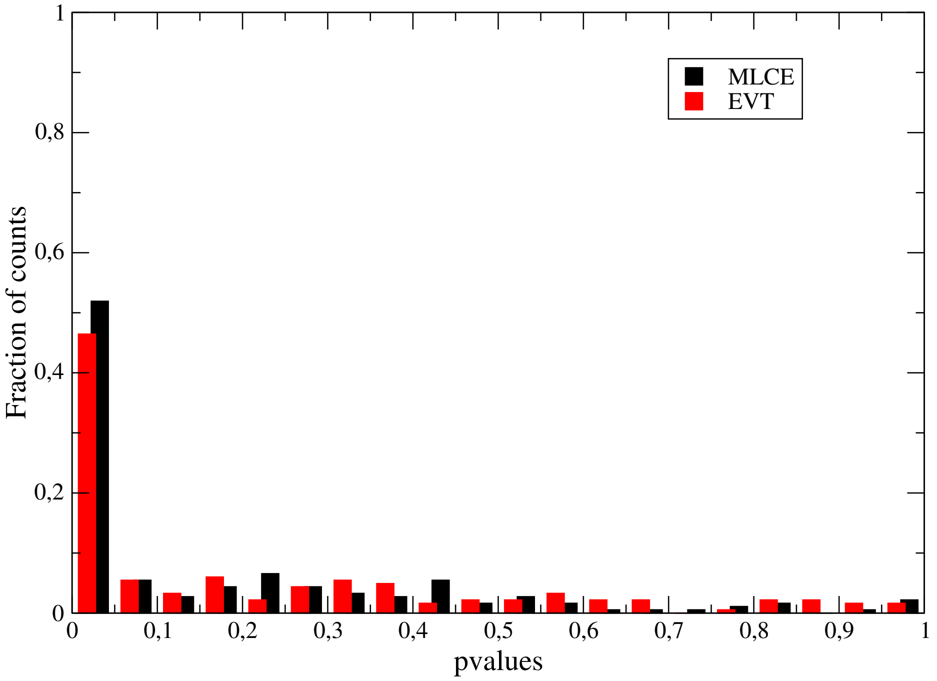


Figure 2.1: Here are showed the histograms of the pvalues obtained in the Kolmogorov-Smirnov test. Every count reflects a protein-protein complex included in the benchmark. Both the EVT and MLCE profiles display a pronounced signal in the region where pvalues are below 0.05, meaning that in both cases the measures are capable to discriminate real binding regions for several complexes.

After the profiles of EVT and MLCE are obtained for the whole dataset of 183 systems, for every protein the data are split in sample and reference distribution for the statistical test. The Kolmogorov-Smirnov test obtains a significant fraction of instances with a good confidence level (pvalue smaller than 0.05) that the two distribution, the sample and the reference one, are different: in the case of EVT values approximately 46% have a p-value below 0.05, percentage that grows to 51% in the case of energetic data (MLCE), Figure 2.1. If the limit for good separation is raised to a pvalue of 0.1 the percentage increase to 51% and 57%, for EVT and MLCE respectively. Those results are checked doing again the test with the Anderson-Darling (AD)[9] statistical approach, another test to compute if a data set is probably drawn from a reference distribution. The number of well-separated interfaces are comparable to the KS ones, obtaining a percentage of cases below a pvalue of 0.05 near 50% for both EVT and MLCE (see Figure 2.2), for a full list of the values see appendix A.

The preceding analyses exhibit a signal for the proposed measures in the case of interface residue, suggesting that zones of the surface, in a consistent number of cases, are likely to be split in sites of interaction and non-interactions ones. The following step is to ask if this knowledge may be helpful for interface prediction. To answer this question, all the database of structure is analysed ranking all residues with an EVT or a MLCE metric. Observing the procedure described

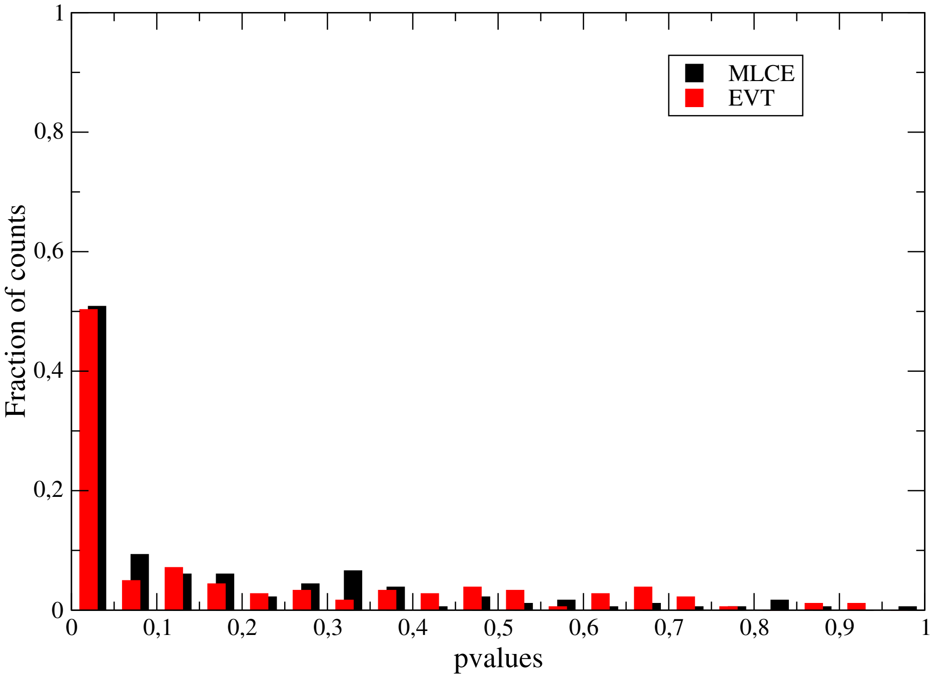


Figure 2.2: Here are showed the histograms of the pvalues obtained in the Anderson-Darling test. Every count reflects a protein-protein complex included in the benchmark. Both the EVT and MLCE profiles display a pronounced signal in the region where pvalues are below 0.05, reproducing the similar behaviour that is observed for the KS test.

before hypothetical surface patches of interaction zones are generated and three different scores are used for the ranking: EVT, MLCE and EVT + MLCE scores. The predicted binding regions vary in number and size, depending on the score chosen. The reference interfaces obtained from crystals have an average dimension of 25 residues. The computed patches superimpose to the area of the reference ones in more than half the cases, for EVT score it happens 77% of the times, 75% for MLCE and up to 85% with the union of EVT and MLCE. The mean size of the generate patches is of 20 residues for MLCE and 10 for EVT score, with a mean of 2 and 2.4 numbers of patches, respectively. EVT + MLCE score generates 2.5 putative sites per case on average, those sites have a mean dimension of 12 residues. Despite the extent of generated patches is on average lower than the actual interface, is observed an overlap between putative and “real” patches in a consistent percentage of systems. Occasionally, a bigger overlap is observed with the combination of two or more computed binding sites that are close in the protein surface (see figure 2.3).

The assessment of the quality of the value of the obtained predictions is conducted through a Receiver Operating Characteristics (ROC) analysis, run over the whole structure databased considered. The generated patches are compared with the experimental ones counting the predicted spots that lie on the interfaces. To be more precise, the value of true positive (TP), false positive (FP), false negative (FN) and true negative (TN) is calculated, with the definitions:

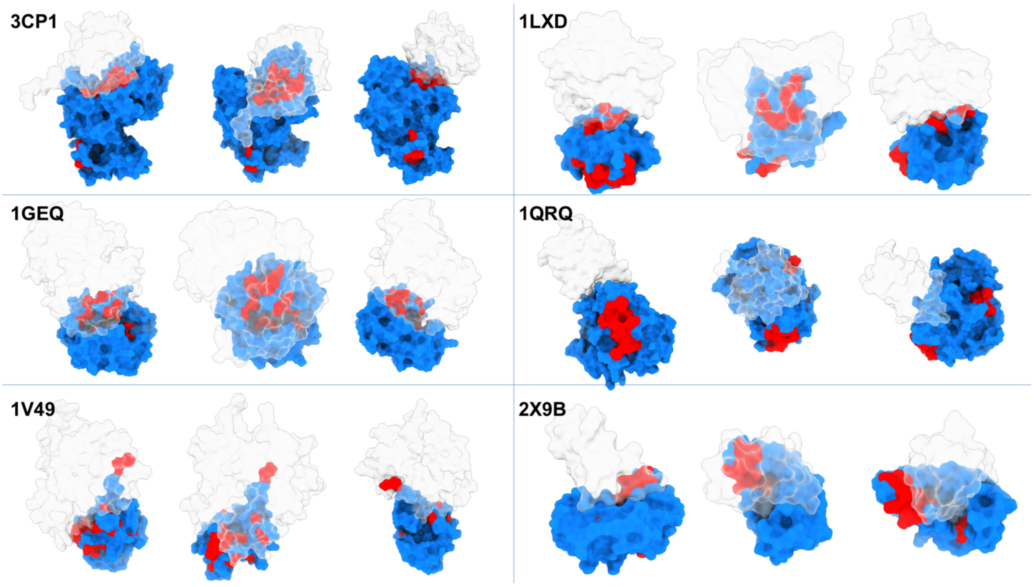


Figure 2.3: Here are illustrated some examples of generated binding sites, computed with the score MLCE + EVT. In blue colour is indicated the target protein while the binding partner of the experimental complex is represented in transparency, and the predicted patches are coloured red. In every case three distinct orientations are presented. The choice of the systems covers good, medium, and bad predictions. The PDB codes of the structures are listed from the left to the right and top to bottom: 3CPI, 1LXD, 1GEQ, 1QRQ, 1V49, and 2X9B.

- **TP**, the count of experimental binding residues that are properly predicted.
- **FP**, the count of non-interfacial residues that are predicted as hot spots.
- **FN**, the count of experimental binding residues that are not predicted.
- **TN**, the count of non-interfacial residues that are not predicted.

Two variables that help in the estimation of the goodness of the prediction are the true positive rate (TPR) and false positive rate (FPR) defined by equations 1.23 and 1.24. A higher value for the TPR rate means encouraging prediction, in this case the FPR should be small. The ROC curve gives the TPR rate in function of FPR, the slope of the curve will indicate if is more favourable to find TP respect FP when the margins of the patches are increased. The ROC curves are computed varying the cut-off quantity in the building patches phase of the interface prediction process, the interval of variation are those previously indicated. A simple number that discriminates if a ROC curve describes a good or a bad predictor is the Area Under the Curve (AUC): if the value of AUC is greater than 0.5 we can estimate that the predictor is superior respect to a random one. The percentage of cases with AUC above 0.5 is 54% for MLCE score and 71% for EVT score and the mean value is 0.51 and 0.58 respectively. The combination of EVT + MLCE gives an AUC bigger than 0.5 in 70% of the systems, with a mean value of 0.58. If the curves are constructed considering the predictions of all the residues of the dataset at once is easier to visualize the global performances of the three scores, with a confirmation that EVT have a good AUC in more case than MLCE, EVT + MLCE do not improve globally (see Figure 2.4).

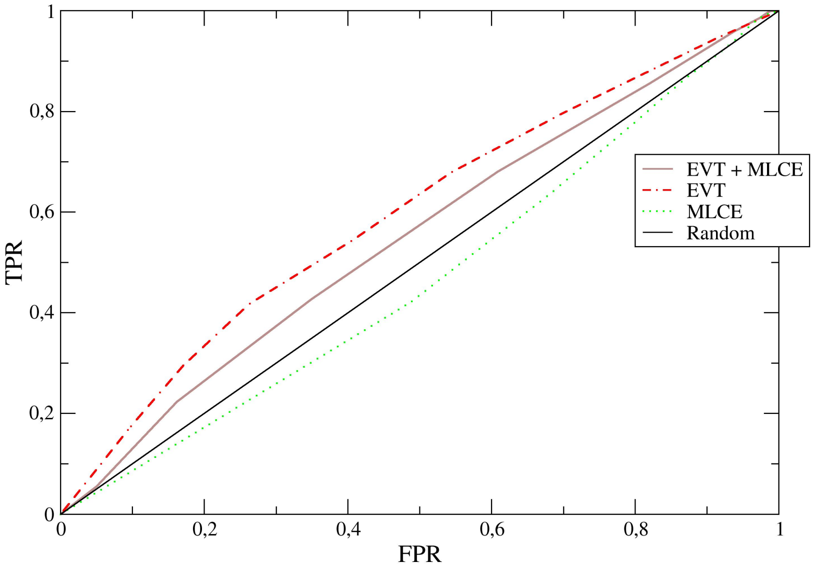


Figure 2.4: Examples of ROC curves. These are the curve produced with the addition of the predictions made considering all the systems as a one. ROC curves are coloured in red for EVT, light green for MLCE, and purple for EVT+ MLCE, while in black is signed the theoretical curve produced by a random predictor.

It is important to remember that not all the protein-protein interfaces are designed in a lock and key fashion, but in many cases the region suitable for binding is subjected to a wide structural reorganization while changing from unbound to the docked state. In our model the monomers are evaluated only in the state prior to binding and the possible conformational modifications are not taken into account. This could limit the prediction ability. Anyway, the analyses conducted ascertain that performances even improve when the protein undergoes a substantial structural variation. In order to verify this condition, it is possible to split the data set in three groups according to the respective rearrangement observed in binding and the correspondent level of complication for the prediction: rigid-body (no reorganization), medium (light reorganization) and difficult (large reorganization). The grouping criteria were previously defined. In the database the rigid-body class has a percentage of good prediction of 72.8% for EVT score and 53.2% for MLCE, those values raise to 83% for EVT and 58% for MLCE in medium class and to 80% in EVT and 60% MLCE scores for difficult cases. The EVT + MLCE score have a similar trend: 72.8% of positive predictions in the rigid-body class, whereas in medium and difficult classes the rate goes, respectively, to 80.6% and 84% (for a complete list of the values see appendix B. Since EVT score is sequence-dependent it is not surprising that the performances could increase for difficult targets. On the other hand, it is interesting that this behaviour also emerges for the MLCE score, that is structure-dependent.

These results suggest those descriptors are capable to take advantage of the information related to the evolutionary pressure for the definition of an optimized region as a binding one, unbeknown to others. In addition, since MLCE is focused on the determination of poorly coupled regions provided with favourable structural organization that encouraging adaptability, arise the hypothesis that such energy-based measure can localize the segments that are flexible enough to rearrange their structure in the moment that the partner establishes an interaction.

Impact of protein function on performances

Starting from the line of reasoning stated above, it is reasonable to ask the question if EVT and MLCE display a different behaviour in the case of a particular protein family and if the existence of a link to the biological role can be hypothesised. In order to answer these questions, the entire database is split into 6 functional classes: antibodies, enzymes, antigens, inhibitors, signal transduction, and structural proteins (that are molecules that bind to the extra cellular matrix (ECM) and assemble the cytoskeleton, such as actin, profilin, metavinculin and twinfilin).

For enzymes and signal transduction classes similar trends are observed, the EVT score gives a percentage of good predictions, respectively, of 81% and 83% that are greater than 51% and 47% for the MLCE score, see Figure 2.5. In the case of enzymes the binding site requires to be intensely preserved in order to maintain those residues indispensable for the correct interaction with the substrate, compel it in the right spot during the pre-reactive phase and complete the desired chemical reaction. For this kind of problem the MLCE score works worse that the EVT metrics, for the reason that it searches for segments that are detached from the energetic core of the structure and display adaptability, whereas the active site of enzymes usually shows rigidity and structural coordination. Signal transduction cases face a comparable problem, with the sequence that need to be preserved and the configuration require a defined coordination for the correct regulation of the reactions (like sulfation, phosphorylation, hydrolysis, etc.) that carry the information during the pathway.

For inhibitor proteins, both the scores present good results: 75% for EVT and 80% for MLCE, suggesting that for these molecules there is a conservation of amino acids without losing conformational flexibility. The connection of the inhibitory function, that may modulate the activity of enzymes and signal transmission, with a little evolutionary entropy in the binding site could suggest a pressure for the choice of the optimal organization of residues to obtain a favourable interaction with the partner. On the other hand, the interfaces have weaker energetic pairings compared to the other parts of the protein sustains the concept that inhibitors can bind to two or more partners with the use of a versatile remodelling. These observations can still be applicable in the case of structural class that present 72% of positive predictions for EVT and 81% for MLCE, by which the sequence preservation in the binding site allows correct identification of the partner and construction of the complex at the base of cellular skeleton. At the same time, weak couplings help in preserving the necessary flexibility to match different binding targets (occasionally in crowded surroundings).

class of protein function	fraction of positive cases			number of elements
	EVT	EVT + MLCE	MLCE	
antibody	0.3	0.2	0.4	10
enzymes	0.83	0.77	0.51	53
antigens	0.33	0.41	0.5	24
inhibitors	0.75	0.8	0.8	20
signal transduction	0.81	0.78	0.47	65
structural	0.72	0.81	0.81	11

Table 2.2: Summary of the counts of positive predictions for EVT, MLCE, and EVT + MLCE scores, separated for the defined classes of activity of the proteins.

Lastly, the antigens and antibodies classes stand out in a noticeable way. For antigenic proteins, the region of interaction has commonly a low evolutionary pressure, likely because a conserved trait could limit the activity of the infectious agent in the body. The resilience to mutations is a weapon for an antigen to avoid the control of the host immune system, while keeping the structural characteristic needed for the pathogen life cycle. On the other hand, antibodies show a preserved conformation for the binding region, with the participation of an hypervariable zone that is necessary to follow antigens modifications. The existence of that highly mutable region could interfere with the EVT predictions, whereas the organized energetic couplings may be individuated by the MLCE. For these classes the performances are greater for MLCE that the EVT score: 50% (antigens) and 40% (antibody) compared with 33% and 30%, the results are summarized in the table 2.2.

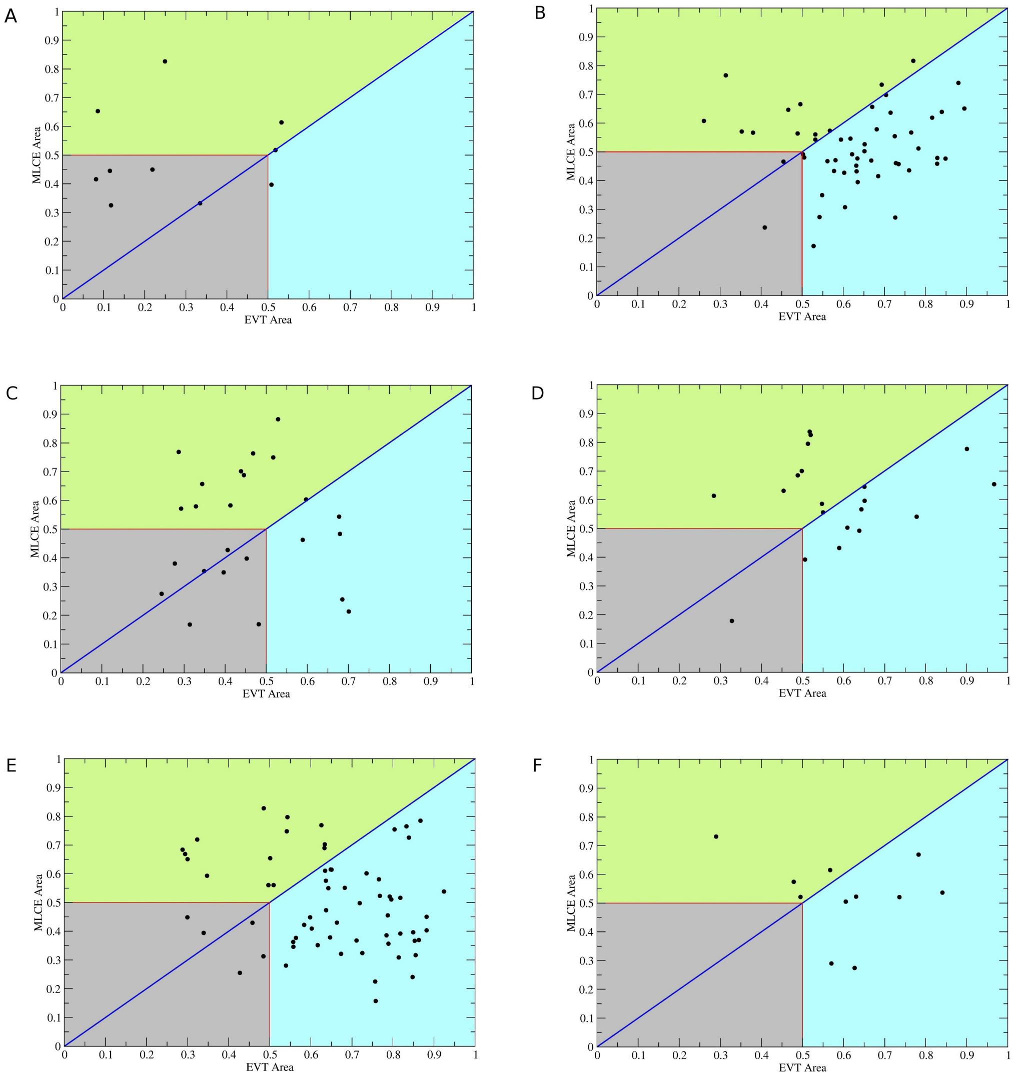


Figure 2.5: Scatter plot of the values of AUC computed with MLCE score versus EVT score for every class, the letters refer to: (A) antibodies, (B) enzymes, (C) antigens, (D) inhibitors, (E) signal transduction, and (F) structural. The diagonal line (coloured blue) split the points that have a MLCE score greater than EVT (green region above) from the cases in which EVT does a finer work (underlying light-blue region). The grey square circumscribes the region of point missed by both measures, i.e. the systems in which the prediction with our approach is harder. Enzymes and signal protein exhibit not many cases in the grey square and the majority of points are situated in the light-blue zone. Antigens have the majority of the cases in the green zone and 8 cases in the grey square, indicating that for EVT and MLCE scores it is harder to predict this class. Inhibitors are even distributed among above and down zones and only 1 case is situated in the grey square. Structural proteins exhibit an empty grey square and many of the cases are in the light-blue region. In the end, antibodies result the most challenging to be predicted, in fact with fifty percent of the points in the grey square, almost all of the correctly predicted are in the green region.

2.3 Conclusions

In this chapter, I showed and discussed the viable exploitation of evolutionary and structural stability knowledge for the detection and characterization of the surface regions designated for protein-protein interactions. Firstly, the hypothesis is formulated that the site of the protein intended for the selection of the specific target in an important pathway could need an evolutionary pressure to preserve the characteristics that allow his activity. In addition, the binding process is regulated by the physical and chemical features of the amino acid involved. Successively, we suppose that with the definition of an evolutionary and an energetic measure, for the consideration of the set of problems previously mentioned, it is possible to construct a score that helps in the individuation of protein-protein interfaces.

It is intriguing that for some functional classes (such as the one responsible for signal transduction and the enzymes), evolutionary pressure is the driving factor for the regulation of the binding site, whereas in other classes energetic couplings have a relevant influence, like the cases of antigen-antibody recognition, inhibitors and proteins with a structural role. With those results, it is possible to argue on the existence of a balance between the measures. With this background it is suggested that when studying a new system, without any knowledge about the interactions, it may be helpful to make use of a combined approach, of the scores discussed, in order to obtain a first guess on the placement of the putative binding region. Successively, from the experimental point of view, it is possible to intervene with perturbations of those sites with a mutation analysis or the development of new compounds suitable to bind in the interface[13]. As a final clarification, it is important to notice that the methods presented in this chapter are not intended to evaluate binding energies or affinities among the proteins forming a complex: for that scope there are various supplementary properties to be considered, for example the intensity of the interactions on the interfaces involved, the dynamical aspects of the interacting site compared between the complex and the monomer taken alone, the consequences in the inner motions of the protein and, lastly, the solvation activity of the features used. The method defined in this work serves the important purpose to help to formulate initial guess of putative patches on the surface involved in binding and, for example, could be complementary to other approaches like protein docking algorithms[48][174][11], with the use of binding constraints in support of the process.

Chapter 3

Detection of PPIs in antigenic proteins

There is a little number of events that have a such huge impact on public health and economic system as the insurgence of infectious diseases[156]. Moreover, in spite of the great improvements brought to pharmacology, the design of a cure for many known pathogens is still a challenging task.

Vaccination and antibody based therapies represent the best options to cure and prevent the spread of diseases for which drugs are not available. At the cell level, the function of the human immune system is based on a series of protein-protein and peptide-protein interactions. The study and characterization of those interactions could help in the development of engineered proteins (or peptides) as antigens to be used in diagnostics or for vaccines production.

During my PhD I was interested in the computational characterization and prediction of antigenic interaction sites of the SARS-CoV-2 Spike Protein.

3.1 Antigen recognition in the immune system

The core defence to fight with pathogens is a sophisticated mechanism, composed by cells and various molecules, called immune system[154]. The correct functioning of this system requires complex processes, like learning and saving in memory precious information, that entail the participation of a vast amount of factors.

Firstly, it is possible to define two kinds of response of the immune system: innate immunity and adaptive immunity. Innate immunity is an early reaction to infectious agents that is not specialized for the pathogen to fight and, also, does not develop resistance to secondary exposures to the same agent. On the other hand, adaptive immunity requires time to spring into action but it is specific to the invading agent with the generation of an immunological memory. The vitality of both the reactions is necessary for the proper operation of the immune system. In order to achieve adaptive immunity the organism possesses an important typology of white blood cells, called lymphocytes[186]. In particular the activity of these kind of cells could be distinguished in two distinct reactions: a cellular response that is carried out by the cytotoxic T lymphocytes (CTL), named also killer T cells, and a humoral response that consists in the secretion of antibodies by B type lymphocytes. The activity of these cells is supported by an additional class of lymphocytes called T-helper cells.

A fundamental step for both the responses is the individuation of antigens and the region of the antigen that is recognised by lymphocytes, called "*epitope*". In the case of T cells, the lymphocytes recognize the major histocompatibility complex (MHC), a specific protein type widespread on the membrane of the host cells. In immune responses, pathogenic protein antigens are commonly degraded and the resulting fragments are exposed in the form of peptides on the membrane with the help of MHC, to prompt inspection T cells. When a CTL recognizes a specific peptide in a MHC, a response is activated which starts to induce apoptosis (programmed cell death) to the infected cell. For the recognition of the epitope in complex with MHC, the T cells possess receptors molecules, T Cell Receptors (TCR). Since there are many possible epitope to identify the receptors are expressed in a huge variety, it is estimated that any human presents at least 10^8 different receptors at the same time.

On the other hand, the humoral response is based on the activity of B cells. Those lymphocytes are able to produce antibodies (Abs), proteins composed by immunoglobulin domains (Ig). There are five types of Abs, IgA, IgG, IgD, IgE and IgM, and all of them share the same essential subunit formed by a tetramer composed by two heterodimers; each heterodimer is divided in a light chain (LC) and a heavy chain (HL). These chains form the typical "Y" shape, with the binding region on the terminal arms of the Y. This site is called "hypervariable loop" because it is produced in a multitude of variations in the organism, in order to better adapt to the antigen.

One B cell is specialized in production of a single Ab type, that is located on the membrane as long as the cell remains inactive. In that position an Ab is able to bind to antigenic material thus activating the B cell that starts proliferating and producing other Abs that are eventually secreted. With respect to T cells, Abs have an increased recognition capacity because in addition to peptides they can bind to whole antigens, such as proteins or viruses. The epitopes recognized by Abs can be linear or conformational, the latter consisting of residues that are physically situated at distinct position in the protein sequence but are grouped in a contiguous surface in the 3D fold.

Once an Ab is bound to the antigen, it may neutralize its activity, and this inhibitory effect is considered of significant relevance in controlling viral infections[64][56]. In other cases the binding is not sufficient for the inhibition of the infectious agent but those non-neutralizing antibodies can still be protective, for example acting as a marker for the intervention of other immune cells.

In an organism the collection of the available antibodies depends on environ-

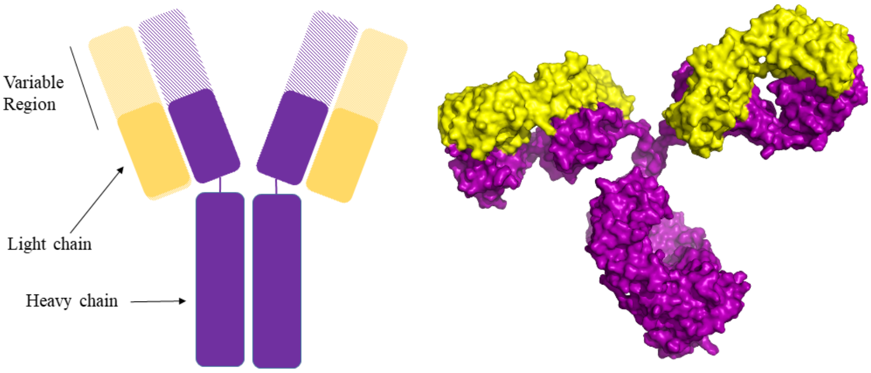


Figure 3.1: On the left there is a scheme of the essential subunit of an antibody. Different colours highlight the separation among light and heavy chains. With dotted areas are filled the variable region of both the chains, on the far end of the structure in correspondence of the antigen binding site. On the right there is a surface representation of the subunit (pdb code: 1igt) with the typical Y shape.

mental influences endured during lifetime. The first time an organism is exposed to a pathogen the B cells start proliferating and, thanks to the high propensity of antibodies to mutate and adapt, the response becomes more and more efficient. The information acquired is then stored in memory cells. In the case of a secondary exposure to the same antigen, memory B cells are activated resulting in the secretion of an increased number of high-affinity antibodies compared to the first time. This learning ability is known as “ secondary immune response ” and is the core mechanism of vaccination.

Vaccines production started with the inoculations of attenuated living pathogens. Over the decades this method revealed some limitations, such as the impossibility to grow *in vitro* the infectious agent or a pathogen that do not activated humoral response because his life cycle is inside the cell, therefore it requires T cells to be eliminated. With the improvement of sequencing techniques it became easy to obtain information about the pathogen proteome. At this point it was possible to develop a vaccine without the employment of the whole pathogen, in fact only the antigen that activates the immune reaction is required. This novel technique is known as Reverse Vaccinology [162] and was successfully used for the first time for the vaccine against Meningococcus B[71].

3.2 Epitope prediction in multidomains proteins

The approach employed for the prediction of the epitope in an antigenic protein refers to the idea used before for general protein-protein interfaces: the residues necessary for the maintenance of the core stability of the system have little role in establishing protein interactions.

The MLCE method is developed for the detection of the residues that are poorly energetically coupled to the structural core: in fact the individuation of such residues pairs permits to define regions with frustrated bonds. However, in the previous use of this method the decomposition of the matrix of non-bonded interaction energy E^{NB}(van der Waals, electrostatic) is truncated only to the first eigenvector (vector with the lowest eigenvalue). In many cases this approximation is capable to detect residues important for the stabilization[193][37][143] and it generally holds in case of monodomain proteins. For multidomain systems it has been shown that an increased number of eigenvector is required. Therefore the analysis of these cases requires further improvements introduced by Genoni et al.[95]. If we think that an eigenvector defines a region of well-connected residues pairs (a block in the interaction matrix) for one domain proteins, when more domains are considered the interaction matrix becomes composed by more blocks, each one of them defined by different eigenvectors. If we select for every domain the eigenvector that optimally represents, it is possible to obtain the approximated matrix:

$$E^{NB} \approx M^{NB} = \sum_{i=1}^{N_D} \lambda_i \omega_i \omega_i^\intercal \tag{3.1}$$

N_D is the number of proteins domains, this matrix will have a block structure corresponding to the domains position. To deal with more realistic cases, it is necessary to consider that the eigenvectors of the E^{NB} matrix consist in a superposition of different components (groups of distinct blocks). Therefore, more than a single eigenvector will be necessary to obtain a coverage of every domain. In order to use the bare minimum number of eigenvectors, it is necessary to minimize the overlap between vectors reducing the redundancy of the representation. The eigenvectors are filtered in order to highlight only the most relevant elements, selecting the components that have a value greater than a threshold (median of the distribution of the absolute values). Once all the relevant elements of every vector are computed it is possible to select the minimum subset that permit a complete coverage. The vectors are selected starting from the one that have the lowest eigenvalue, then the others are chosen extracting from the whole set the vector that display the smallest overlap of the relevant components with the vectors that are previously selected. These precautions help in the reduction of the redundancy of domain coverage, obtaining a final matrix:

$$E^{NB} \approx M = \sum_{i=1}^{N_e} \lambda_i \omega_i \omega_i^\intercal \tag{3.2}$$

with N_e the number of essential eigenvectors, that is usually larger than N_D. The essential non bonded matrix (equation 3.2) is then filtered with a symbolization scheme to highlight the most relevant couplings, obtaining a symbolized matrix that is further submitted in a cluster algorithm to delimit domains borders. At the end of all the processes the final matrix M^S will possess only those coupling that have the mildest and the more intense energetic interactions.

The epitope prediction is achieved following the MLCE procedure, where the decomposed energetic matrix is filtered removing non local couplings. A contact map C is built labelling a residue pair as neighbours if their three dimensional distance is below the arbitrary limit of 6 Å; the contacts are computed among H atoms for glycine, Cβ atoms for non glycine types and C1 atoms if we are dealing with glycans. The matrix of the local pairwise coupling energies (MLCE) is then obtained through the entrywise product of the approximated interaction matrix and the contact map:

$$MLCE = M^S \circ C \tag{3.3}$$

Once the MLCE matrix is computed the values of the local couplings are ordered obtaining a rank of energetic interactions according to their robustness, from the mildest to the more intense interaction. The extraction of putative epitopes is obtained picking only the coupling that are under a certain threshold defined in percentage, the width of the threshold will define the extent of the area of the epitopes.

3.3 Case of study: fully glycosylated SARS-CoV-2 spike protein

SARS-CoV-2 is the etiological agent of COVID-19, a severe respiratory disease with seriously disruptive socioeconomic impacts[24]. The spike protein (S) is the pivotal intermediary of the virus transmission and thus has captured most of the attention for the production of vaccinees and diagnostics, being the first point of contact between the virus and the host. In this section, I reported an application of the energetic decomposition method described before for the detection of subnetworks of frustrated interprotein interactions as putative epitopes for the fully glycosylated S protein.

3.3.1 Introduction

The insurgence of COVID19 syndrome begun in Wuhan, China, and then diffused around the world becoming a universal pandemic[69]. The pathogen causing this severe disease, the new coronavirus SARS-CoV-2, has infected more than 46.000.000 people and caused more than 1.200.000 deaths to date (source `https://coronavirus.jhu.edu/`). Since the infection is widespread on the globe the use of social distancing precautions may not be enough to control the diffusion worldwide. For this reason there, it is urgent to design new drugs or vaccines, that are the unique measures for disease containment to bring social life back to normality. There are various trials in progress for the development of vaccines (as reported in `https://www.nytimes.com/interactive/2020/science/coronavirus-vaccine-tracker.html`) or new tests for repurposing compounds known for the treatment of different syndromes[119][127][168]. The SARS-CoV-2 is surprisingly efficient in taking advantage of the cell host pathways to infect and replicate. This feature is typical in the *Coronaviridae* family, a group of viruses specialized in the infection of different species of animals and, in the human case, cause of various pathologies that afflict the liver, the nervous and respiratory system[8][210], as well as responsible for epidemics in the past[34][40]. Common to SARS-CoV and MERS-CoV, other members of the *Coronaviridae* family, the surface spike protein (S) has a central role in the process of cell infection, that happens through the interaction with the human receptor of angiotensin-converting enzyme 2 (ACE2)[215][194]. After molecular recognition the S protein is divided by the activity of serine protease, such as trypsin and cathepsins, facilitating the entrance of the virus in epithelial cells[168]. It is relevant that a great part of the vaccines studied for COVID-19 are based on recombinant expressions of the spike protein. The resolution of the complete structure of the protein S heterotrimer has been obtained with cryogenic electron microscopy (cryo-EM)[213][205][204] giving improvement in the comprehension of the molecular individuation of the receptor biding domain (RBD) of S protein by ACE2[215]. On the other side, *in silico* analyses bring the focus on the dynamics of the spike protein at an atomic level, elucidating the function of polysaccharides widespread on the protein surface[29][97][181]. Other studies put the attention on the factors that drive the viral binding to the human cell, i.e. the interactions among spike and ACE2 proteins[220][183][206].

This enrichment of information will assist the comprehension of the mechanism underlying recognition of the spike protein, acquiring details of the necessary traits to elicit effective Abs. The details gathered may be used for the study and development of enhanced antigenic material starting from S protein, such as the definition of domains or epitopes that can be reproduced with engineered peptides. This phase would occupy a central part in the selection and improvement of promising vaccines and Abs used for treatment and, moreover, the enhancement of diagnostic devices. Furthermore information gained regarding molecular interactions and recognition may be suitable in the future in the event of comparable epidemics with the employment of optimized methods to novel antigens. To be more precise, in case of an outbreak of a new infectious agent, versatile computational approaches could be quickly adapted to select and design engineered protein or peptide based vaccine candidates.

In this section, I present a study on three dimensional structures of the S protein with its glycosylation resolved. The conformers are taken from the atomistic molecular dynamic (MD) trajectories supplied by the Woods group[97][96] for the detection of putative epitopes. To this purpose, the *ab initio* predictor of interacting regions presented before is applied with the addition of glycans treatment. This approach is grounded on the hypothesis that the antigen binding site for Abs could be formed by a subnetwork of residues with weak energetic pairings with respect to the stability core of the protein: such segments should display frustrated interactions among each other and the rest of the protein. The obtaining of a proper docking of another protein is possible if the free state results in an higher free energy respect to the docked one through opportune interactions among the molecules[173][95]. Moreover, low energetic pairings with the remainder of the structure supply these segments with a large structural flexibility that may be necessary for the rearrangement of the interface during the binding event[177][58]. Another consequence of weak couplings is the possibility to mutate with minimum energetic cost in the absence of alteration in protein's conformation and stability.

These are peculiar traits of the antigenic epitopes.

In this work the used method is shown to be capable to detect sites that are comparable with the Abs binding sites discovered in latest structural immunology works, even in the presence of glycans. In addition other putative epitopes are predicted (still not investigated at the moment), these sites may be employed for the design of improved antigens, in both engineered peptides or extracted domains. In the end, the results obtained support the analysis about the molecular roles underlying the structural reorganization supporting the activity of the protein. As far as I know, this methodology is among the first that allows epitope detection considering glycans based just on the characterization of the energetics of the antigen taken alone. It is significant that the method works in the absence of experimental information about Ab recognition and epitope location of homologous proteins and does not require a specific parametrisation, such as structural or sequence features. The process is versatile and transferable to different systems.

3.3.2 Protein structure and the role of the glycan envelope

According to their structures, the viral fusion proteins can be classified in three different types[211]. The SARS-CoV-2 spike is a protein of the first class composed by three monomers of 1273 residues length. As described by Casalino et al.[29] it is possible to distinguish three topological parts of the protein: the head, the stalk and the cytoplasmic tail. In the head is located the subunit S1 that is formed by the N-terminal domain (NTD) and the receptor binding domain (RBD), the determinant of the recognition of the ACE2 through the receptor binding motif (RBM)[76]. To reach a proper binding with ACE2 the RBD can swing from a bent closed to an extended open position. These two alternative conformations are referred as “ up ” and “ down ”. The RBM is hidden to the solvent in the “ RBD down ” position and become exposed when the conformation switch to the “ RDB up ”. After the RBD, there is a furin cleavage site regarded as important to prepare the S protein for membrane fusion and cell infection[19]. This site separates the S1 from the S2 subunit. The S2 has a relevant participation in connecting and merging the host membrane with the viral one. In this subunit there is a second cleavage site that frees the fusion peptide (FP), a segment that induce membrane merging after its insertion in the cell[12]. After the FP are located the central helix (CH) and the connecting domain (CD). In the final residues there is the stalk part, that consist of the heptad repeat 2 (HR2) and the transmembrane domain (TM), and the cytoplasmic tail (CT).

Another relevant structural property of this protein is the dense coverage of glycosyl-moieties on the surface. This envelope is a common trait of fusion proteinz in viruses[52][38], and it is believed to have an active role in viral infection[84][207]. In this context the case of HIV-1 envelope spike (Env) is emblematic. The Env shows and extended area of glycosylation sites[82] and the carbohydrates are so compacted that they occupy a significant fraction of the whole Env molecular weight[65]. It is supposed that the widespread presence of glycosylated residues, inert from an immunological point of view, allows the pathogen to avoid the Abs and lymphocytes recognition[6][38][207]. This deceit that is particularly efficient for HIV-1 could be more vulnerable in the case of Coronaviruses S protein, in fact

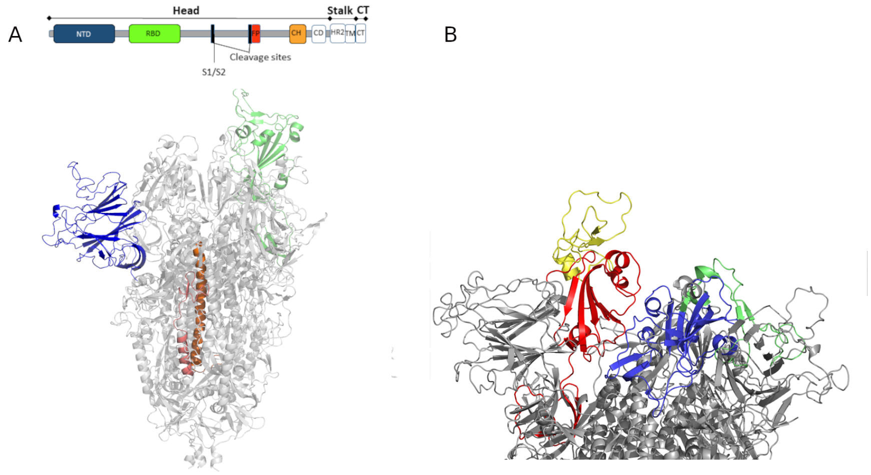


Figure 3.2: In figure A there is a scheme with the relevant parts of the full S protein with the three dimensional representation of the Head part. There are highlighted the NTD in blue, the RBD in green, the CH in orange and the FP in red. In B there is a focus on the protein S top with the “ RDB up ” coloured red and “ RBD down ” in blue, the structure show a different exposition for the RBM, respectively in yellow and in green.

these proteins exhibit a lower glycan coverage[84]. The SARS-CoV-2 S protein displays around 22 glycosylation sites[12] that were recently investigated by the Amaro group[29]. The Authors computed the SASA of the full-length protein comparing the system in the absence of glycans with a full glycosylated system. The results show that the SASA reduction due to the glycan shield is higher in the stalk compared to the head part, meaning that the latter presents more weakness in the coverage that can be helpful to be employed as a target. It is significant that the RBD in the down conformation has a greater SASA reduction than the upward position, suggesting an increased role of swing mechanism between the two states in hiding the RDB domain. In addition, it is found that two glycosylation sites (N165 and N234) are relevant for the stability of the “ RBD up ” conformation.

3.3.3 Implementation

The structures used are extracted from the trajectories of fully glycosylated SARS-CoV-2 S protein (PDB code 6VSB) provided by the Woods group[97], six distinct molecular dynamics runs of 400 ns for a total 2.4 μs. In the elaboration of the structures, we used the parameters originally chosen by Woods et al. in the simulations: the force field ff14SB[133] is employed for every residue save glycosylated asparagines, in that case the GLYCAM_06j is used[114]. In order to obtain representative conformations to use for the energetic calculations the trajectories are concatenated and aligned, then clustered referring to the root-mean-square deviation of Cα atoms in the RBD domain using the cpptraj tool of Ambertools17[30] . The selected algorithm is the hierarchical agglomerative[44] with threshold ϵ

equal to 0.5. In the three most numerous clusters the central structure is selected, solvent and ions were removed. The trimer displays two monomers in a “ RBD down ” and a single monomer in a “ RBD up ” conformations, only the monomer with the domain in the upward position is considered for the analysis. The three monomers chosen are successively minimized with the standard procedure for the tool sander of the AMBER package. The minimization of 200 step is performed with the generalized Born implicit solvent model in the version of Onufriev et al.[149] and the cut-off for both Lennard-Jones and Coulomb terms is set to 12 Å. At this step the counterions (implicit) concentration is 0.1 M and the solvent accessible surface area is evaluated with the LCPO algorithm[209]. In the end, the interaction matrix is computed, for every structure, with the MM-GBSA approach, keeping the same parameters of the previous step except the zeroing of counterions concentration and the changing of the solvent accessible algorithm with the ICOSA. The interaction matrix is then processed with the MLCE method for multidomain proteins described before.

3.3.4 Results and Discussion

The MLCE method computes the interactions among neighbour residues and for the selections of the weakest residue pairs the couplings are ordered according to the strength, from the weakest to the strongest. Then the regions with a poor energetic coupling are detected, selecting only the pairs under a cut-off expressed in percentile. In order to obtain a more complete scenario, two different cut-offs are chosen: one set of pairing is extracted from the 15th percentile and another set from the 5th. The variation of this “ margin of softness ” will change the extent of the detected region, helping in the individuation of the most relevant parts. As a clarification, it is important to stress that variations of the S protein structures can reflect in distinct outcomes, since the matrix of energetic couplings relies on the physical conformation of the molecule. The predicted areas of poorly coupled residues are considered for the determination of immunological relevant domains and epitopes. Starting with the broader margin (cut-off of 15%) the whole monomer of S protein is subdivided in putative immunoreactive domains[70]. The aim is to discover the regions that have the potential to interact with Abs but are usually concealed from recognition when situated inside the whole structure. Neutralizing epitopes with great reactivity may be displayed in just few transient conformers that X-ray and cryo-EM models are unable to highlight. The exposure to Abs interaction of these elusive areas via a single separated domain could bring new opportunities in the design of new antigens[70]. The use of more restrictive cut-off of 5% brings the attention to those little fragments that are central to establish interactions with Abs, information that is helpful for the design of new antigenic peptidomimetics. For this scope the selected fragment needs to have a minimal length of six residues. The selection with higher cut-off detect a wide group of frustrated energetic couplings in the RBD domain, indicating that this domain is the part, of the “ RBD up ” monomer used, more prone to interact with an Ab (see Figure 3.3). The regions found are obtained through consensus among the three structures extracted from molecular dynamics and, noteworthy, they have a good superposition with epitopes known to interact with nanobodies

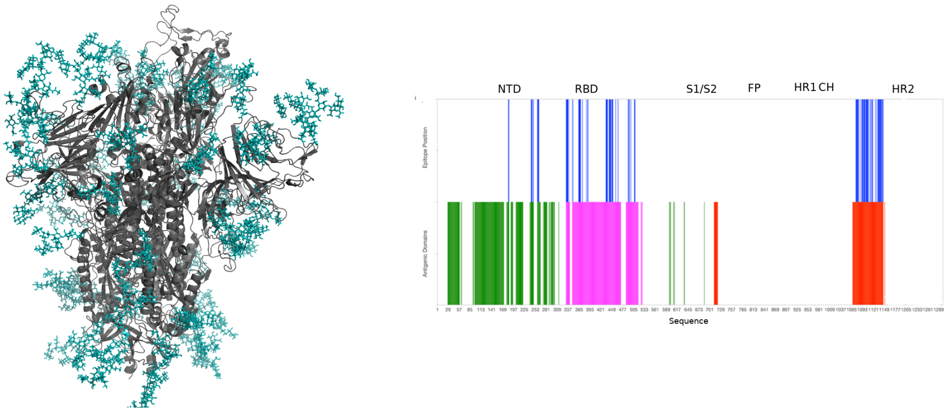


Figure 3.3: In this figure is shown the three dimensional structure of the protein S head with the glycan envelope coloured blue. On the right there is the projection on the protein sequence of the antigenic domains and epitopes detected. In above line there are shown the epitopes found with the narrow margin (5th percentile), while in the lower line there are the antigenic domains detected with the broader margin (15th percentile), the three different colours refers to different domains.

or Abs (as it is shown by latest X-ray and cryo-EM models). An interesting example is the good match with the binding site of the monoclonal Ab CR3022 [72], which dock in an elusive epitope that is visible for the recognition after relevant conformational reorganizations[29], see Figure 3.4.

Another antigenic domain is identified in a region that covers great part of the N-terminal domain. In a paper by Chi et al.[81] is reported that the epitope recognised by novel Ab 4A8 is in this domain. A third area with low energetic couplings is located in the terminal zone of the NTD domain. This part is in the neighbourhood of the density for fragments of antigen binding (Fab) of COV57, a newly studied Ab with a neutralization property that is alternative to the action of RBD-binding Abs, as it is verified with latest cryo-EM study of Abs-S protein interaction[62]. A work by Zhou et al.[222] shows that Ab 7D10 for MERS interacts in the same region. In addition the local coupling method detect a putative area with good reactivity into the S2 subunit in the CD domain. In this domain, the region of binding of 1A9 is detected, an interesting Ab able to cross-react with the spike proteins of coronavirus in different species, including human and bats. In this antigenic domain is also contained a groups of glycan. This could be linked to a novel discovery where an epitope that includes carbohydrates is positioned around this region and is bonded by an Ab that recognises carbohydrates HIV-1 bnAb 2G12[74] (see Figure 3.4).

The definition of domains with poorly coupled groups of residues can have other functional involvement in addition to epitope detection. In fact the part of the protein that are not necessary for structural stabilization can undergo wide

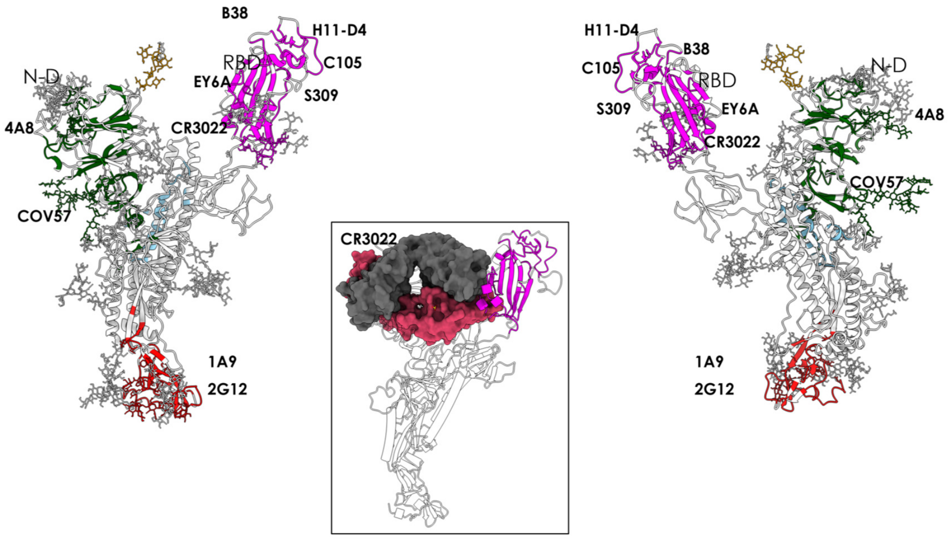


Figure 3.4: Predicted interactions zones and known epitopes. The groups of residues of the detected antigenic domains are highlighted with colours, green for the NTD, magenta in the case of RBD and red in the end of the S protein head. The location where is experimentally determined the interaction with antibodies is labelled with the respective antibody name, the structrure is represented in two specular images. In the centre is shown the interaction site of antibody CR3022 in correspondence with the antigenic domain coloured magenta.

conformational rearrangements that support the biological activity. The borders of the NTD antigenic domain (Figure 3.4, green colour) are located in vicinity to the furin-cleavage site (indicated by the motif RRAR), necessary to prime the fusion mechanism of S protein. Therefore, this wide area of weak contacts in the NTD can perform a movement in order to support the cut of this motif, inducing the separation of the S1 subunit from the S2[204][190]. Moreover, there is a β-sheet in the border of the CD domain that is in vicinity of the fusion peptide. It could be rational to presume that the presence of a cluster of frustrated couplings in the terminal of the S protein head could assist the structural changes necessary to give to the FP a better exposure, thus helping the peptide insertion into the cell membrane[190].

On the whole, these results sustain the feasibility of the MLCE method to detect parts of the protein that can display an interesting antigenic activity, since they are more liable to be involved in the humoral response respect to the rest of the protein. There are observed putative antigenic domains different from the RBD that can be bound for the neutralization the the virus. This aspect is particularly relevant considering that the RBD is a target of Abs without neutralization activity (for example, CR3022 [72]). Therefore, it is proposed as a feasible therapeutic option the employment of a mixture of antibodies that bind in distinct areas of S protein [81]. This consideration may find a basis in the cure of others epidemics where a cocktail of Abs was employed. In this scenario, Regeneron Pharmaceuticals is chasing a cocktail based treatment for COVID-19 that is being tested in clinical trials[57].

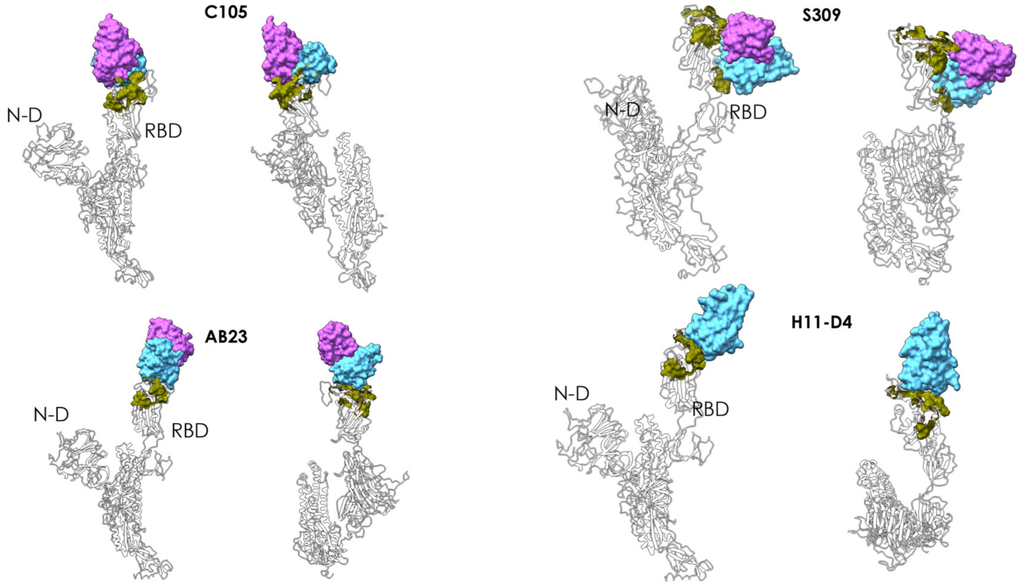


Figure 3.5: Putative epitopes detected in the RBD domain with the lower cut-off (5%) are set against experimentally determined Ab docking. The Abs binding (C105, S309, AB23 and nanobody H11-D4) are shown with a surface representation (coloured blue or blue/pink) and the monomer used in the analyses is showed in grey. The green area highlight the position of the predicted epitope.

When the selection of the areas with poor couplings is made with a more restrictive margin (cut-off of 5%) the attention goes to the fragments of S protein that establish interactions with Abs, i.e. conformational epitopes. Remarkably, one of the epitope detected, made by residues (348)A-(352)A-(375)S-(434)IAWNS(438)-(442)-DSKVGG(447)-(449)YNYL(452)-(459)S-(465)E-(491)-PLQS(494)-(496)Q-

(507)PYR(509), includes zones of the protein that make interactions with Abs C105 (PDB code 6XCN)[62], S309 (PDB codes 6WPT and 6WPS)[63], AB23 (PDB code 7BYR)[83] and with the nanobody H11-D4 (PDB code 6Z43) and with another engineered nanostructure (PDB code 7C8V), see Figure 3.5.

Moreover, it is interesting to notice that also an epitope with glycans is detected and it is observed to be included in the interaction site of the Ab S309[63]: the predicted fragments consists of residues (332)ITNLC(336)-(361)C with the addition in position N334 of the fucosylated N-glycan chitobiose core (Manβ1-4GlcNAcβ1-4[Fucα1-6]GlcNAcβ-Asn)[114]. The detected area is noteworthy in the proximity of the RBM. Another individuated region, with residues (365)YSVLYN(370)-(384)PTKLN(388), overlap with a significant fraction of the evasive epitope delineated by Wilson et al.[72] (see Figure 3.6), that is the binding site of the Ab EY6A (PDB code 6ZDH). It is stressed, one more time, that the detection of those reactive areas is just conducted with the use of plain structural data produced from a monomer of the glycosylated spike protein conformation that is extracted from MD trajectories.

The narrow margin in the epitope area definition (cut-off 5%) permits to detect an antigenic fragment, that includes the residues (184)GN(185)-(242)- LAL(244)-(246)R-(248)Y-(258)WTAGA(262), in the NTD domain. Inside this epitope are found sites R246 and W258, both known as crucial factors for the interaction among NTD and the Ab 4A8[81], see Figure 3.7. Lastly, there is the prediction of immunogenic segments in the zone that goes through residue 1076 to 1146, where is located the segment 1111-1130 corresponding to the experimental epitope of the mAb 1A9[85]. To be more precise the identified patch is: (1076)TTAPAICH(1083)-(1087)A-(1092)-REG(1094)-(1096)FVSNGHWFVTQR N(1108)(1112)P-(1114)I-(1116)T-(1118)DN(1119)-(1126)C-(1129)V-(1132)IVN NTVYDPLQPELD(1146).

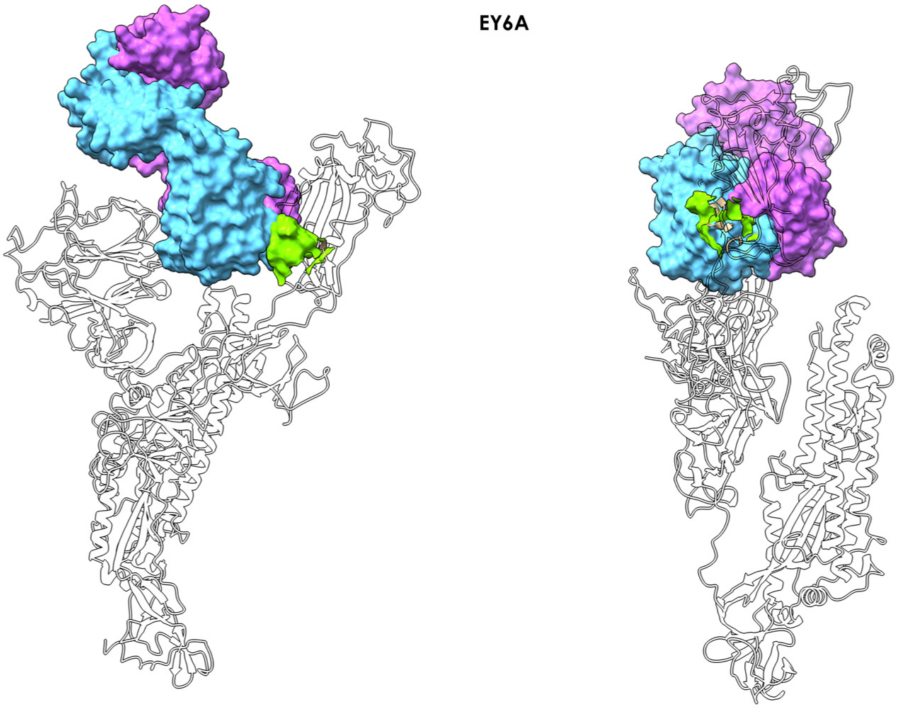


Figure 3.6: Here is showed the Ab EY6A in complex to spike protein monomer. The Ab (PDB code 6ZDH) is in a surface representation coloured blue/pink and binds to the RBD of the spike monomer, coloured white. The putative epitope is coloured green, the binding is shown in two distinct views, highlighting the goodness of the overlap.

Overall, the applied method is capable to individuate putative antigenic domains or epitopes in the S monomer using only structural and energetic data: in the comparisons the technique predicts accurately from 20 to 80% of the residues involved in mAbs recognition extracted from X-ray complexes. Given that various (mixtures of) reactive residues of the antigen could be target of distinct Abs during a polyclonal proliferation (like what happens in the human organism), the individuation of just a minimum segment provided with immunological activity may supply the necessary help to the design of new compounds able to elicit neutralizing antibodies.

In this context, the sequences detected with the narrow margin (5th percentile)

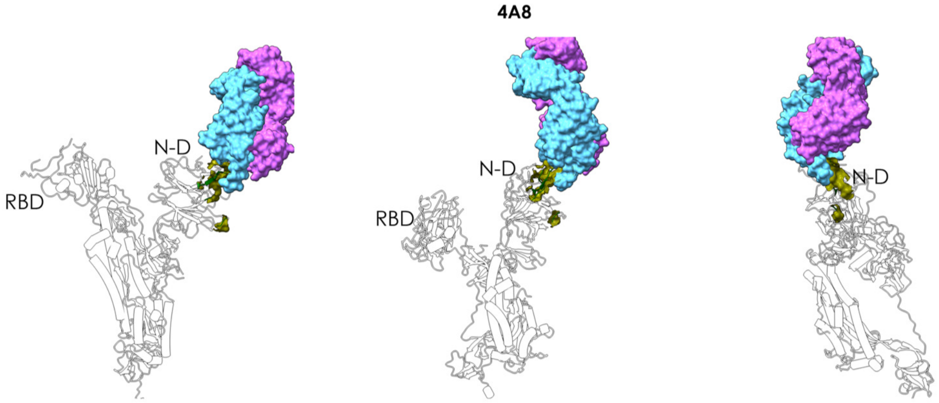


Figure 3.7: Here is showed the Ab 4A8 in complex with the spike protein monomer. The Ab, depicted in surface representation, is coloured blue/pink and the spike monomer white. The detected epitope is coloured green. The Figure shows how the Ab interacts with the NTD domain and the overlap with the green region suggests the formation of contacts.

may be employed for the creation of engineered epitopes in a peptide format. The production of epitopes of this kind would require the design of structured peptidomimetics of the original antigenic segment, including optimisation of the peptide stability via the use of non-natural amino acids. In this way it is possible to recreate the structural circumstances that cause the humoral immune response. The peptides detected could be exploited as models for novel vaccines or probes for serologic diagnostic tools (such as ELISA or microarray tests) for the individuation of Abs that flow in the blood which are secreted after the exposure to SARS-CoV-2, in particular those that can operate a proper neutralization. In addition, this probes may be a convenient instrument for the individuation of novel mAbs and testing compounds for drug design. A relevant characteristic of the used MLCE method is that the whole spike protein carbohydrates envelope is considered for the detection of active epitope. The glycans do not display an uniform function but they tend to have a distinct behaviour, depending on the case. On the protein there are groups of glycans that exhibit a strong energetic interactions with the rest of the protein, these segments are not predicted as binding sites but contributes to the global stability of the S protein. On the other hand, the decomposition method is able to detect groups of carbohydrate chains that are weakly coupled to the structure core and are labelled as putative binding sites for Abs (or fragments of them), these areas suggest probable imperfections in the carbohydrates shield that may be target of new small molecules or used as a starting point for new vaccines. The parts of this envelope that are in the first group present a dual role, to hide the pathogen from immune recognition, therefore increasing its dangerousness, and to supply an additional mechanical activity. An example is represented by a couple of glycans that is showed in Figure 3.8. One is the complete carbohydrate segment linked in N234 (see Figure 3.8 A), that is reported by the Amaro's group to be relevant in the stabilization of the “ RBD up ” conformation[29]. Site-directed mutation of this site with alanine produces a relevant population shift at the expense of the “ RBD up ” state[100]. The other case is the fragment of carbohydrate chain of the residue N165 that is coloured in yellow in Figure 3.8 B. The glycan in the site N165 has an active role in regulation conformational transitions, principally for the “ RBD up ” state[100]. Interestingly this fragment of the glycan is not predicted as a binding site and therefore it acts to hide the virus from Abs in the area close to N165, but the rest of the residue chain (in orange in the Figure 3.8 B) is detected to have probably immunogenic activity, in fact it has weak energetic coupling to the monomer.

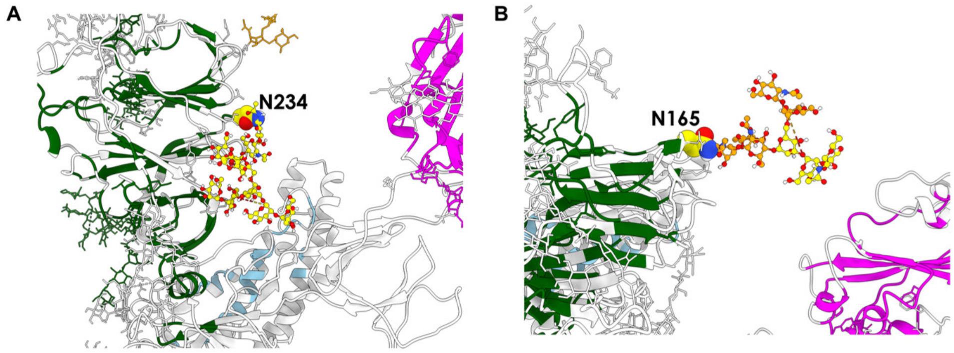


Figure 3.8: Carbohydrates chains with distinct activity. In A there is the glycan of the residue N234, that is individuated as a stability agent for the protein structure. In B is shown the glycan linked to N165, here the prediction finds a dual function of the chain, a part works for stability (yellow) and another acts as immunogenic agent (orange).

It is remarkable that the used method is able to predict Ab-binding sites close to no immunogenic glycans despite (or in some aspects thanks to) its simple theoretical basis relying on the individuation of frustrated contacts in the protein. According to this result it is plausible to think that the shielding activity of these specific oligosaccharides could be bypassed and made ineffective through the disclosure of the subjacent peptide structure. Moreover, information about the glycans that are recognised into the epitopes can be employed for the study of engineered antigens in the form of glycosylated peptides. This last concept is significant: in fact novel synthetized compounds capable of imitating the specific features of original antigens (behaving as a proper substitute) introduce attractive chance to increase the properties of the immune reagents designed considering the simpler purification and manageability, reduced expenses and optimized stability in different states. In addition, the employment of this kind of compounds can decrease the possibility of cross-reactivity with other antigens present, a common situation when recombinant proteins are used. An expressed viral protein (and every protein grounded diagnostic tool) usually needs more limiting environmental parameters (like temperature) for the warehousing, carriage and supervision for the maintenance of the antigens in its characteristic functional fold. This line of reasoning still holds for vaccines made by inert infectious agents.

In conclusion, these results show that the prediction of immunoreactive epitopes of the SARS-CoV-2 spike protein is possible with the analysis of unbiased structural data. The confirmation with experimental complexes expands the comprehension of the molecular basis for interactions, which can be transferred into novel vaccine candidates or testing tools. Moreover, the exposed method gives possible functional involvement as it is suggested from the individuation of the domains and zones that are significant for correct behaviour of the biological system. For this reason the technique seems adequate for the detection of potential modulation of the protein activity in presence of mutations caused by future viral spread and host adaptation. In the end, the opportunity to cluster this multifaceted molecular machinery into functional parts may be used for the definition of a coarse-grained system able to simulate the protein for an higher timespan. The development of structure-driven *in silico* methods of this type may expand the applicability of unbiased analysis and molecular simulations. At a practical level, the creation of libraries of detected immunoreactive peptides (also glycosylated) would increase the efficiency of the screening of candidates for vaccine design.

Part II

Consequences of Binding

Chapter 4

Computational study of allosteric proteins

Allosteric regulation is the modulation of protein functions with the binding of a ligand in a region distinct from the active site. The reasearch of the mechanism that can illustrate the allosteric effect has required scientists' dedication in the past 50 years[32]. Since the description of the basic two-state system by Leff[126], two main approaches were adopted: a thermodynamic approach, consisting in the mathematical modelling of the reactions using the concentration of the states as variables[54][73], and a structural approach, that is based on the analysis of residues connection and comunication among different areas of the protein[45]. These complementary criteria can be joined under a free energy landscape model, that is capable to merge the analysis of structural induced modulation with quantification of the population shift using shared descriptors[198], in the same way of the unification of conformational selection and induced fit in protein binding models.

During my PhD I was involved in the study of two allosteric systems, adopting structural approaches. Firstly I developed a method for classification of the activity of allosteric ligands of Hsp90 protein, inserting learning algorithms in a docking protocol. The secondary task is the study of allosteric effect in the αvβ6 integrin with extended molecular dynamic simulations.

4.1 An allosteric model: Human Hsp90

The Heat Shock Protein 90 is a molecular chaperon essential for the cell life cycle[175]. This molecule is widespead in the organism and constitutes around 1–2 % of the dry cell mass, percentage that grows to 4–6 % in stressed cells[122][89]. Recently it is showed that Hsp90 become a nucleating site for the costruction of robust multiprotein complexes that improve cell survival, presenting tumor-specific

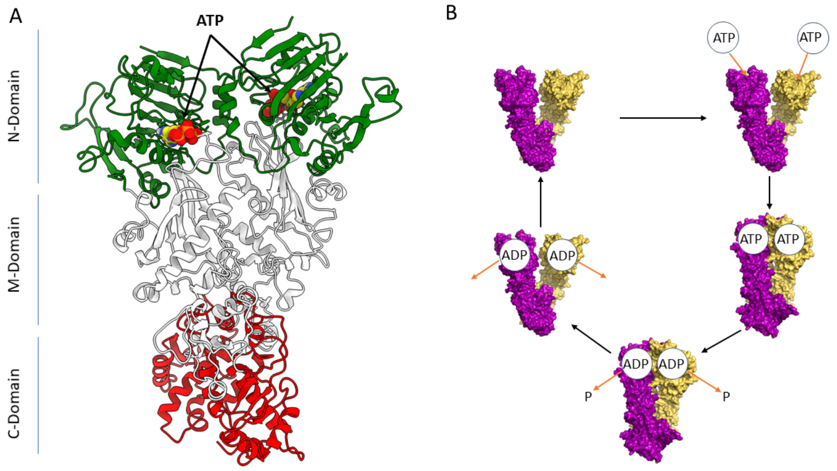


Figure 4.1: In figure A there is a representation of the Hsp90 dimer, the distinct domains are highlighted: the NTD in green, MD in white and CTD in red. On the right side, Figure B depicts a schematic view of the chaperon cycle.

patterns that are missing in cell in usual condition[59].

The Hsp90 family can be distinguished in three isoforms. Cytosolic Hsp90 that is involved in signal transduction, cell preservation and growth, has clients like hormone receptors and kinases. The mitocondrial isoform is TRAP1 responsible of the folding of mitocondrial proteins necessary for homeostasis and respiratory cycles. The last isoform, GRP94, is present in the endoplasmic reticulum serving proteins like IgGs and integrins. The protein has a homodimeric structure with monomers of around 700 residue lenght, subdivided in three principal domains: the N-terminal domain (NTD), the middle-domain (MD) and the C-terminal domain (CTD). The two monomers form a stable dimerization interface in the CTD forming a V-shaped complex.

For the function of its chaperone activity the homodimer undergoes to a conformational cycle[160], moving from an open state to closed one with the NTDs binding to each other's and then back to the open state. It is demonstrated that to obtain the reaction cycle is required the hydrolysis of ATP in a binding pocket situated in the NTD[104]. For this reason the complete cycle is depicted in five steps: firstly ATP molecules dock in the active site of NTD, this event induces the dimerization of the Nlobes of the monomers, after the dimerization takes place the hydrolysis of ATP molecules with the consequence of conformation opening and, finally, nucleotide release.

This kind of sophisticated machinery needs the formation of structurated signaling pathways between different parts of the protein. Prodromou et al. showed that the deletion of the CTD inhibits the ATPasic activity of the NTD, highlighting an active role of Clobe in the protein modulation and providing a proof of allosteric regulation[161]. Further information are provided by the release of a Cryo-EM structure[201] of Hsp90 in complex with cochaperon Cdc37 and a client, Cdk4 kinase. The complex unveils Hsp90-client interactions in the region between CTDs interface and MD. Studies of the dimer in different nucleotide concentration bing additional insight on the reaction cycle. It is found that the conformation with a double ATP molecule docked is poorly represented in solution[164] and, also, it is obtained a mixed (ATP-ADP) closed conformation, showing that a double ATP state is not necessary for NTD dimerization[55]. This negative cooperation among the NTD interfaces is an additional proof of the comunication between CTD.

4.2 Building a learning classificator for ligand activity

4.2.1 Introduction

The improved knowledge of gene organization coupled with the advances in gene editing and structural analysis methods can potentially start a whole new era in drug discovery[199][212]. In particular, improved target identification can shed light on biomolecules whose perturbation via small molecule binding results in a functional response, transforming a disease phenotype into a normal one. The extraordinary complexity of biochemical networks in healthy and disease conditions[151][106] and the costs associated to drug discovery are however hampering the advent of this new era of therapeutics, as shown by the relatively low numbers of new drugs approved in the last few years[60][75]. Most drug discovery efforts aim at targeting the active sites of enzymes or the orthosteric sites of regulatory proteins. Because of the evolutionary and structural conservation of such sites across the proteome, issues related to selectivity, off-target effects and development of drug resistance have started to appear.

In this context, allosteric ligands have recently emerged as a viable complement or alternative to active- site directed molecules, with novel potential as drug candidates or chemical tools[78][177][88]. Allosteric ligands bind to sites that are generally distinct and distal from the classic orthosteric ones. In doing so, they can perturb the target not only by inhibition but also through modulation or activation of specific functions. This represents an advantage in terms of fundamental and applicative perspectives. In fundamental research, chemical modulators (effectors) can be used to direct signaling pathways and whole cells towards desired functional states, representing important tools for understanding the roles of specific biomolecules in complex biochemical networks[189][223]. In biomedical applications, since they target sites that are generally less evolutionarily conserved, allosteric ligands can be highly selective, even among different members of the same protein family[171], providing new opportunities for therapeutic discovery.

To date, most (non-natural) allosteric ligands/drugs have been discovered using high-throughput screening. The ever growing amount of sequences and structural information combined with the increases in computing power and the improvement of predictive algorithms are starting to facilitate the discovery of allosteric modulators, but major challenges remain to develop approaches focused on rational drug design.

Computational approaches to the problem have focused on variations on the theme of molecular docking. Binding affinities predicted by docking simulations are routinely used in virtual screening to estimate relative ligand rankings and to inform further steps in lead identification[129][20]. Efficient screening of large libraries of compounds is achieved by the use of approximate scoring functions and simplified strategies for conformational sampling[68]. Typically a static model of the target structure is used. However, recently the influence of protein dynamics on the recognition process has been more accurately modelled using ensemble strategies[25][148][41][79]. These strategies involve the docking of a molecular ligand libraries over an ensemble of selected geometries of the protein, creating a more realistic representation of the ligand bound to the different expected conformations of the target. The use of an ensemble of conformations reduces the dependence of the docking results on the target structure[28]. Ensembles can be extracted from unbiased molecular simulations of apo structures[7] and more often by sampling of protein conformations from holo structures containing first-generation ligands[33]. Under the assumption of conformational selection, a set of different ensembles representing different binding states would have selective preferential binding for different ligands. Based on this hypothesis, previous studies have used ‘a panel of ensembles’ for virtual screening[192], whereby a vector of binding affinities against the panel is used to generate a specific fingerprint for each ligand.

This type of data has high dimensionality both in the chemical and conformational space and is best suited for analysis using Machine Learning (ML) methods, that have been increasingly adopted in drug discovery studies. Indeed, they contributed to improvement of performance in virtual screening studies[115][18][136] and they have been effectively used in the enhancement of structural based virtual screening and scoring[145][61]. ML methods are mostly data-driven and their performance is often dependent on the size and quality of the dataset. To this end, they may present limited transferability and care is required in reporting results and scope of applicability.

The combination of ML with molecular simulations can dramatically advance the process of selection of allosteric ligands with a desired impact on the function of the target. Indeed, a major limitation in docking simulations is the lack of information on the functional consequences of the allosteric binding event. While relative binding affinities and geometries can be reproduced close to experimental accuracy, there is no predictive score to discriminate inhibitors from activators, agonists from antagonists or partial agonists[203]. Experimental assays typically report on the orthosteric function, in most cases by direct measurement of a relevant biochemical parameter that involves the active/orthosteric site. This may not necessarily reflect the affinity of binding at the allosteric site [159][165][67]. In most cases, binding is only one aspect of an intricate interplay of structural and dynamic factors that emerge from the cross-talk between the allosteric ligand and the protein and define functional responses. As a consequence, the derivation of structure–activity relationships (SARs) for allosteric ligands is typically much more complex than for orthosteric ones. This unmet challenge calls for new approaches that integrate information on binding, conformational dynamics, and biological activity because the desired readout of the binding event is a change of functional state in the protein, that is not directly or easily modelled by single docking calculations.

Here, to progress along this fascinating avenue, the potential of ML models trained on molecular simulations is explored in order to predict the functional effect of allosteric ligands on proteins. Allosteric ligands can either activate or inhibit protein function. As a test case, the attention is focused on the difficult case represented by the Hsp90 chaperone system, a molecular machinery essential for cell development and maintenance, that works by facilitating the folding of a broad spectrum of clients [92][179][163][175]. Proteins of the Hsp90 family (Hsp90 in the cytosol, Grp94 in the ER and Trap1 in mitochondria) are homodimers with two chains consisting of three globular domains, the N-terminal (NTD), Middle and C-terminal (CTD). The functions of the chaperone are regulated by ATP hydrolysis in the N-term domain, where ATP processing is coupled to Hsp90 conformational reorganization and consequent client remodelling. Early work by Neckers' group and recent computational studies reported an allosteric site at the boundary between the M- and C-terminal domains that modulates ATP-related functionalities[88][134]. The discovery of this allosteric site facilitated the development of different series of allosteric ligands that are able to perturb Hsp90 mechanisms, by either inhibition or activation of ATP processing. Kinetic and biochemical data indicated that the functional effects of the ligands are critically coupled to their influence on the conformational dynamics of the protein.

In this work, we ask whether it is possible to develop a reliable predictor of activation/inhibition for Hsp90 allosteric ligands. Model training is driven by ensemble-based structural, dynamic and energetic characterization of allosteric binding.

4.2.2 Computational Implementation

Molecular Dynamics simulation and analysis

The protein structure coordinates (PDB ID: 2CG9) for yeast Hsp90 were downloaded from the Protein Databank (`https://www.rcsb.org/`). Initial pose for ligand docking were derived from previously published models[79][144][142]. MD simulations were run with Gromacs 2018.2[2] with Amber03 force field[22]. The protein-ligand complex was solvated with TIP3P water model in a dodecahedral box with minimal distance from the solute of 14 Å, counterions were added to neutralize the system. After a minimization the molecules were equilibrated for 100ps in NVT ensemble and successively in NPT ensemble for 100ps. The simulations were conducted at constant temperature of 300K and at constant pressure of 1 bar, with a coupling time of 2 ps. The electrostatic term was described by using the particle mesh Ewald algorithm[42], the LINCS algorithm[103][87][102] was used to constrain all bond lengths. Available ATP parameters for amber force field[137] were used and ligands topologies were generated using AnteChamber software with AM1-BCC charge model. For each ligand-protein complex a 400ns of simulation was run. Cluster analysis was performed on a combined metatrajectory of all simulations with the representative ligands. Rigid roto-translation fitting and RMSD calculations were made on alpha carbon atoms of secondary structure segments extracted with VMD software[107]. Clustering was performed with Gromos algorithm[43] using a cut-off between 2 and 2.5 Å.

Molecular Docking and fingerprint analysis

All systems were prepared using the Schrodinger Suite: bond orders and atomic charges were assigned and the hydrogens were added, protonation states were evaluated on acid and basic enzymes and hydrogen bonds were optimized. The protein was then minimized with a cutoff of 0.3 respect to starting configuration. The Glide[77] software was used for molecular docking: the putative binding site was mapped on a grid with dimensions of 48 Å, enclosing box, and 28 Å, inner box. Calculation with fixed receptor and flexible ligand were made with standard precision (SP) modality with OPLS3e Force Field. No additional changes to default settings were made. Fingerprint similarities were computed with the Canvas program of the Schrodinger Suite, typing scheme is atom distinguished by functional type with no scaling in 32 bit.

Supervised and Unsupervised Learning

In house scripts for cluster analysis and supervised learning prediction were developed in Python using scikit-learn functions (`https://scikit-learn.org/stable/index.html`). Source code is available at `https://github.com/alepandini/LIGXF`.

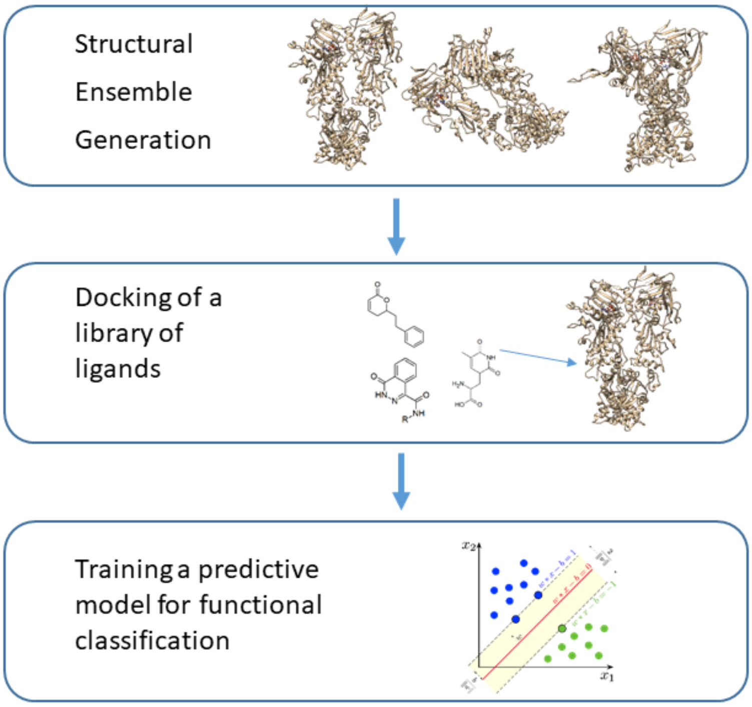


Figure 4.2: Here there is a scheme of the procedure implemented for this work.

4.2.3 Results and Discussion

The approach used to classify allosteric ligands as activators or inhibitors of ATP hydrolysis in Hsp90 entails three steps (see Figure 4.2). First, a panel of structural ensembles is generated by cluster analysis of conformations from molecular dynamics simulations of representative holo structures, in which Hsp90 is bound to ATP in the N-terminal domain and to an allosteric effector in the allosteric site. Then a library of allosteric compounds is docked against the Hsp90 structural panel. Finally, a predictive model for functional classification of the allosteric ligands is trained keeping into account the structural, dynamic and energetic properties of the resulting complexes. From the literature, 133 compounds with known activity against Hsp90 were collected, comprising 49 inhibitors and 84 activators (for the complete list see Appendix C). This dataset was used to train and test the predictive model. The protein conformational ensembles for docking were generated by atomistic molecular dynamics simulations in explicit water of Hsp90 in complex with three different ligands: one activator (CC26) and two inhibitors (ND2 and Novobiocin) (see Figure 4.3).

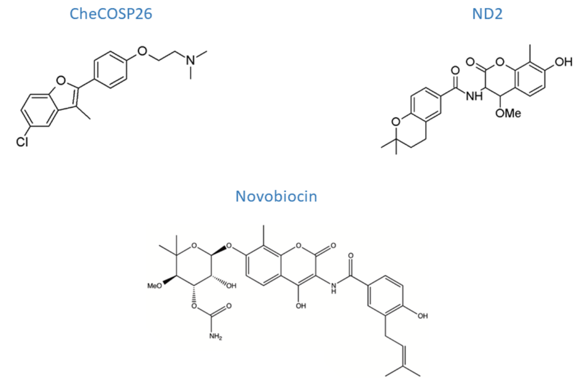


Figure 4.3: Here is showed the 2D structure of the ligands used in the simulations.

To keep the generation of the structural ensembles independent from the dataset used for training, these ligands were not included in the training and test datasets. Each replica of molecular dynamics was run for 400ns saving structures every 100ps and the resulting trajectories were combined into a single metatrajectory. The panel of structural ensembles for docking was built to approximate the most relevant states in Hsp90 functionally oriented dynamics. To this end, geometrical cluster analysis of the metatrajectory was repeated using four different reference frameworks: the backbone atoms of N-terminal domain (Clust-N); the backbone atoms of Middle domain (Clust-M); the backbone atoms of N-terminal and Middle domains (Clust-NM) and the backbone atoms of Middle and C-terminal domains (Clust-MC). In addition to these domain-based frameworks, a cluster analysis of the allosteric site was performed, where the ligand binding site was defined as the ensemble of residues that are within 1nm of any bound allosteric ligand in at least 75% of all visited structures collected in the metatrajectory.

Next, the three most representative structures from each of the four domain-based ensembles were selected as a target for docking experiments. In addition, the two main representative structures resulting from the allosteric-site based clustering were added. Two structures were enough to recapitulate more than 95% of the structural variability observed in the pocket. Cluster analysis of the molecular dynamics metatrajectory yielded 14 representative protein structures for the following step of docking. This collection was generated to capture the propensity of Hsp90 to populate conformations potentially endowed with different functional properties. After docking the ligand library to each of the selected representative structures, three measures were calculated for every resulting complex: the docking score of the best pose for every representative structure, the root mean square (RMS) of the docking score for the 10 best poses and, finally, the RMSD on the atomic positions of the first 10 poses, reporting on structural adaptation within the pocket. A total of 42 features was thus used for ML prediction.

The underlining hypothesis of this study is that features describing the docking results of a ligand against a panel of distinct conformational ensembles can be used as 'dynamic fingerprints' of its functional effect on the protein. This hypothesis were tested under three assumptions: 1) the separation of activators and inhibitors cannot be directly detected in the feature space by cluster analysis; 2) the separation of activators and inhibitors requires modelling a complex relationship by supervised learning; 3) the separation cannot be trivially obtained by use of small molecule fingerprints in the absence of information on the protein structure and dynamics.

None of the features described above can independently be used as a classifier and directly separate inhibitors from activators. This is evident from the distribution of values for every single feature against the two known ligand classes: in all the cases the pair of per-class distributions overlap (see boxplot in Appendix D). This suggests that a model based on the combination of these features is required to discriminate between the two classes. The first step is to test if the separation of the two groups of ligands can be directly detected with an unsupervised learning approach.

To this end cluster analysis was performed. Two different algorithms were used: k-means and agglomerative hierarchical clustering with target cluster numbers ranging from 2 to 4. The ability to correctly separate ligand classes in the clusters was estimated by cluster purity, that has values between 0 (when the class labels are completely mixed in the clusters) and 1 (clusters composed by only one class). Both algorithms have similar purity values, in particular when 2 clusters are considered the purity is low (0.66 for K-means and 0.69 for hierarchical) and with more clusters the purity raises, remaining below 0.80 (for 4 clusters: k-means have 0.78 of purity and hierarchical have 0.79). The increased purity is due to the reduced size of clusters that helps adapt to the class separation. Yet, the value in the case of 2 clusters reveals that is difficult to detect a segmentation of the compounds in the functional classes directly by cluster analysis. This suggest that

Measure	LR	SVM	RF
Balanced Accuracy	0.88	0.89	0.74
Precision	0.92	0.96	0.81
Recall	0.88	0.85	0.85

Table 4.1: In the Table are reported the performances of the three models tested.

it is not possible to automatically partition the space of the data to identify inhibitors and activators. A model trained on properties from the different binding conformations is therefore needed.
In this framework, a classification model was built using supervised learning. Three widely used algorithms were compared: Logistic Regression (LR) as a baseline, Support Vector Machine (SVM) and Random Forest (RF). The performances of the three methods were compared after training and test using the holdout method, where the dataset is randomly split in training set and test set with the proportion of 70% and 30% respectively. The performance in prediction is reported in Table 4.1.

LR and SVM show similar performances while RF has poorer performance. Nevertheless, all three methods show a better classification power compared to the cluster segmentation. 10-fold cross-validation without shuffling was performed to exclude any bias due to the simple holdout split and to further compare the methods. This approach also highlighted possible variability across the datasets and facilitated interpreting the performance with more insight on the chemical features of the molecules (see below). SVM shows the best performance with

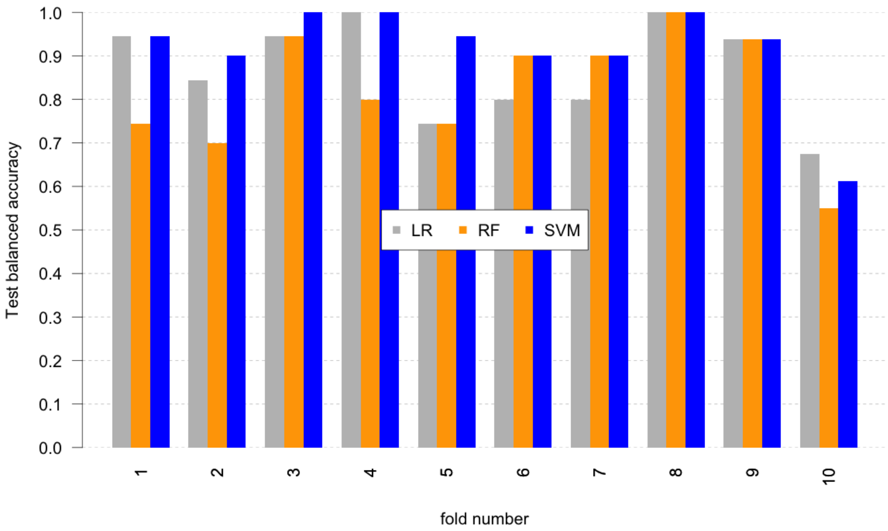


Figure 4.4: Here is the performances of the 10-fold crossvalidation for all the models. The value of balanced accuracy for every fold is noted, the values for Logistic Regression are in grey, Random Forest are in orange and SVM in blue.

an average balanced accuracy of 0.90, compared to 0.87 (LR) and 0.79 (RF). In Figure 4.4 per-fold balanced accuracy is reported. Only for one fold, values are below 0.8. The results show consistency in performance by SVM across the set. Finally, the possible dominance of one type of ensemble features (docking score, rmsd, rms) in the prediction was assessed by selectively excluding each feature in turn and repeated cross-validation. In each case the variation in the average performance was not statically significant, z test score below 1 (see Figure 4.5), therefore no feature was detected as dominant.

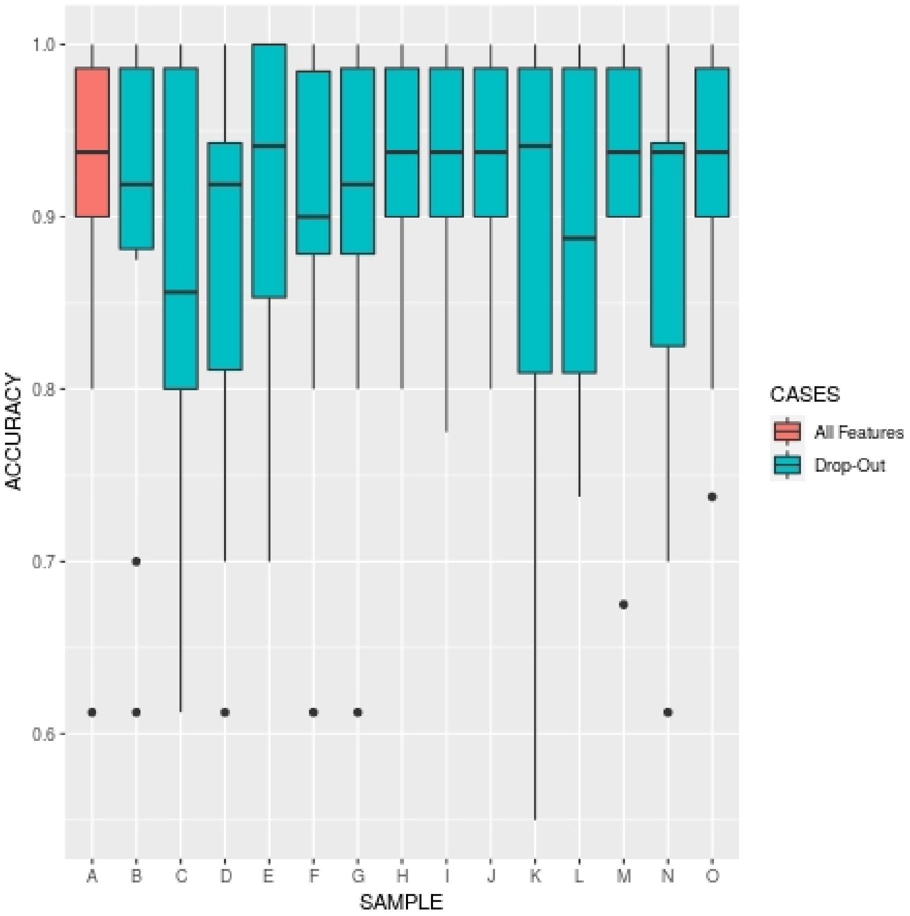


Figure 4.5: Boxplot of the distributions of the accuracy values for SVM. The A sample represents the SVM trained on all the variables, in the other samples the measure are obtained with the drop-out of one of the variables. All the samples display an overlap on the distribution of values.

The classification model trained on docking against the panel of representative conformations does not directly account of chemometrics properties of the ligands. In the context of compound selection, it is interesting to compare the classification model with a direct analysis of the chemical properties of the compounds. This is to assess if correct classification can be obtained by small molecule fingerprints in the absence of information on protein structure or dynamics. Our dataset comprises molecules representative of different chemotypes (see Appendix C). It may be possible to qualitatively cluster these molecules with respect to shared scaffolds: in our case, this results in eight different groups (Figure 4.6).

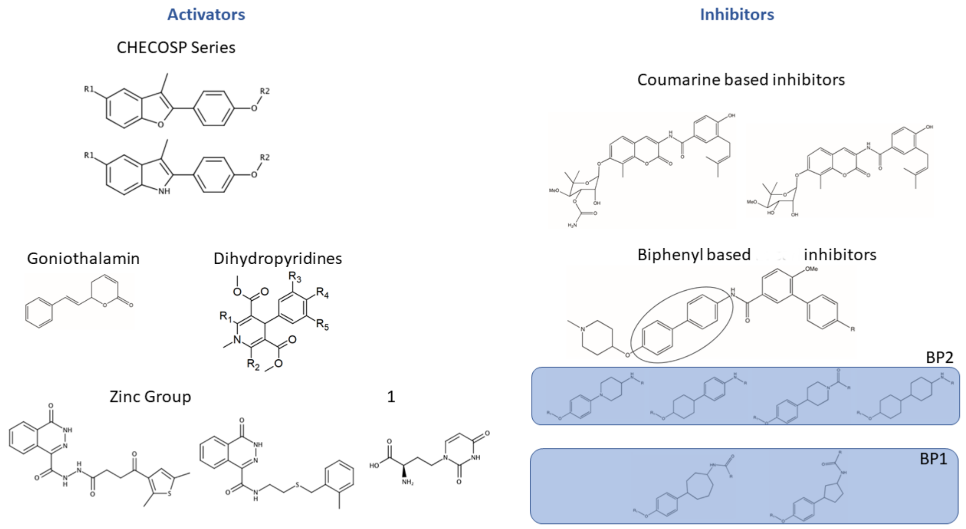


Figure 4.6: Here is shown the subdivision of the employed molecules in distinct groups, according to a shared scaffold. The 2D structure of the scaffolds are divided in eight groups, from left to right: CheCOSP molecules (CC), coumarina based inhibitors (CB), goniothalamin (GT), dihydropyridines (DP), the biphenyl inhibitors set is splitted in two groups (BP1 and BP2), the Zinc Group (Z) and lastly the compound labelled with 1 make its own group (Unk).

Yet, the compounds can still display substantial differences in their substituents in terms of dimensions, charges and functional groups. Therefore, a classification based only on the core of the molecules would give only a rough estimate of the chemical variability in the dataset. For this reason, to explore the possibility of classifying the function of molecules based only on their chemical properties, we used a more quantitative method based on cheminformatics similarity criteria. A common method to evaluate the similarity among compounds is to compute the Tanimoto coefficient on molecular fingerprints[16]. The efficacy of similarity algorithms tends to vary with biological activity; therefore the choice of the fingerprint model usually depends on the system under study. Here, our aim is specifically to introduce a metric for the comparison with our ML-dynamics based predictions. Since the best fingerprint model for our dataset is not known we tried two widely used methods: ECFP, a method that maps a molecule with a set of fragments radially grown from each heavy atom; and MACCS, that accounts for the presence/absence of specific structural features[167][1]. In both cases the molecules are clustered using k-means algorithm with a cluster number varying from 2 to 4 (Table 2).

The ECFP fingerprint works better in separating the compounds between activators and inhibitors when only 2 clusters are chosen, while with 3 or 4 clusters the separation is similar. With 3 clusters we found that the CheCOSP[79] group is separated. This consists only of activators validated by experimental characterization. The result shows that, despite a shared scaffold, there is a substantial

	MACCS		ECFP	
	Activators	Inhibitors	Activators	Inhibitors
K=2				
Cluster 1	33	48	12	49
Cluster 2	51	1	72	0
K=3				
Cluster 1	45	1	67	0
Cluster 2	0	47	17	2
Cluster 3	39	1	0	47
K=4				
Cluster 1	16	0	42	0
Cluster 2	0	47	14	2
Cluster 3	36	1	0	47
Cluster 4	32	1	28	0

Table 4.2: In the Table is shown the separation among Activators and Inhibitors in different clusters for both MACCS and ECFP fingerprints. Using a k-mean algorithm the dataset is splitted with different values of K, i.e. the number of clusters. With K=2 ECFP have a better separation of Inhibitors from Activators respect to MACCS.

chemical variety in the group.

The best result obtained by ECFP fingerprint on two clusters was compared with the best ML predictive model obtained by SVM (all data in Appendix E). The comparison was broken down by chemical groups to explore how the two approaches perform on different subclasses of ligands. In Figure 4.7 we report the fraction of correct classifications for every group in our dataset. For the three inhibitors group (BP1, BP2 and CB) a high fraction of correctly classified is observed for ECFP, meaning that inhibitors have good chemical similarity, whereas for activators the fraction for ECFP is high only for CC group. In all the other groups (Z, DP, Unk and GT) the fraction is 0. Interestingly, the SVM model correctly predicts as activators even the groups with low similarity with CC (the group most extensively characterized at the experimental level). In this context, we notice that SVM still correctly predicts group Z to 0.3 (0.0 in the case of fingerprints), DP to 0.6, Unk and GT to 1. In contrast, inhibitors of the CB groups have good similarity with the rest of inhibitors but they are not correctly predicted by SVM.

Overall, the results of this comparative analysis suggest that the characterization of allosteric binding with the partner protein, which reverberates the cross-talk between the ligand and the receptor, captures the main structural and dynamic determinants at the basis of allosteric modulation. On the one hand, this approach is not dependent on the similarities among the molecular structures of the libraries of compounds under exam. Yet, considering that specific functionalities may determine recognition, binding and the successive functional regulation, it is important to underline that the relevance of specific chemotypes for functional modulation emerges from the ML analysis. This aspect is aptly captured by the suitable combination of docking and Molecular Dynamics.

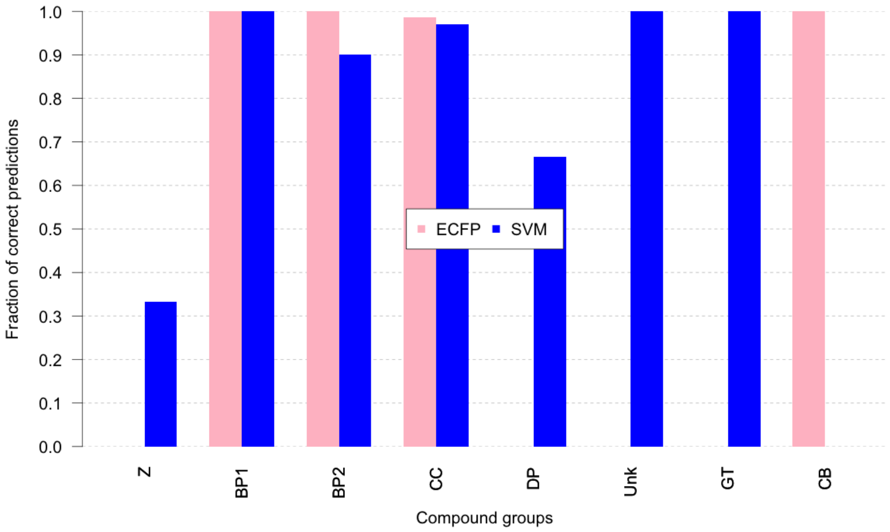


Figure 4.7: In this Table are reported the fraction of correct predictions obtained with SVM method compared with the cluster separation of ECFP values. An entry of the ECFP cluster is considered correctly separated if is situated in the cluster that contains the majority of his class. The values are in pink for ECFP and in blue for SVM, the fractions are evaluated for every scaffold group separately.

The most successful predictor is a learnt supervised model built on features describing the protein-ligand interaction across the whole set of representative structures from the conformational panel. Attempts to use only some features or some structures leads to poorer performance. This is consistent with the current understanding that functional activation by allosteric ligands is often mediated by the ligand "selecting" some of the conformational states. Information on both selected and non-selected states is required to identify effective binding. This also suggests that the model has learnt the relationship between selective binding patterns and functional effect. Therefore, the need for more sophisticated unsupervised algorithm is explained: this relationship is multivariate, not known in its analytical form and complex.

It can be suggested that the ML strategy we have presented here, while demonstrated on a specific but highly challenging case, is not system-specific and could be extended to the study of other allosterically regulated systems: in this context, this method can be proposed as a valid complement to the selection of allosteric leads for potential drug-development.

4.3 Allosteric effects in integrin αvβ6

The integrins are a class of transmembrane receptors that permit cell anchoring to the extracellular matrix (ECM) and also contribute to the connection among cells in metazoan organisms. Moreover, these proteins are able to induce important intracellular pathways through the binding to the cytoskeleton[108]. Here we focus on a specific type that is αvβ6. This integrin has the peculiarity to be present only on epithelial cells. Throughout embryo development integrin αvβ6 has high expression levels in forming lungs, skin and kidney, while its expression is considerably reduced in developed healthy tissues[17]. In contrast, increased expression of αvβ6 has been related to an enanched aggressiveness in different forms of tumor, such as colon and ovarian cancer or endometrial carcinoma[17]. In addition integrin αvβ6 is involved in the recognition of the Transforming Growth Factor β (TGFβ), a family of molecular factors that are central for organism development and are linked to cancer generation[166], and activation of the associated pathways.

This integrin is a molecule of significant size, composed by two distinct subunits, an αv subunit of 1048 residue length and a β6 subunit of 788 residue with a size similar to other integrins. The two subunits extend from the cell membrane and form non covalent bonds in the apical domains; the extracellular structure could be divided in an head part and two legs connected to the transmembrane domains. The Headpiece consists of two terminal domains of αv subunit (the b Propeller and the thigh) and two domains of β6 (bI and the hybrid). In the between of bPropeller and bI interface there is the site for the recognition of a residue motif, Arg-Gly-Asp (RGD), that is conserved in many partners, such as fibronectin, vitronectin and TGFβ proteins[191]. Integrin activation, defined as an increase of the affinity for RGD ligands, is regulated by a swing between two conformations: when inactive, the structure forms a bent conformation with the head pointing towards the membrane, while when it is activated the protein is extended with the loop between thigh (or hybrid) domain and the leg working as a “knee”. From a structural point of view the activation mechanism has partially become accessible since the release of X-ray crystal of protein headpiece in complex with the binding region of the Latency Associated Peptide (LAP) of pro-TGFβ3[49].

In the past years different studies have started to shed light on the activation of integrin αvβ6 and binding to pro-TGFβ complexes, with the consequent TGFβ release. Recent work by Dong et al. have shown the importance of the hybrid domain in regulating binding affinity, hybrid domain removal produce a 50 fold higher activity[51]. Moreover, the study of the pro-TGFβ activation induced by integrin αvβ6 has shown, in cell experiments, that deletion of the β6 cytoplasmic domain inhibits the activation, suggesting that a mechanical force might be necessary to TGFβ release. The role of the two subunits in the traction-based pro-TGFβ activation has been investigated with steer molecular dynamics (SMD), where a force is applied to αvβ6 pro-TFGβ1 complex on one subunit terminal residue and resisted by the pro-TGFβ1 molecule[50]. The authors showed that the only subunit that resist to the mechanical traction required for TFGβ1 release is the β6, suggesting that this subunit is principally involved in allosteric regulation. In spite of these results it is still difficult to depict the effect of the peptide on the integrin dynamics.

In order to acquire further details of the effect of the peptide on the structure and its dynamic, it could be interesting to analyse the effect that the LAP peptide produce on the structure and its dynamics. The main scope of the study is to analyse the dynamical and structural properties that could be connected to the allosteric (de)activation of the integrin. To this end, molecular dynamics simulations are attempted. Starting from the crystal structure (pdb code 4UM9), two systems are prepared: one with the bounded peptide and one in the apo form. For the simulations the Amber18 package with the ff99SBIldn force field is used[128]. Both the structures were solvated, with an explicit solvent box of the TIP3P model, Na+ counterions are added to obtain electroneutrality. In the metal ion binding sites Mg2+ are modelled with the parameters derived by Allner et al.[5]. The final systems consist of more than 200000 atoms. Each system was submitted to a 100 ps equilibration (NVT and NPT) then simulated in unrestrained conditions, for both the cases (APO and peptide bounded) 4 replicas were conducted of 1 µs each in the NPT ensemble; temperature and pressure were set to 300K and 1atm with the use of Berendsen and Langevin algorithms[21][225].

As a first thing we observe the effect of the ligand on the global protein structure, measuring the variation of the radius of gyration (RG) of the molecule (see Figure 4.9). The starting crystal has a radius of gyration of 34.1 Å, the trajectories of APO and LAP depict a distribution of the RG that has a two peaks shape for LAP case, with a small peak around 33.5 Å and a larger peak around 35.8 Å, whereas the APO case has a broader distribution with a well-defined peak around 37.5 Å and a shoulder at smaller values. The results clearly show that the starting configuration is not represented in the simulated ensembles, in addition the protein in absence of ligand (APO) seems to possess a greater freedom of movements arriving to extend over 37 Å of RG. Since the central rearrangement happens in the protein legs, in particular in the swing-out of the hybrid domain, the analysis is then focused on the specific domain motion and internal reorganisation. The RMSD values on Cα atom are computed for every domain. Moreover, in order to observe the mobility of the thigh and hybrid domain in the respective subunit, two measures of their displacement are also computed, i.e. the rmsd of the thigh domain upon structural alignment on the bP domain and the rsmd of the hybrid domain after the alignment on the bI domain. The values show similar distributions for both APO and LAP configuration in all of the four domains, with the thigh and the hybrid having a broader distribution with respect to the other two. The measures of the displacement show that the thigh is extremely flexible in the subunit. In fact it reaches large rmsd values, with the APO and the LAP systems having similar distributions. The displacement of the hybrid domain from the starting closing position is more contained, but the distribution of the values in APO configuration show a peak in higher values respect to the LAP case (see Figure 4.11 F).

The thigh domain gives most of the contribution to the motion of the protein, and this is not surprising because the domains in the β6 subunit have a double reinforcement, having a pair of polypeptide connections strengthen by disulfide bonds, that is necessary to resist the mechanical traction. Nevertheless, this domain has little contribution to the modulation of the dynamic occurred with the peptide binding, in fact the hybrid domain has a relevant decrease in the rmsd values range, suggesting an increased rigidity of the subunit when the peptide is bound.

Further confirmation of this result can be obtained computing the distance fluctuations (DF), i.e. the fluctuations of the distances among residue pair over the whole trajectory, for a pair of residues i and j:

$$DF_{ij} = \langle (r_{ij} - \langle r_{ij} \rangle)^2 \rangle \tag{4.1}$$

The values form a matrix with the DF value of every pair, and to evaluate the contribution of every residue to the DF of the rest of the protein is computed the sum of every column of the DF matrix (formula 4.1), obtaining a per residue score. The DF profiles for APO and LAP configuration are very close except for the residues of the hybrid domain, were the APO case has more DF score respect to the LAP, meaning that with the peptide bound the domain has lower fluctuations (see Figure 4.10, for the complete matrices see Appendix F).

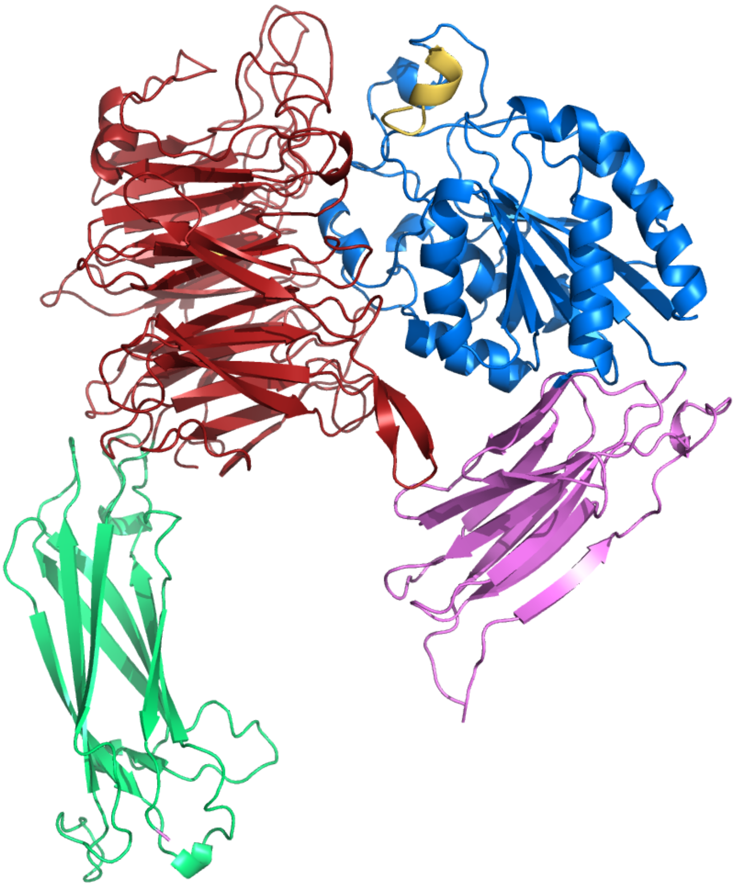


Figure 4.8: Here is represented the structure of αvβ6 integrin. The β subunit is formed by bI, colored in blue, and the hybrid, in pink, while the α subunit consist of the b Propeller, in red, and the thigh, in green. In yellow is showed a peptide of pro-TGFβ3.

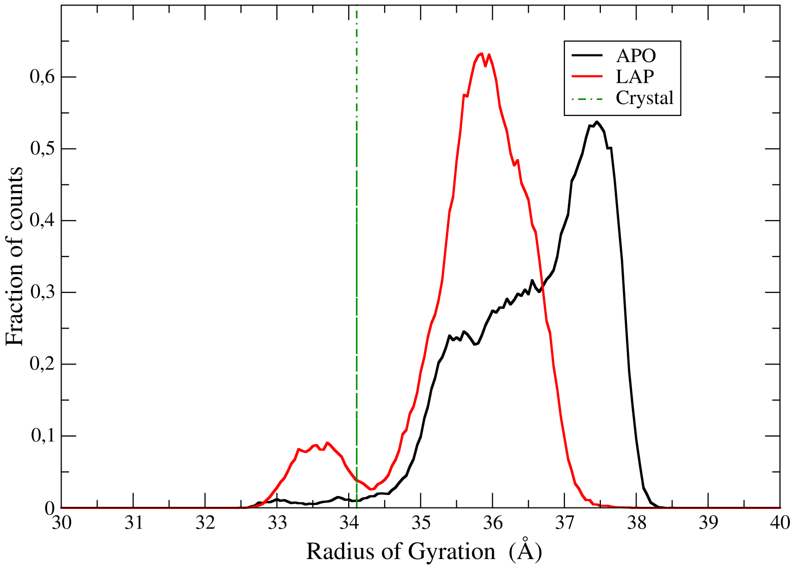


Figure 4.9: Here are represented the distribution of Radius of Gyration for the APO (black) and the LAP bound (red) configurations. The value of RG for the starting structure is colored in green.

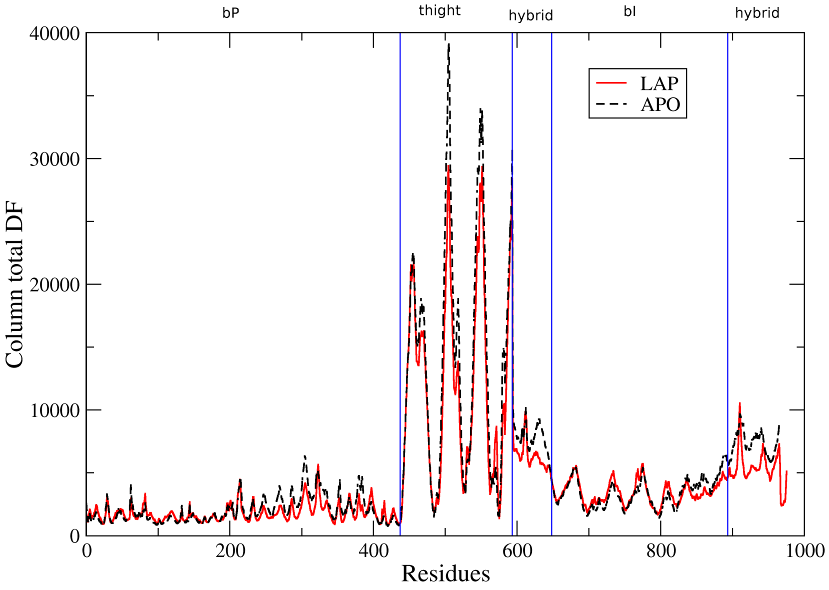


Figure 4.10: Here are reported the total column counts of the DF matrix. The residues of the hybrid domain display the major differences between APO and LAP state.

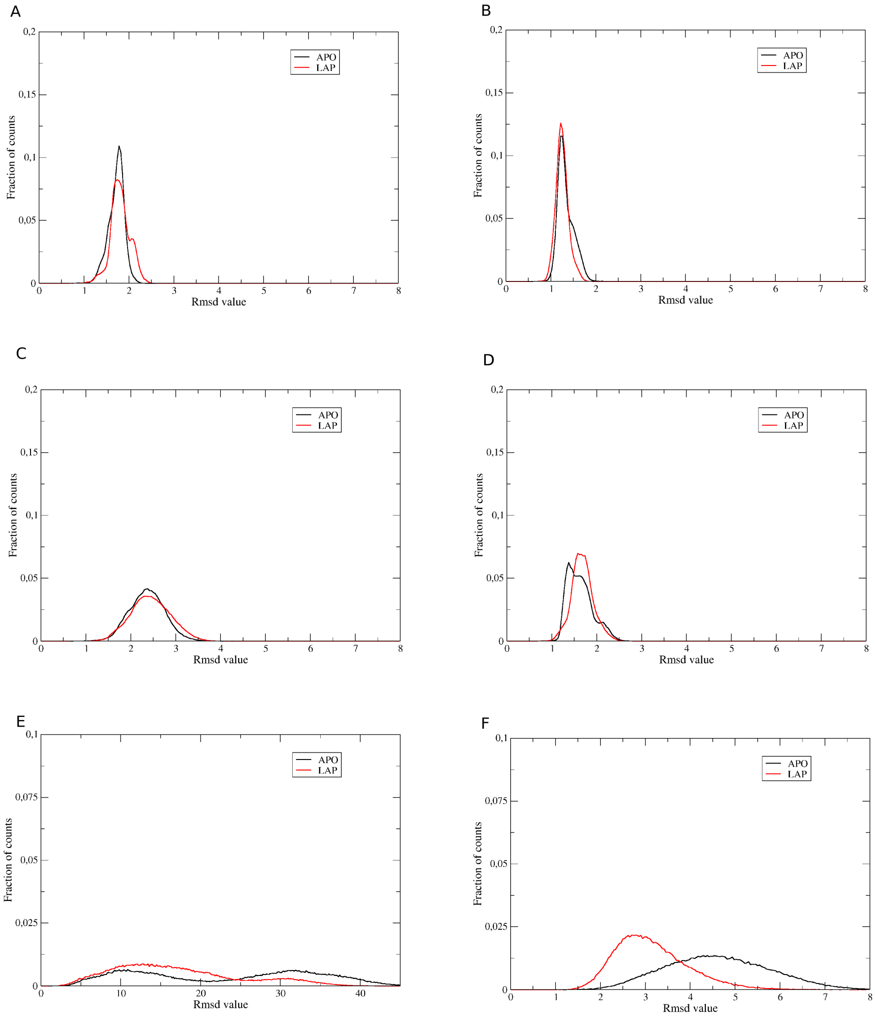


Figure 4.11: In the Figure are reported the distribution of the rmsd in the bP (A), bI (B), thight (C) and hybrid (D) domains, displaying similar distributions of the values for APO and LAP configurations. The measures of the displacement from the rest of the subunit is showed in E for thigh domain and in F for hybrid.

Chapter 5

Conclusions

The study of protein-protein interactions remains a complicated challenge, in spite of the continuous progresses that are being made in both experimental and theoretical fields. In this thesis these problems are investigated from a computational chemistry point of view using theoretical tools and simulating molecular models at an atomistic level.

In the first part of the thesis, I studied methods for the prediction of the residues involved in protein-protein interactions. In chapter 2, I presented two different scores, one based on evolutionary information and one based on the energetics of the protein, on a dataset of crystal structures. Both scores have the capability to discriminate the interface region from the rest of the protein in a relevant fraction of cases. Moreover, a comparison of the scores efficacy on distinct protein classes highlights the importance of considering the biological function of the protein on the performance of the method used for the prediction of interface residues.

In chapter 3 the energetic method for interface residues prediction is used for the detection of antigenic epitopes on the spike protein of SARS-CoV-2. The regions predicted were confirmed against experimental complexes expanding our understanding of the molecular basis for interactions. In perspective, the acquired knowledge could be used for the design of novel vaccine candidates and diagnostic tools and to increase our readiness in the case of future epidemics.

In the second part there, I focussed on the study of two allosteric systems, presented in chapter 4. Firstly a method is presented for the integration of an ensemble docking protocol with a learning classifier for allosteric ligands of the protein Hsp90. The method reaches a good accuracy in classifying the activity of these ligands and this approach seems to reduce the dependency on the chemical similarity of the compounds used for the training. The method is tested on a limited dataset and further developments could be achieved in the future if the library of compounds is increased. In the end, I presented the initial analysis of an allosteric signal for integrin αvβ6 in complex with a pro-TGFβ peptide, with the use of molecular dynamics simulations. The data suggest that the presence of the peptide induces a rigidification of the legs of the structure, in particular for a specific domain.

In conclusion, with my thesis I showed how the development and use of computational and theoretical approaches can be very helpful in the study of the structure-dynamics-function relationships in biological problems. The hope is that with the increase of accessible structural data the strategies presented can be further optimised and find extended reach. Combining the application of computational chemistry techniques with the improvement of models and atomistic simulations will permit to acquire new knowledge on the fundamentals trait of protein-protein interactions.

Publications

As of November 2020

Refereed Publications

1. The Subtle Trade-Off between Evolutionary and Energetic Constraints in Protein-Protein Interactions

   Marchetti F. and Capelli R. and Rizzato F. and Laio A. and Colombo G.

   Journal of Physical Chemistry Letters, 2019, 10(7), pp. 1489-1497

2. The Answer Lies in the Energy: How Simple Atomistic Molecular Dynamics Simulations May Hold the Key to Epitope Prediction on the Fully Glycosylated SARS-CoV-2 Spike Protein

   Serapian S.A. and Marchetti F. and Triveri A. and Rasola A. and Colombo G.

   Journal of Physical Chemistry Letters, 2020, 11(19), pp. 8084-8093

Appendices

Appendix

A KS and AD values complete list

Here is the complete list of the p-values, for EVT and MLCE data, produced with Kolmogorov-Smirnov (KS) and Anderson-Darling (AD) methods. The entries are sorted referring to the pdc code of the "Complex", i.e. the protein-protein crystal used as a benchmark for the evaluations. In "Monomer" are listed the pdb code the structure used for the calculations.

Complex	Monomer	EVT pvalues		MLCE pvalues	
		KS	AD	KS	AD
1A2K	1OUN	0.01962	0.04335	0.01285	0.01182
1A2K	1QG4	0.00012	0.00001	0.39582	0.39449
1ACB	1EGL	0.53601	0.67010	0.00015	0.00087
1ACB	2CGA	0.01588	0.05452	0.27513	0.41297
1AFV	1GWP	0.51531	0.87830	0.00003	0.00000
1AHW	1TFH	0.72129	0.81813	0.00072	0.00003
1AK4	2CPL	0.00212	0.00225	0.00403	0.00243
1AK4	4J93	0.43648	0.36281	0.35533	0.21066
1ATN	1IJJ	0.54841	0.60690	0.08677	0.19492
1ATN	3DNI	0.40755	0.57711	0.22334	0.28559
1AVX	1BA7	0.39693	0.29394	0.12339	0.21474
1AVX	1QQU	0.11743	0.27805	0.80819	0.92044
1AY7	1A19	0.00002	0.00000	0.32815	0.21121
1AY7	1RGH	0.03612	0.09591	0.00114	0.00188
1AZS	1AB8	0.49945	0.44841	0.43939	0.57670
1AZS	1AZT	0.00009	0.00001	0.04903	0.01198
1B6C	1D6O	0.02585	0.03839	0.09286	0.27594
1BGX	1AY1	0.00001	0.00002	0.00010	0.00000
1BGX	1TAQ	0.03654	0.07526	0.08114	0.26200

Continued on next page

1BUH	1DKS	0.10842	0.14496	0.23894	0.30447
1BUH	1HCL	0.16773	0.07548	0.00565	0.01036
1DQJ	1DQQ	0.03724	0.12974	0.69725	0.75170
1DQJ	3LZT	0.08909	0.31898	0.21692	0.20593
1E6J	1A43	0.47983	0.55374	0.00042	0.00021
1E6J	1E6O	0.00002	0.00003	0.09128	0.02651
1EAW	1EAX	0.55519	0.59646	0.63219	0.83276
1EAW	9PTI	0.08161	0.17814	0.15819	0.12234
1EXB	1QDV	0.20405	0.21769	0.71220	0.78937
1EXB	1QRQ	0.13564	0.19309	0.03218	0.02760
1EZU	1ECZ	0.26734	0.30878	0.22186	0.37589
1EZU	1TRM	0.11009	0.14224	0.75884	0.97566
1FE8	1AO3	0.52854	0.31669	0.00128	0.00291
1FFW	1FWP	0.00016	0.00064	0.30779	0.40460
1FFW	3CHY	0.66576	0.88569	0.00073	0.00127
1FSK	1BV1	0.01523	0.03330	0.00018	0.00012
1GL1	1K2I	0.34100	0.37286	0.29214	0.27828
1GL1	1PMC	0.47252	0.54939	0.28694	0.23952
1GPW	1K9V	0.00007	0.00006	0.07724	0.18138
1GPW	1THF	0.39655	0.42723	0.00858	0.04146
1GRN	1A4R	0.00002	0.00003	0.00103	0.00127
1GRN	1RGP	0.00001	0.00000	0.00031	0.00125
1H0D	1K59	0.86116	0.88208	0.00235	0.00133
1H9D	1EAN	0.04704	0.01585	0.03231	0.06949
1H9D	1ILF	0.01088	0.00438	0.38177	0.42839
1HCF	1B98	0.04061	0.15595	0.00558	0.00549
1HCF	1WWB	0.05915	0.17135	0.04613	0.01031
1I4D	1I49	0.00913	0.00192	0.00706	0.00124
1I4D	1MH1	0.00121	0.00059	0.10560	0.09176
1IC4	3LZT	0.05501	0.37783	0.28795	0.32815
1IQD	1D7P	0.02772	0.03810	0.00027	0.00012
1JMO	2CN0	0.27065	0.39004	0.01077	0.02362
1K74	1MZN	0.00001	0.00000	0.00040	0.00162
1K74	1ZGY	0.00001	0.00000	0.00058	0.00149
1KAC	1F5W	0.72954	0.90662	0.36949	0.59829
1KAC	1NOB	0.23251	0.31988	0.09749	0.08832
1KKL	1JB1	0.34940	0.33881	0.00565	0.01252
1KKL	2HPR	0.00564	0.00830	0.00058	0.00537
1KTZ	1M9Z	0.14936	0.37139	0.16925	0.11011
1KTZ	1TGJ	0.69346	0.53826	0.35224	0.32961
1LFD	1LXD	0.37617	0.29214	0.49505	0.47838
1LFD	5P21	0.00029	0.00012	0.00320	0.04796
1MHP	1CK4	0.01924	0.01401	0.18202	0.18616
1ML0	1DOL	0.42853	0.46861	0.51361	0.59236
1MQ8	1IAM	0.02133	0.08125	0.00012	0.00001

Continued on next page

1MQ8	1MQ9	0.11879	0.08609	0.32910	0.42444
1NMC	7NN9	0.02230	0.01792	0.28224	0.24707
1NSN	1KDC	0.00513	0.02633	0.04089	0.07887
1NW9	1JXQ	0.04124	0.00552	0.08926	0.21750
1NW9	2OPY	0.11492	0.08058	0.03374	0.07119
1OAK	1AUQ	0.04964	0.15531	0.00002	0.00000
1OC0	1B3K	0.63578	0.67480	0.17483	0.26741
1OC0	2JQ8	0.47189	0.57640	0.02693	0.01532
1OPH	1QLP	0.00118	0.00062	0.08080	0.12449
1OPH	1UTQ	0.26066	0.61346	0.11282	0.16771
1PKQ	1PKO	0.61045	0.67323	0.00001	0.00000
1PPE	1BTP	0.12891	0.15744	0.32738	0.51777
1PPE	1LU0	1.01072	0.84332	0.02788	0.03958
1QFW	1HCN	0.64957	0.65235	0.11709	0.38176
1R0R	1SCN	0.00038	0.00901	0.37580	0.51847
1R0R	2GKR	0.87298	0.99291	0.07437	0.04644
1R8S	1R8M	0.00001	0.00000	0.03815	0.01546
1RJL	1P4P	0.27244	0.22556	0.00001	0.00000
1RKE	1SYQ	0.23640	0.12208	0.11097	0.39965
1RKE	3MYI	0.00026	0.00038	0.00133	0.00149
1SYX	1L2Z	0.05921	0.31324	0.15449	0.45788
1SYX	1QGV	0.15881	0.15615	0.00056	0.00533
1TPX	1UW3	0.01224	0.00373	0.34985	0.44862
1WDW	1GEQ	0.00001	0.00000	0.00276	0.01109
1WDW	1V8Z	0.04085	0.01765	0.05198	0.00000
1XU1	1U5Y	0.42915	0.35373	0.00017	0.00005
1XU1	1XUT	0.16269	0.16247	0.04574	0.04707
1Y64	1UX5	0.00583	0.00407	0.00265	0.00205
1Y64	2FXU	0.53906	0.38687	0.00198	0.00000
1YNT	1KZQ	0.00020	0.00023	0.00135	0.00019
1YVB	1CEW	0.02825	0.04426	0.49367	0.36539
1YVB	2GHU	0.08346	0.45829	0.33124	0.44914
1Z0K	1YZM	0.00034	0.00740	0.07974	0.16594
1Z0K	2BME	0.00003	0.00014	0.00136	0.00018
1Z5Y	1L6P	0.00005	0.00003	0.17624	0.24693
1Z5Y	2B1K	0.01074	0.06878	0.00448	0.03525
1ZHH	1JX6	0.61539	0.62488	0.00097	0.00070
1ZHH	2HJE	0.79629	0.80256	0.31544	0.33690
2A9K	1U90	0.00412	0.00508	0.00326	0.00266
2A9K	2C8B	0.12875	0.15513	0.00051	0.00350
2AJF	1R42	0.00024	0.00151	0.00004	0.00000
2AJF	2GHV	0.01480	0.02068	0.31109	0.32797
2B4J	1BIZ	0.18230	0.39511	0.00708	0.00824
2B4J	1Z9E	0.21073	0.08210	0.05896	0.04140
2BTF	1IJJ	0.17502	0.10844	0.30214	0.19975

Continued on next page

2BTF	1PNE	0.47349	0.11455	0.00339	0.00638
2CFH	1SZ7	0.00002	0.00001	0.00036	0.00012
2CFH	2BJN	0.00009	0.00069	0.07291	0.06046
2FJG	2VPF	0.52374	0.52660	0.00011	0.00002
2GTP	1GFI	0.00031	0.00359	0.00072	0.00495
2GTP	2BV1	0.00002	0.00003	0.01066	0.04432
2H7V	1MH1	0.00001	0.00000	0.10136	0.03832
2H7V	2H7O	0.69620	0.90647	0.00005	0.00000
2HLE	1IKO	0.00018	0.00085	0.25018	0.24883
2HLE	2BBA	0.03280	0.02089	0.01684	0.02552
2I25	2I24	0.00240	0.00167	0.00001	0.00000
2I25	3LZT	0.36576	0.20966	0.00004	0.00012
2IDO	1J54	0.74710	0.93509	0.65756	0.82705
2IDO	1SE7	0.69562	0.26026	0.58654	0.65668
2J0T	1D2B	0.92797	0.98533	0.00002	0.00021
2J0T	966C	0.00022	0.00041	0.25123	0.14485
2JEL	1POH	0.03568	0.03833	0.00014	0.00011
2MTA	2BBK	0.02665	0.03854	0.00529	0.03594
2MTA	2RAC	0.00490	0.00136	0.01522	0.02655
2O3B	1J57	0.02193	0.02708	0.25435	0.31461
2O3B	1ZM8	0.00051	0.00545	0.45767	0.52061
2O8V	1SUR	0.36092	0.33398	0.53863	0.64993
2O8V	2TRX	0.00002	0.00001	0.00467	0.02048
2OUL	2NNR	0.00009	0.00003	0.14317	0.41136
2OUL	3BPF	0.06921	0.45741	0.48415	0.70292
2OZA	3FYK	0.00721	0.01902	0.08607	0.03172
2OZA	3HEC	0.00002	0.00000	0.87345	0.96957
2PCC	1CCP	0.40793	0.50139	0.16408	0.26045
2PCC	1YCC	0.00651	0.00398	0.35859	0.42749
2SIC	1SUP	0.00204	0.01089	0.16062	0.15035
2SIC	3SSI	0.10401	0.32513	0.00611	0.02508
2SNI	1UBN	0.00015	0.00222	0.56325	0.50348
2SNI	2CI2	0.46675	0.83527	0.13578	0.23430
2VIS	1GIG	0.00001	0.00000	0.00661	0.00038
2VIS	2VIU	0.04892	0.04816	0.02195	0.01436
2X9A	1S62	0.49912	0.62379	0.00267	0.00242
2X9A	2X9B	0.06878	0.44064	0.05481	0.08842
2Z0E	1V49	0.04379	0.02336	0.00003	0.00015
2Z0E	2D1I	0.00002	0.00000	0.30022	0.42872
3A4S	1A3S	0.62811	0.38344	0.84766	0.84853
3A4S	3A4R	0.02010	0.01580	0.13788	0.25319
3BIW	2R1D	0.18900	0.26298	0.00223	0.00146
3BIW	3BIX	0.91645	0.75337	0.02667	0.04173
3CPH	1G16	0.00074	0.00040	0.00418	0.00110
3CPH	3CPI	0.00005	0.00000	0.00034	0.00026

Continued on next page

3DAW	1IJJ	0.02222	0.02752	0.31038	0.43423
3DAW	2HD7	0.00001	0.00002	0.30420	0.47156
3F1P	1P97	0.00407	0.00128	0.12952	0.07122
3F1P	1X0O	0.00138	0.00223	0.17999	0.28016
3FN1	2EDI	0.13243	0.16093	0.02183	0.00223
3FN1	2LQ7	0.34223	0.59690	0.03940	0.14853
3H11	3H13	0.13523	0.31637	0.01487	0.01261
3H11	4JJ7	0.00035	0.00343	0.00898	0.03993
3H2V	1WI6	0.73605	0.57458	0.02929	0.01276
3H2V	3MYI	0.21574	0.22754	0.56814	0.52861
3LVK	1DCJ	0.00029	0.00003	0.06052	0.02743
3LVK	3LVM	0.67765	0.46698	0.00036	0.00074
3PC8	3PC6	0.68738	0.89082	0.05843	0.01920
3PC8	3PC7	0.00452	0.03788	0.00217	0.01746
3SGQ	2OVO	0.10716	0.29674	0.83309	0.98360
3SGQ	2QA9	0.02797	0.02228	0.02925	0.04686
3V6Z	3KXS	0.65654	0.99508	0.01775	0.05772
3V6Z	3V6F	0.00001	0.00000	0.03455	0.05582
4CPA	8CPA	0.07229	0.15148	0.02781	0.19487
4FZA	1UPL	0.00001	0.00000	0.00005	0.00000
4FZA	3GGF	0.25245	0.09444	0.00009	0.00014
4GXU	1RUZ	0.25298	0.31639	0.00400	0.00297
4GXU	4GXV	0.00001	0.00001	0.10958	0.01827
4HX3	1C7K	0.00006	0.00004	0.97481	0.98191
4HX3	4HWX	0.36607	0.29602	0.01291	0.01001
4IZ7	1ERK	0.00153	0.00604	0.01787	0.03846
4IZ7	2LS7	0.00355	0.00369	0.00080	0.00024
4M76	1C3D	0.16389	0.27475	0.16671	0.21423
4M76	1M1U	0.18682	0.03974	0.17159	0.21588

Appendix B

AUC values complete list

Full list of the AUC value computed for EVT, MLCE and EVT+MLCE methods. The list is sorted respect to the pdb code of the “Complex“. It is also indicated the degree of flexibility of the interface: R is for rigid body, M for medium, D for difficult and N for not labelled. It is also reported the group of activity of the protein with the labels: antibody (Ab), enzymes (Ez), antigens (Ag), inhibitors (In), signal transmission (ST) and structural (Sr).

Complex	Monomer	Hardness	Group	EVT	MLCE	EVT+MLCE
1A2K	1OUN	R	ST	0.63315	0.68931	0.72917
1A2K	1QG4	R	ST	0.92419	0.53791	0.87818
1ACB	1EGL	D	In	0.51990	0.82526	0.70675
1ACB	2CGA	D	Ez	0.63461	0.47668	0.58771
1AFV	1GWP	N	Ag	0.48195	0.16880	0.28083
1AHW	1TFH	N	ST	0.53959	0.27993	0.40755
1AK4	2CPL	R	Ez	0.69385	0.73333	0.81692
1AK4	4J93	R	Ag	0.41279	0.58188	0.40892
1ATN	1IJJ	D	Sr	0.49537	0.52116	0.51240
1ATN	3DNI	D	Ez	0.53209	0.54224	0.53975
1AVX	1BA7	R	In	0.58947	0.43217	0.55071
1AVX	1QQU	R	Ez	0.66838	0.46936	0.58774
1AY7	1A19	R	In	0.96667	0.65370	0.90463
1AY7	1RGH	R	Ag	0.68556	0.25506	0.49258
1AZS	1AB8	R	Ez	0.56151	0.46711	0.60000
1AZS	1AZT	R	ST	0.81818	0.51578	0.83633
1B6C	1D6O	M	Ez	0.65222	0.52632	0.66512
1BGX	1AY1	D	Ab	0.21875	0.44935	0.22543

Continued on next page

1BGX	1TAQ	D	Ag	0.58905	0.46239	0.52774
1BUH	1DKS	R	ST	0.60251	0.40873	0.61442
1BUH	1HCL	R	ST	0.58406	0.42177	0.56584
1DQJ	1DQQ	R	Ab	0.51845	0.51716	0.50833
1DQJ	3LZT	N	Ez	0.38024	0.56640	0.45336
1E6J	1A43	R	Ag	0.46804	0.76316	0.63628
1E6J	1E6O	R	Ab	0.08149	0.41544	0.12990
1EAW	1EAX	R	Ez	0.65168	0.50176	0.60714
1EAW	9PTI	R	In	0.65079	0.64484	0.59821
1EXB	1QDV	R	ST	0.33929	0.39397	0.33259
1EXB	1QRQ	R	Ez	0.48919	0.56376	0.54623
1EZU	1ECZ	R	In	0.54758	0.58548	0.57540
1EZU	1TRM	R	Ez	0.62151	0.49117	0.60201
1FE8	1AO3	N	Ag	0.45267	0.39736	0.43753
1FFW	1FWP	R	ST	0.79196	0.52054	0.85357
1FFW	3CHY	R	ST	0.48561	0.82727	0.65682
1FSK	1BV1	N	Ag	0.70103	0.21263	0.45773
1GL1	1K2I	R	Ez	0.59463	0.54245	0.60849
1GL1	1PMC	R	In	0.50625	0.39167	0.32083
1GPW	1K9V	R	Ez	0.78319	0.51176	0.76891
1GPW	1THF	R	Ez	0.54880	0.34888	0.49311
1GRN	1A4R	M	ST	0.81422	0.30889	0.63711
1GRN	1RGP	M	ST	0.84786	0.24013	0.68586
1H0D	1K59	N	Ag	0.51734	0.74919	0.66035
1H9D	1EAN	R	ST	0.29427	0.66840	0.41884
1H9D	1ILF	R	ST	0.68286	0.55086	0.65029
1HCF	1B98	R	ST	0.63725	0.47304	0.57892
1HCF	1WWB	R	Ab	0.33483	0.33227	0.28169
1I4D	1I49	R	ST	0.32338	0.71852	0.53519
1I4D	1MH1	M	ST	0.78763	0.45498	0.73333
1IC4	3LZT	N	Ez	0.35281	0.57013	0.43896
1IQD	1D7P	N	Ag	0.28701	0.76809	0.46382
1JMO	2CN0	D	Ez	0.57780	0.43320	0.49700
1K74	1MZN	R	ST	0.86302	0.36940	0.75833
1K74	1ZGY	R	ST	0.81817	0.39144	0.73513
1KAC	1F5W	R	Ab	0.53279	0.61326	0.60469
1KAC	1NOB	R	Ag	0.59709	0.60259	0.77379
1KKL	1JB1	M	Ez	0.60422	0.30687	0.55045
1KKL	2HPR	M	Ez	0.77047	0.81652	0.87865
1KTZ	1M9Z	R	ST	0.61686	0.35153	0.63123
1KTZ	1TGJ	R	ST	0.56410	0.37628	0.50128
1LFD	1LXD	M	ST	0.59860	0.44798	0.58075
1LFD	5P21	M	ST	0.80390	0.75411	0.86450
1MHP	1CK4	N	Ag	0.67760	0.54286	0.65455
1ML0	1DOL	R	ST	0.64848	0.61449	0.66816

Continued on next page

1MQ8	1IAM	M	Sr	0.28962	0.73113	0.44843
1MQ8	1MQ9	M	ST	0.63701	0.57558	0.60387
1NMC	7NN9	N	Ag	0.29310	0.57101	0.34497
1NSN	1KDC	N	Ag	0.27771	0.37967	0.27082
1NW9	1JXQ	M	Ez	0.71521	0.63582	0.69967
1NW9	2OPY	M	In	0.63838	0.49164	0.60493
1OAK	1AUQ	N	Ag	0.32895	0.57895	0.45263
1OC0	1B3K	R	In	0.55013	0.55556	0.57296
1OC0	2JQ8	R	ST	0.55714	0.36250	0.32679
1OPH	1QLP	R	In	0.28436	0.61351	0.38649
1OPH	1UTQ	R	Ez	0.63538	0.39428	0.59385
1PKQ	1PKO	N	ST	0.48462	0.31255	0.34777
1PPE	1BTP	R	Ez	0.60254	0.42679	0.57892
1PPE	1LU0	R	In	0.51339	0.79464	0.64732
1QFW	1HCN	N	Ag	0.40640	0.42672	0.36946
1R0R	1SCN	R	Ez	0.72533	0.55412	0.70977
1R0R	2GKR	R	In	0.49848	0.70000	0.66970
1R8S	1R8M	D	ST	0.88194	0.40262	0.79940
1RJL	1P4P	N	Ag	0.44602	0.68750	0.57623
1RKE	1SYQ	D	Sr	0.60549	0.50498	0.60192
1RKE	3MYI	D	Sr	0.78261	0.66848	0.79176
1SYX	1L2Z	M	ST	0.29980	0.65039	0.44531
1SYX	1QGV	M	Ez	0.67050	0.65625	0.73024
1TPX	1UW3	N	Ag	0.67954	0.48302	0.64856
1WDW	1GEQ	R	Ez	0.89485	0.65025	0.87723
1WDW	1V8Z	R	Ez	0.61758	0.54545	0.64146
1XU1	1U5Y	R	ST	0.54330	0.79657	0.68832
1XU1	1XUT	R	ST	0.67361	0.32118	0.46007
1Y64	1UX5	D	ST	0.64259	0.54941	0.66924
1Y64	2FXU	D	Sr	0.56740	0.61438	0.63252
1YNT	1KZQ	N	Ag	0.24543	0.27449	0.20303
1YVB	1CEW	R	In	0.64366	0.56641	0.58420
1YVB	2GHU	R	Ez	0.58109	0.47077	0.59783
1Z0K	1YZM	R	ST	0.84933	0.39600	0.74400
1Z0K	2BME	R	ST	0.83869	0.72561	0.87777
1Z5Y	1L6P	R	Ez	0.82846	0.45833	0.86797
1Z5Y	2B1K	R	Ez	0.72661	0.27131	0.60759
1ZHH	1JX6	R	ST	0.49680	0.56040	0.55321
1ZHH	2HJE	R	ST	0.45797	0.42940	0.41097
2A9K	1U90	R	ST	0.71146	0.36731	0.61434
2A9K	2C8B	R	Ez	0.40919	0.23658	0.21104
2AJF	1R42	R	Ez	0.26081	0.60749	0.36149
2AJF	2GHV	R	ST	0.29954	0.44815	0.28449
2B4J	1BIZ	R	Ez	0.31402	0.76601	0.49009
2B4J	1Z9E	R	ST	0.63404	0.70213	0.69574

Continued on next page

2BTF	1IJJ	D	Sr	0.63056	0.52222	0.63089
2BTF	1PNE	R	Sr	0.57063	0.28986	0.44545
2CFH	1SZ7	M	ST	0.83299	0.76468	0.83766
2CFH	2BJN	M	ST	0.28798	0.68367	0.33985
2FJG	2VPF	N	Ag	0.52905	0.88176	0.80473
2GTP	1GFI	R	ST	0.72565	0.32386	0.59713
2GTP	2BV1	R	ST	0.85526	0.31652	0.77485
2H7V	1MH1	M	ST	0.88207	0.45000	0.82989
2H7V	2H7O	M	ST	0.50949	0.56013	0.57278
2HLE	1IKO	R	ST	0.76842	0.52295	0.84511
2HLE	2BBA	R	ST	0.63487	0.61034	0.63333
2I25	2I24	R	Ab	0.24955	0.82576	0.48173
2I25	3LZT	N	Ez	0.52797	0.17213	0.26760
2IDO	1J54	D	Ez	0.50498	0.48007	0.48090
2IDO	1SE7	D	Ez	0.50208	0.49115	0.45469
2J0T	1D2B	R	In	0.51766	0.83634	0.75126
2J0T	966C	R	Ez	0.76540	0.56703	0.74517
2JEL	1POH	N	Ag	0.31404	0.16744	0.15895
2MTA	2BBK	R	Ez	0.81646	0.61835	0.62152
2MTA	2RAC	R	ST	0.78896	0.35649	0.66753
2O3B	1J57	D	In	0.65123	0.59627	0.67089
2O3B	1ZM8	D	Ez	0.72835	0.46078	0.66687
2O8V	1SUR	R	Ez	0.63254	0.43214	0.57302
2O8V	2TRX	R	Ez	0.88024	0.73953	0.92374
2OUL	2NNR	R	In	0.77813	0.54088	0.73201
2OUL	3BPF	R	Ez	0.63177	0.45149	0.61823
2OZA	3FYK	M	ST	0.65075	0.61376	0.69923
2OZA	3HEC	M	ST	0.71931	0.49722	0.71966
2PCC	1CCP	R	Ez	0.56745	0.57329	0.61398
2PCC	1YCC	R	ST	0.76582	0.58047	0.77358
2SIC	1SUP	R	Ez	0.68167	0.57802	0.69506
2SIC	3SSI	R	In	0.32820	0.17797	0.24268
2SNI	1UBN	R	Ez	0.76027	0.43511	0.73903
2SNI	2CI2	R	In	0.45415	0.63062	0.47145
2VIS	1GIG	R	Ab	0.11805	0.32481	0.11905
2VIS	2VIU	R	Ag	0.34448	0.65672	0.53086
2X9A	1S62	R	ST	0.42732	0.25464	0.28306
2X9A	2X9B	R	Ag	0.34911	0.35357	0.30089
2Z0E	1V49	M	ST	0.62607	0.76854	0.74239
2Z0E	2D1I	M	Ez	0.84896	0.47623	0.85136
3A4S	1A3S	R	Ez	0.53198	0.56007	0.50216
3A4S	3A4R	R	ST	0.79613	0.51042	0.70089
3BIW	2R1D	R	Sr	0.62724	0.27384	0.50717
3BIW	3BIX	R	Ez	0.45464	0.46564	0.38788
3CPH	1G16	M	ST	0.86688	0.78450	0.91802

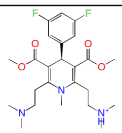
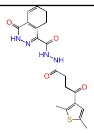
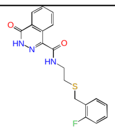
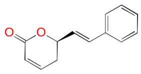
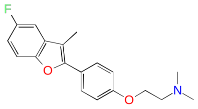
Continued on next page

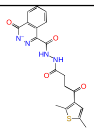
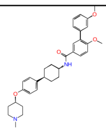
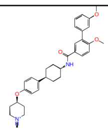
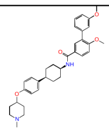
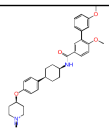
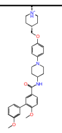
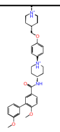
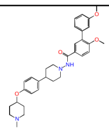
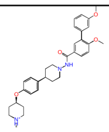
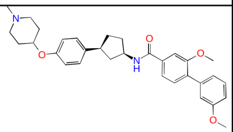
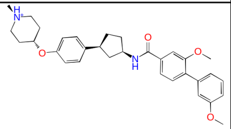
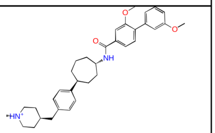
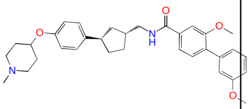
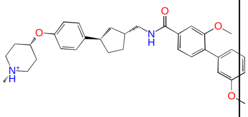
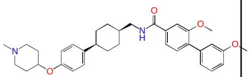
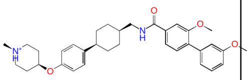
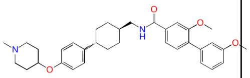
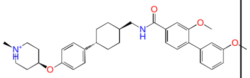
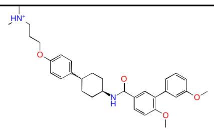
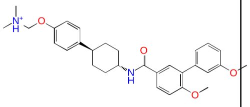
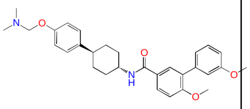
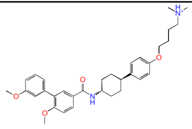
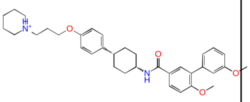
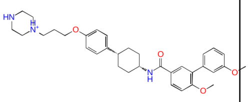
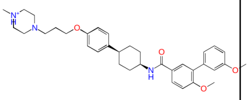
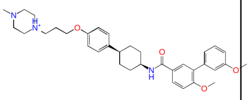
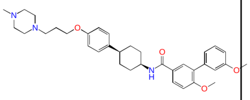
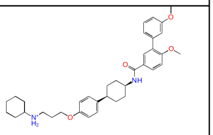
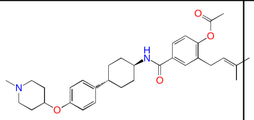
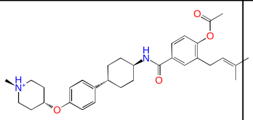
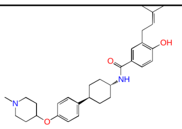
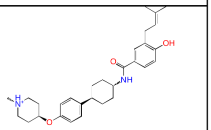
3CPH	3CPI	M	In	0.90026	0.77623	0.92416
3DAW	1IJJ	D	Sr	0.73618	0.52073	0.70747
3DAW	2HD7	M	Sr	0.84069	0.53622	0.87718
3F1P	1P97	D	ST	0.66346	0.42962	0.58115
3F1P	1X0O	D	ST	0.73550	0.60133	0.78616
3FN1	2EDI	D	Ez	0.49584	0.66536	0.57817
3FN1	2LQ7	D	ST	0.50119	0.65357	0.62222
3H11	3H13	D	Ez	0.46617	0.64575	0.51574
3H11	4JJ7	D	Ez	0.73490	0.45714	0.66406
3H2V	1WI6	R	ST	0.34764	0.59259	0.52273
3H2V	3MYI	D	Sr	0.47892	0.57342	0.53464
3LVK	1DCJ	R	Ez	0.83989	0.63864	0.84079
3LVK	3LVM	R	Ez	0.54245	0.27290	0.42309
3PC8	3PC6	R	ST	0.64709	0.37833	0.56598
3PC8	3PC7	R	ST	0.75676	0.22475	0.48222
3SGQ	2OVO	R	In	0.60937	0.50240	0.76923
3SGQ	2QA9	R	Ez	0.70443	0.69774	0.70025
3V6Z	3KXS	M	Ag	0.39663	0.34890	0.45330
3V6Z	3V6F	M	Ab	0.08580	0.65278	0.12623
4CPA	8CPA	R	Ez	0.68522	0.41489	0.57338
4FZA	1UPL	M	ST	0.85280	0.36624	0.71361
4FZA	3GGF	M	ST	0.54139	0.74752	0.66838
4GXU	1RUZ	R	Ag	0.43866	0.70109	0.52808
4GXU	4GXV	R	Ab	0.11548	0.44464	0.13274
4HX3	1C7K	R	Ez	0.82902	0.47845	0.75096
4HX3	4HWX	R	In	0.48864	0.68455	0.64545
4IZ7	1ERK	M	ST	0.78449	0.38563	0.77717
4IZ7	2LS7	M	ST	0.75818	0.15699	0.40923
4M76	1C3D	R	Ab	0.50880	0.39680	0.45893
4M76	1M1U	R	ST	0.55769	0.34568	0.46013

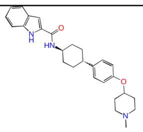
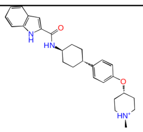
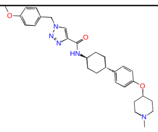
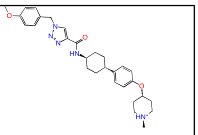
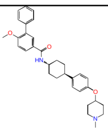
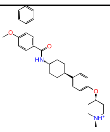
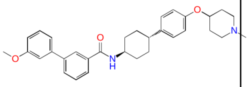
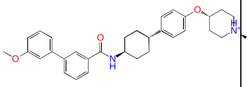
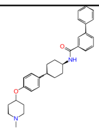
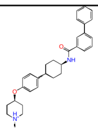
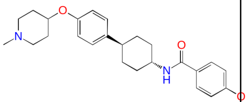
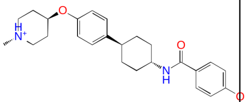
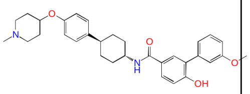
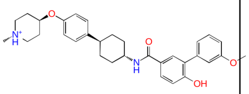
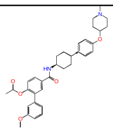
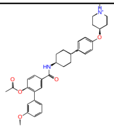
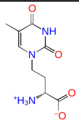
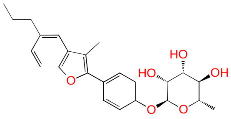
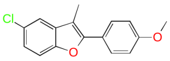
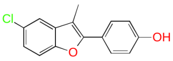
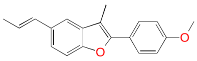
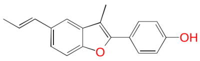
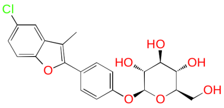
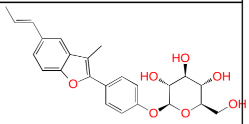
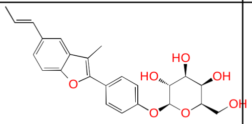
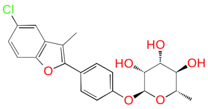
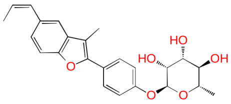
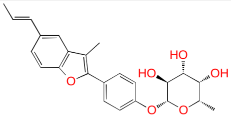
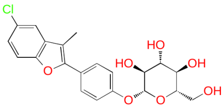
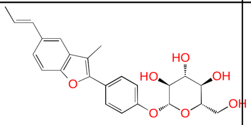
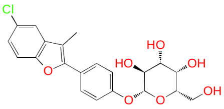
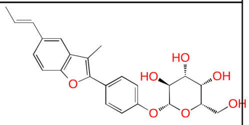
Appendix

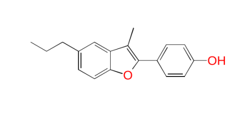
C Molecular structure of all compounds

2D representation of the structure of the ligand used for the ensemble docking scoring. Every molecule is labelled with an entry ID.

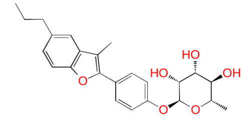
 entry id: 129	 entry id: 130	 entry id: 131	 entry id: 132
 entry id: 133			

 entry id: 1	 entry id: 2	 entry id: 3	 entry id: 4
 entry id: 5	 entry id: 6	 entry id: 7	 entry id: 8
 entry id: 9	 entry id: 10	 entry id: 11	 entry id: 12
 entry id: 13	 entry id: 14	 entry id: 15	 entry id: 16
 entry id: 17	 entry id: 18	 entry id: 19	 entry id: 20
 entry id: 21	 entry id: 22	 entry id: 23	 entry id: 24
 entry id: 25	 entry id: 26	 entry id: 27	 entry id: 28
 entry id: 29	 entry id: 30	 entry id: 31	 entry id: 32

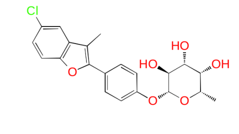
 entry id: 33	 entry id: 34	 entry id: 35	 entry id: 36
 entry id: 37	 entry id: 38	 entry id: 39	 entry id: 40
 entry id: 41	 entry id: 42	 entry id: 43	 entry id: 44
 entry id: 45	 entry id: 46	 entry id: 47	 entry id: 48
 entry id: 49	 entry id: 50	 entry id: 51	 entry id: 52
 entry id: 53	 entry id: 54	 entry id: 55	 entry id: 56
 entry id: 57	 entry id: 58	 entry id: 59	 entry id: 60
 entry id: 61	 entry id: 62	 entry id: 63	 entry id: 64



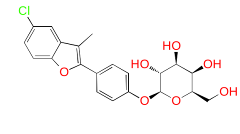
entry id: 65



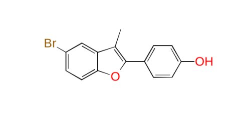
entry id: 66



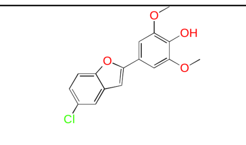
entry id: 67



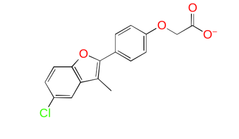
entry id: 68



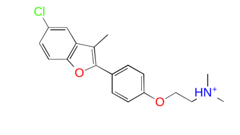
entry id: 69



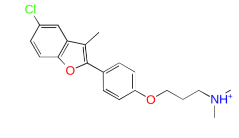
entry id: 70



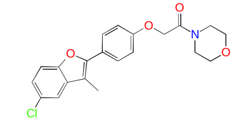
entry id: 71



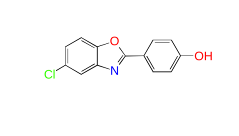
entry id: 72



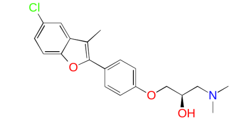
entry id: 73



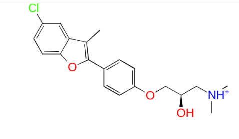
entry id: 74



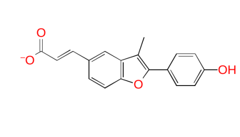
entry id: 75



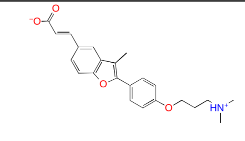
entry id: 76



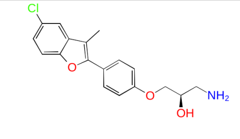
entry id: 77



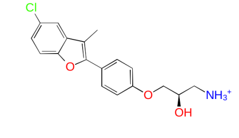
entry id: 78



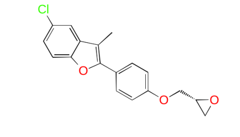
entry id: 79



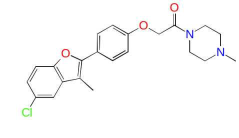
entry id: 80



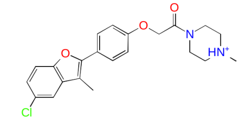
entry id: 81



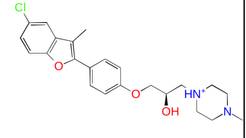
entry id: 82



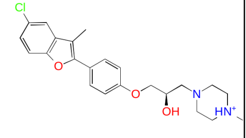
entry id: 83



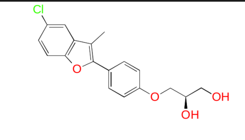
entry id: 84



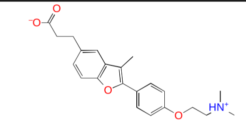
entry id: 85



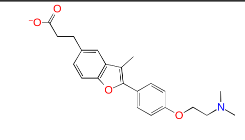
entry id: 86



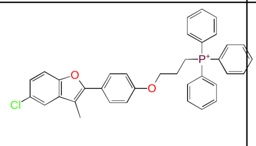
entry id: 87



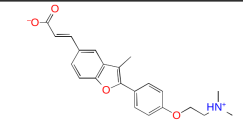
entry id: 88



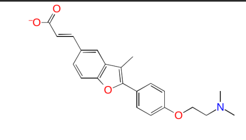
entry id: 89



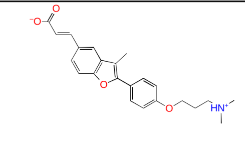
entry id: 90



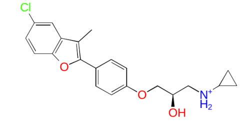
entry id: 91



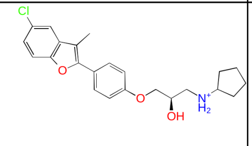
entry id: 92



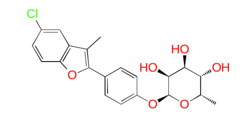
entry id: 93



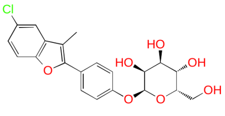
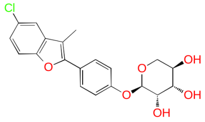
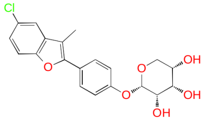
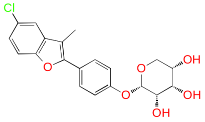
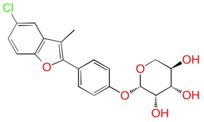
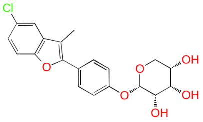
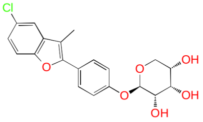
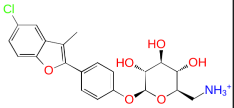
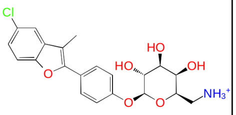
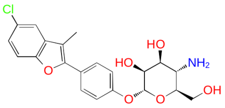
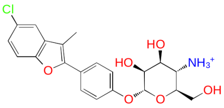
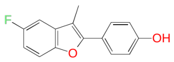
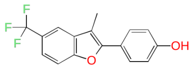
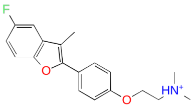
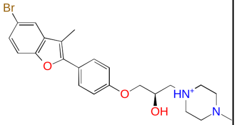
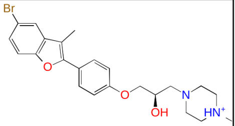
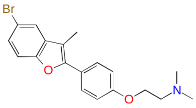
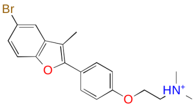
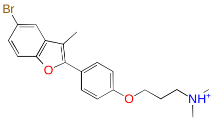
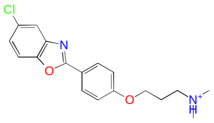
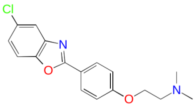
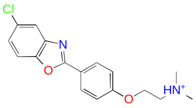
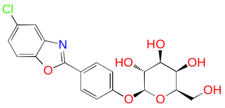
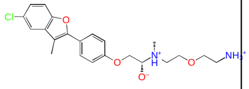
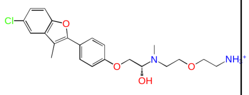
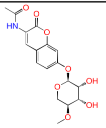
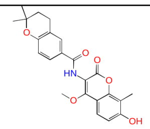
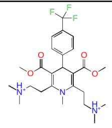
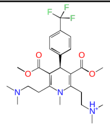
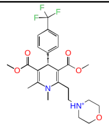
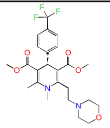
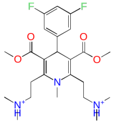
entry id: 94



entry id: 95



entry id: 96

 entry id: 97	 entry id: 98	 entry id: 99	 entry id: 100
 entry id: 101	 entry id: 102	 entry id: 103	 entry id: 104
 entry id: 105	 entry id: 106	 entry id: 107	 entry id: 108
 entry id: 109	 entry id: 110	 entry id: 111	 entry id: 112
 entry id: 113	 entry id: 114	 entry id: 115	 entry id: 116
 entry id: 117	 entry id: 118	 entry id: 119	 entry id: 120
 entry id: 121	 entry id: 122	 entry id: 123	 entry id: 124
 entry id: 125	 entry id: 126	 entry id: 127	 entry id: 128

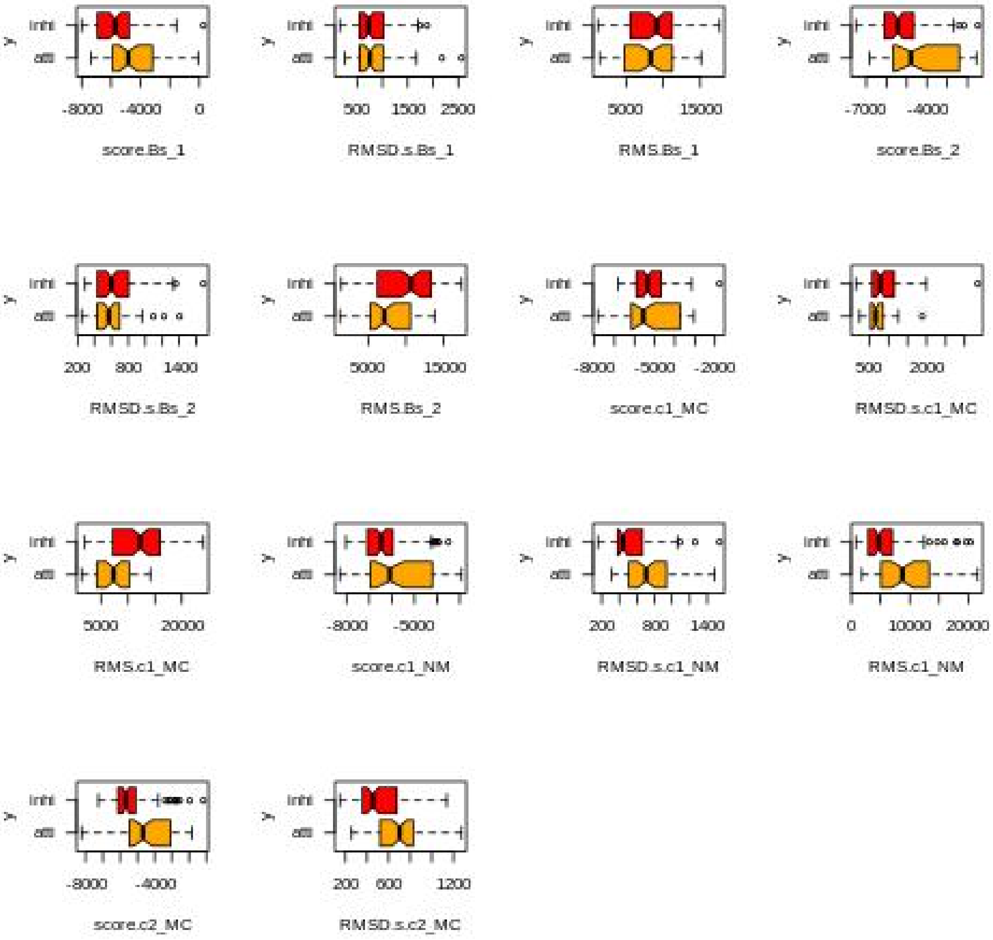
Appendix

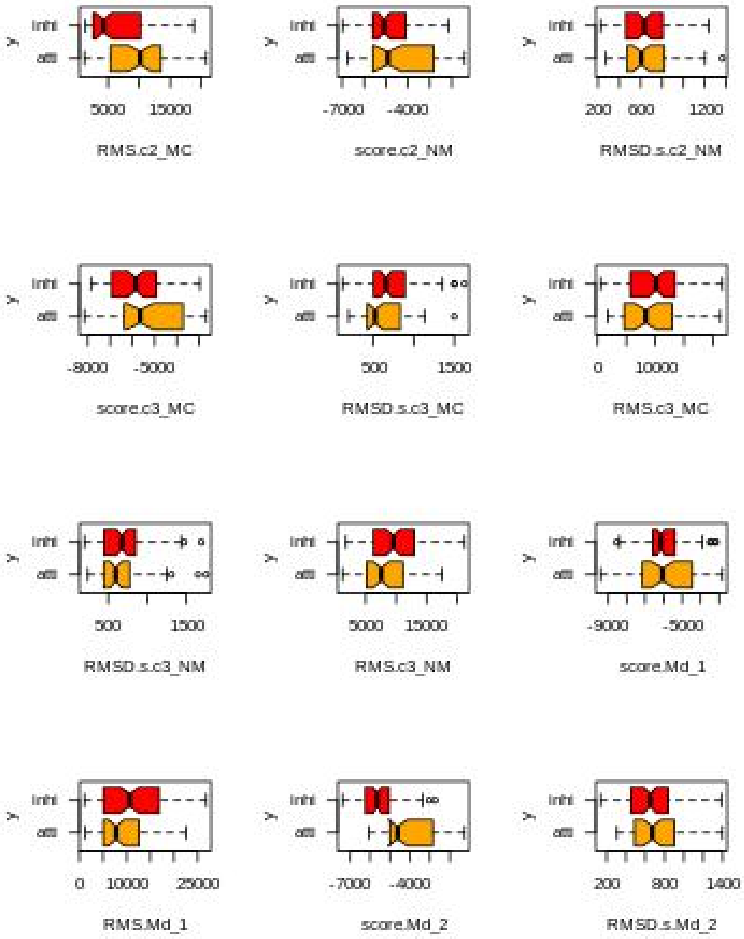
D

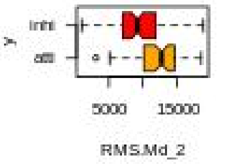
Features comparison

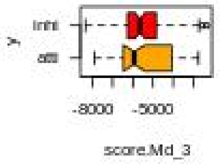
Here are reported the boxplot of the distribution of values for every single feature against the two known ligand classes: inhibitors and activators. Inhibitors are in red and activators are in orange. The legend of the features is:

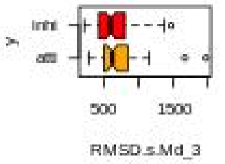
Score	docking score for the best pose.
RMSD	rmsd of the docking score for the ten best poses.
RMS	RMS of atomic position respect to the best pose, averaged over the ten best poses.
Bs	Clustering on the binding site.
Nd	Clustering on the N-term domain.
Md	Clustering on the Middle domain.
NM	Clustering on both N-term and Middle domains.
MC	Clustering on both Middle and C-term domain.
Numbers 1,2,3	First, Second and Third representative structure of the cluster analysis.

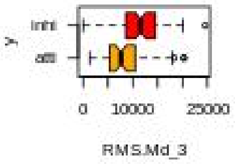

y
inhi
atti
-8000
-4000
0
score.Bs_1
500
1500
2500
RMSD.s.Bs_1
5000
15000
RMS.Bs_1
-7000
-4000
score.Bs_2
200
800
1400
RMSD.s.Bs_2
5000
15000
RMS.Bs_2
-8000
-5000
-2000
score.c1_MC
500
2000
RMSD.s.c1_MC
5000
20000
RMS.c1_MC
-8000
-5000
score.c1_NM
200
800
1400
RMSD.s.c1_NM
0
10000
20000
RMS.c1_NM
-8000
-4000
score.c2_MC
200
600
1200
RMSD.s.c2_MC

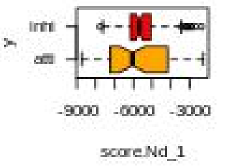

y
infl
atti
5000 15000
RMS.c2_MC
-7000 -4000
score.c2_NM
200 600 1200
RMSD.s.c2_NM
5000 15000
RMS.c2_NM
-8000 -5000
score.c3_MC
500 1500
RMSD.s.c3_MC
0 10000
RMS.c3_MC
-8000 -4000
score.c3_NM
500 1500
RMSD.s.c3_NM
5000 15000
RMS.c3_NM
-9000 -5000
score.Md_1
500 1500
RMSD.s.Md_1
0 10000 25000
RMS.Md_1
-7000 -4000
score.Md_2
200 800 1400
RMSD.s.Md_2

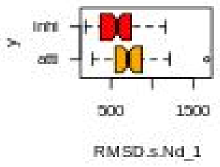

y
infi
atti
5000
15000
RMS.Md_2

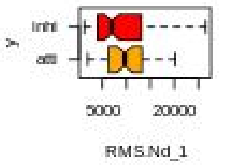

y
infi
atti
-8000
-5000
score.Md_3

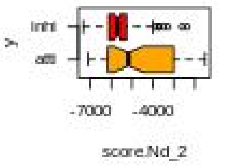

y
infi
atti
500
1500
RMSD.s.Md_3

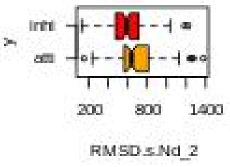

y
infi
atti
0
10000
25000
RMS.Md_3

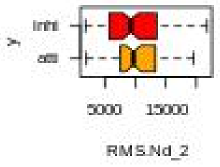

y
infi
atti
-9000
-6000
-3000
score.Nd_1

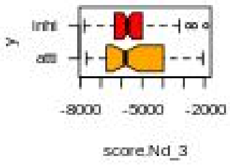

y
infi
atti
500
1500
RMSD.s.Nd_1

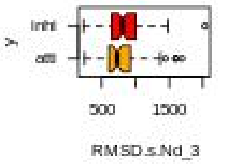

y
infi
atti
5000
20000
RMS.Nd_1

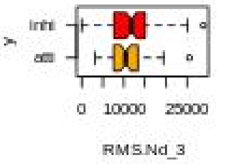

y
infi
atti
-7000
-4000
score.Nd_2


y
infi
atti
200
800
1400
RMSD.s.Nd_2


y
infi
atti
5000
15000
RMS.Nd_2


y
infi
atti
-8000
-5000
-2000
score.Nd_3


y
infi
atti
500
1500
RMSD.s.Nd_3


y
infi
atti
0
10000
25000
RMS.Nd_3

Appendix

E ECFP and SVM comparison

In this table are reported the comparison of the best result obtained by ECFP fingerprint on two clusters with the best ML predictive model obtained by SVM. For SVM the compounds that are correctly predicted are labelled as True, for ECFP we reported the cluster number in which each compound is assigned, the cluster with majority of inhibitors is 1.

ID	Group	Class	SVM	ECFP
1	Z	activator	False	1
2	BP2	inhibitor	True	1
3	BP2	inhibitor	True	1
4	BP2	inhibitor	True	1
5	BP2	inhibitor	True	1
6	BP2	inhibitor	True	1
7	BP2	inhibitor	False	1
8	BP2	inhibitor	True	1
9	BP2	inhibitor	True	1
10	BP1	inhibitor	True	1
11	BP1	inhibitor	True	1
12	BP1	inhibitor	True	1
13	BP1	inhibitor	True	1
14	BP1	inhibitor	True	1
15	BP2	inhibitor	True	1
16	BP2	inhibitor	True	1
17	BP2	inhibitor	True	1
18	BP2	inhibitor	True	1
19	BP2	inhibitor	True	1

Continued on next page

20	BP2	inhibitor	True	1
21	BP2	inhibitor	True	1
22	BP2	inhibitor	True	1
23	BP2	inhibitor	True	1
24	BP2	inhibitor	True	1
25	BP2	inhibitor	True	1
26	BP2	inhibitor	False	1
27	BP2	inhibitor	True	1
28	BP2	inhibitor	False	1
29	BP2	inhibitor	True	1
30	BP2	inhibitor	True	1
31	BP2	inhibitor	True	1
32	BP2	inhibitor	True	1
33	BP2	inhibitor	False	1
34	BP2	inhibitor	True	1
35	BP2	inhibitor	True	1
36	BP2	inhibitor	True	1
37	BP2	inhibitor	True	1
38	BP2	inhibitor	True	1
39	BP2	inhibitor	True	1
40	BP2	inhibitor	True	1
41	BP2	inhibitor	True	1
42	BP2	inhibitor	True	1
43	BP2	inhibitor	True	1
44	BP2	inhibitor	True	1
45	BP2	inhibitor	True	1
46	BP2	inhibitor	True	1
47	BP2	inhibitor	True	1
48	BP2	inhibitor	True	1
49	Ukn	activator	True	1
50	CC	activator	True	2
51	CC	activator	True	2
52	CC	activator	True	2
53	CC	activator	True	2
54	CC	activator	True	2
55	CC	activator	True	2
56	CC	activator	True	2
57	CC	activator	True	2
58	CC	activator	True	2
59	CC	activator	True	2
60	CC	activator	True	2
61	CC	activator	True	2
62	CC	activator	True	2
63	CC	activator	True	2
64	CC	activator	True	2

Continued on next page

65	CC	activator	True	2
66	CC	activator	True	2
67	CC	activator	True	2
68	CC	activator	True	2
69	CC	activator	True	2
70	CC	activator	True	1
71	CC	activator	True	2
72	CC	activator	True	2
73	CC	activator	True	2
74	CC	activator	True	2
75	CC	activator	True	2
76	CC	activator	True	2
77	CC	activator	True	2
78	CC	activator	True	2
79	CC	activator	True	2
80	CC	activator	True	2
81	CC	activator	True	2
82	CC	activator	True	2
83	CC	activator	True	2
84	CC	activator	True	2
85	CC	activator	True	2
86	CC	activator	True	2
87	CC	activator	True	2
88	CC	activator	True	2
89	CC	activator	True	2
90	CC	activator	False	2
91	CC	activator	True	2
92	CC	activator	True	2
93	CC	activator	True	2
94	CC	activator	True	2
95	CC	activator	True	2
96	CC	activator	True	2
97	CC	activator	True	2
98	CC	activator	True	2
99	CC	activator	True	2
100	CC	activator	True	2
101	CC	activator	True	2
102	CC	activator	True	2
103	CC	activator	True	2
104	CC	activator	True	2
105	CC	activator	True	2
106	CC	activator	True	2
107	CC	activator	True	2
108	CC	activator	True	2
109	CC	activator	True	2

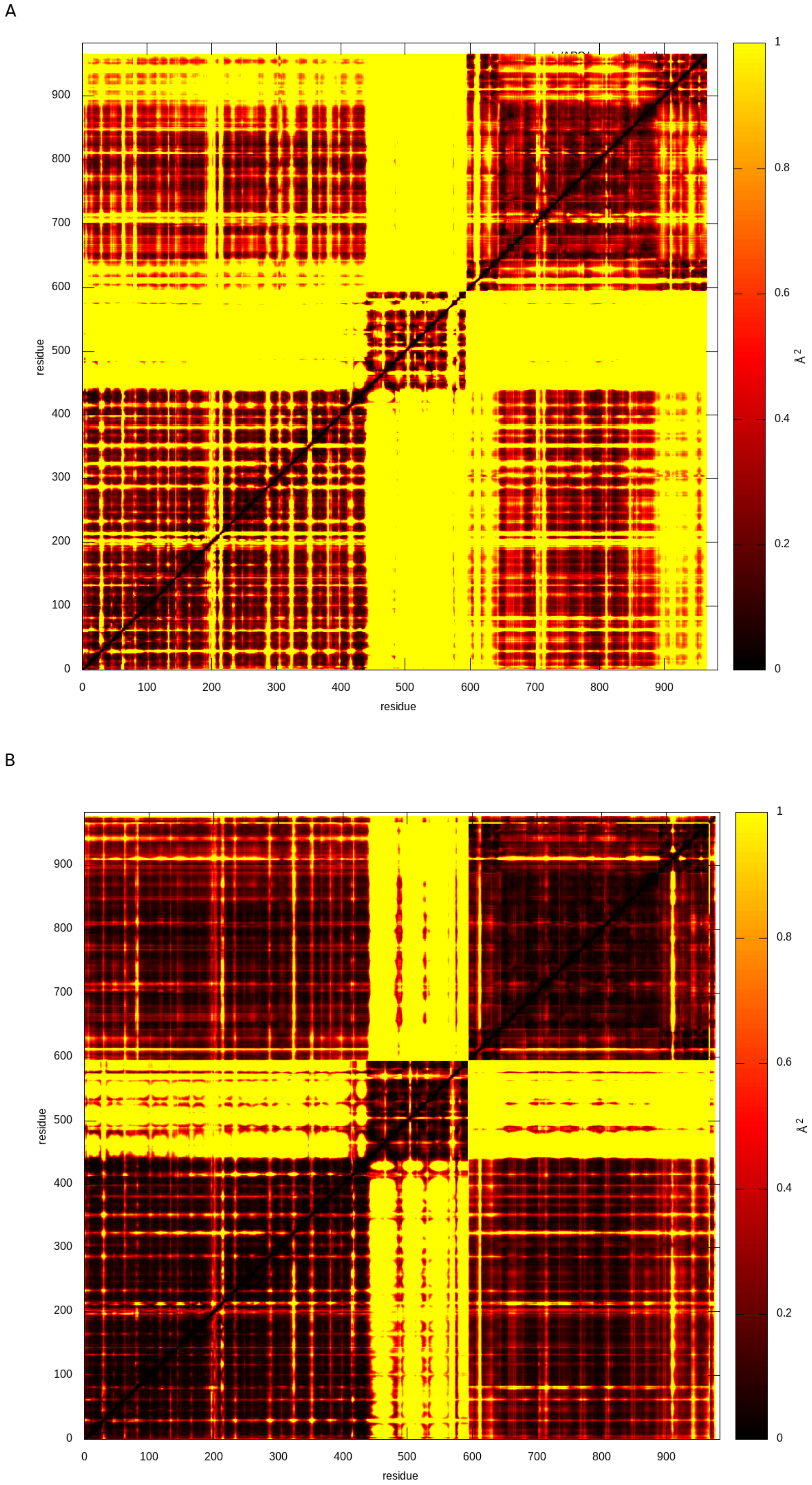
Continued on next page

110	CC	activator	True	2
111	CC	activator	True	2
112	CC	activator	True	2
113	CC	activator	True	2
114	CC	activator	True	2
115	CC	activator	True	2
116	CC	activator	True	2
117	CC	activator	False	2
118	CC	activator	True	2
119	CC	activator	True	2
120	CC	activator	True	2
121	CC	activator	True	2
122	CB	inhibitor	False	1
123	CB	inhibitor	False	1
124	DP	activator	True	1
125	DP	activator	True	1
126	DP	activator	False	1
127	DP	activator	True	1
128	DP	activator	False	1
129	DP	activator	True	1
130	Z	activator	True	1
131	Z	activator	False	1
132	GT	activator	True	1
133	CC	activator	True	2

Appendix

F Distance Fluctuation analyses

The Distance Fluctuation matrix for the APO and LAP state of integrin αvβ6 are computed extracting around 1000 conformations for every replica, the final matrix is obtained averaging over the 4 matrix (one for every replica) for every state. In Figure A there is the matrix for the APO state, in Figure B for the LAP bound conformation.


A
residue
0
100
200
300
400
500
600
700
800
900
residue
0
100
200
300
400
500
600
700
800
900
Å 2
0
0.2
0.4
0.6
0.8
1
B
residue
0
100
200
300
400
500
600
700
800
900
residue
0
100
200
300
400
500
600
700
800
900
Å 2
0
0.2
0.4
0.6
0.8
1

Bibliography

[1] MACCS-II. *MDL Information Systems/Symyx, Santa Clara, CA.*

[2] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J.C. Smith, B. Hess, and E. Lindahl. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 2015.

[3] A.Kolmogorov. Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari*, 4:83–91, 1933.

[4] D.G. Alberg and S.L. Schreiber. Structure-based design of a cyclophilin-calcineurin bridging ligand. *Science*, 8:248–250, 1993.

[5] O. Allnér, L. Nilsson, and A. Villa. Magnesium Ion-Water Coordination and Exchange in Biomolecular Simulations. *Journal of Chemical Theory and Computation*, 8:1493–1502, 2012.

[6] M.O Altman, M. Angel, I. Košík, N.S. Trovão, S.J. Zost, J.S Gibbs, L. Casalino, R.E. Amaro, S.E. Hensley, M.I. Nelson, and J.W. Yewdell. Human Influenza A Virus Hemagglutinin Glycan Evolution Follows a Temporal Pattern to a Glycan Limit. *mBio*, 2019.

[7] R.E. Amaro, J. Baudry, J. Chodera, Ö. Demir, J.A. McCammon, Y. Miao, and J. C. Smith. Ensemble Docking in Drug Discovery. *Biophysical Journal*, 114:2271–2278, 2018.

[8] K. G. Andersen, A. R. W. I. Lipkin, E. C. Holmes, and R. F. Garry. The proximal origin of SARS-CoV-2. *Nature Medicine*, 26:450–452, 2020.

[9] T. W. Anderson and D. A. Darling. Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes. *Annals of Mathematical Statistics*, 23:193–212, 1952.

[10] J. Andreani and R. Guerois. Evolution of protein interactions: from interactomes to interfaces. *Archives of Biochemistry and Biophysics*, 554:65–75, 2014.

[11] N. Andrusier, R. Nussinov, and H.J. Wolfson. FireDock: Fast interaction refinement in molecular docking. *Proteins: Structure, Function and Genetics*, 69:139–159, 2007.

[12] B. Apellániz, N. Huarte, E. Largo, and J.L. Nieva. The Three Lives of Viral Fusion Peptides. *Chemistry and Physics of Lipids*, 181:40–55, 2014.

[13] M.R. Arkin, Y. Tang, and J.A. Wells. Small-Molecule Inhibitors of Protein-Protein Interactions: Progressing toward the Reality. *Chemical Biology*, 21:1102–1114, 2014.

[14] M.R. Arkin and J.A. Wells. Small-Molecule inhibitors of protein- protein interactions: progressing towards the dream. *Nature Reviews Drug Discovery*, 3:301–317, 2004.

[15] F. Bai, F. Morcos, R.R. Cheng, J. Hualiang, and J.N. Onuchic. Elucidating the druggable interface of protein-protein interactions using fragment docking and coevolutionary analysis. *Proceedings of the National Academy of Sciences*, 113:e8051–e8058, 2016.

[16] D. Bajusz, A. Rácz, and K. Héberger. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics*, 20, 2015.

[17] A. Bandyopadhyay and S. Raghavan. Defining the Role of Integrin αvβ6 in Cancer. *Current Drug Targets*, 10:645–652, 2012.

[18] I.I. Baskin, D. Winkler, and I. V. Tetko. A renaissance of neural networks in drug discovery. *Expert Opinion on Drug Discovery*, 11:785–795, 2016.

[19] S. Belouzard, V.C. Chu, and G.R. Whittaker. Activation of the SARS coronavirus spike protein via sequential proteolytic cleavage at two distinct sites. *Proceeding of the National Academy of Sciences*, 106:5871–5876, 2009.

[20] C. Benodand, J. Carlsson, R. Uthayaruban, P. Hwang, J.J. Irwin, A.K. Doak, B.K. Shoichet, E.P. Sablin, and R.J. Fletterick. Structure-based discovery of antagonists of nuclear receptor LRH-1. *Journal of Biological Chemistry*, 288:19830–19844, 2013.

[21] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak. Molecular dynamics with coupling to an external bath. *Journal of Chemical Physics*, 81:3684, 1984.

[22] R.B. Best, W. Zheng, and J. Mittal. Balanced Protein-Water Interactions Improve Properties of Disordered Proteins and Non-Specific Protein Association. *Journal of Chemical Theory and Computation*, 10:5113–5124, 2014.

[23] M. J. Betts and M.J.E. Sternberg. An analysis of conformational changes on protein-protein association: implications for predictive docking . *Protein Engineering, Design and Selection*, 12:271–283, 1999.

[24] G. Bonaccorsi, F. Pierri, M. Cinelli, A. Flori, A. Galeazzi, F. Porcelli, A. L. Schmidt, C. M. Valensise, A. Scala, W. Quattrociocchi, and F. Pammolli. Economic and social consequences of human mobility restrictions under COVID-19. *Proceedings of the National Academy of Sciences*, 117:15530–15535, 2020.

[25] A.L. Bowman, Z. Nikolovska-Coleska, H.Z. Zhong, S.M. Wang, and H.A. Carlson. Small molecule inhibitors of the MDM2-p53 interaction discovered by ensemble-based receptor models. *Journal of the American Chemical Society*, 129:12809–12814, 2007.

[26] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. Charmm: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*, 4:187–217, 1983.

[27] J.D. Bryngelson, J.N. Onuchic, N.D. Socci, and P.G. Wolynes. Funnels, pathways, and the energy landscape of protein folding: A synthesis. *Protein Structure, Function and Bioinformatics*, 21:167–195, 1995.

[28] H.A. Carlson and J.A. McCammon. Accommodating protein flexibility in computational drug design. *Molecular Pharmacology*, 57:213–218, 2000.

[29] L. Casalino, Z. Gaieb, J. A. Goldsmith, C. K. Hjorth, A. C. Dommer, A. M. Harbison, C. A. Fogarty, E. P. Barros, B. C. Taylor, J. S. McLellan, E. Fadda, and R. E. Amaro. Beyond Shielding: The Roles of Glycans in the SARS-CoV-2 Spike Protein. *ACS Central Science*, 6:1722–1734, 2020.

[30] D.A. Case, T.E. Cheatham, T. Darden, H. Gohlke, R. Luo, K.M. Jr Merz, A. Onufriev, C. Simmerling, B. Wang, and R.J. Woods. The Amber biomolecular simulation programs. *Journal of Computational Chemistry*, 26:1668–1688, 2005.

[31] L.C. Cesa, S. Patury, T. Komiyama, A. Ahmad, E. Zuideweg, and J.E. Gestwicki. Inhibitors of Difficult Protein-Protein Interactions Identified by High-Throughput Screening of Multi-protein Complexes. *ACS Chemical Biology*, 8:1988–1997, 2013.

[32] J. P. Changeux. Allostery and the Monod-Wyman-Changeux model after 50 years. *Annual Reviews in Biophysics*, 41:103–133, 2012.

[33] L.S. Cheng, R.E. Amaro, D. Xu, W.W. Li, P.W. Arzberger, and J. A. McCammon. Ensemble-Based Virtual Screening Reveals Potential Novel Antiviral Compounds for Avian Influenza Neuraminidase. *Journal of Medicinal Chemistry*, 51:3878–3894, 2008.

[34] V. C. C. Cheng, S. K. P. Lau, P. C. Y. Woo, and K. Y. Yuen. Severe Acute Respiratory Syndrome Coronavirus as an Agent of Emerging and Reemerging Infection. *Clinical Microbiology Reviews*, 20:660–694, 2007.

[35] C. Chothia and A.M. Lesk. The relation between the divergence of sequence and structure in proteins. *EMBO Journal*, 5:823–826, 1986.

[36] T. Clackson and J.A. Wells. A hot spot of binding energy in a hormone-receptor interface. *Science*, 267:383–386, 1995.

[37] S. Colacino, G. Tiana, and G. Colombo. Similar folds with different stabilization mechanisms: the cases of prion and doppel proteins. *BMC structural Biology*, 6, 2006.

[38] M. Crispin, A.B. Ward, and I.A. Wilson. Structure and Immune Recognition of the HIV Glycan Shield. *Annual Review of Biophysics*, 47:499–523, 2018.

[39] P. Csermely, R. Palotai, and R. Nussinov. Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. *Trends in Biochemical Sciences*, 35:539–546, 2010.

[40] J. Cui, F. Li, and Z. Shi. Origin and evolution of pathogenic coronaviruses. *Nature Reviews Microbiology*, 395:181–192, 2019.

[41] I. D'Annessa, S. Sattin, J.H. Tao, M. Pennati, C. Sanchez-Martin, E. Moroni, A. Rasola, N. Zaffaroni, D.A. Agard, A. Bernardi, and G. Colombo. Design of Allosteric Stimulators of the Hsp90 ATPase as New Anticancer Leads. *Chemistry - A European Journal*, 23:5188–5192, 2017.

[42] T. Darden, D. York, and L. Pedersen. Particle mesh Ewald: An N-log(N) method for Ewald sums in large systems. *Journal of Chemical Physics*, 98:10089–10092, 1993.

[43] X. Daura, K. Gademann, B. Jaun, D. Seebach, W. F. van Gunsteren, and A.E. Mark. Peptide folding: when simulation meets experiment. *Angewante Chemie International Edition*, 38:236–240, 1999.

[44] D. Defays. An efficient algorithm for a complete link method. *The Computer Journal*, 20:364–366, 1977.

[45] A. del Sol, C.J. Tsai, B.Y. Ma, and R. Nussinov. The origin of allosteric functional modulation: multiple pre-existing pathways. *Structure*, 17:1042–1050, 2009.

[46] S.J. deVries and A.M.J.J. Bonvin. How proteins get in touch: interface prediction in the study of biomolecular complexes. *Current Protein and Peptide Sciences*, 9:394–406, 2008.

[47] S. E. Dobbins, V. I. Lesk, and M. J. E. Sternberg. Insights into protein flexibility: The relationship between normal modes and conformational change upon protein-protein docking. *Proceedings of the National Academy of Sciences*, 105:10390–10395, 2008.

[48] C. Dominguez, R. Boelens, and A.M.J.J. Bonvin. HADDOCK: a protein-protein docking approach based on biochemical and/or biophysical information. *Journal of American Chemical Society*, 125:1731–1737, 2003.

[49] X. Dong, N.E. Hudson, C. Lu, and T.A. Springer. Structural determinants of integrin β-subunit specificity for latent TGF-β. *Nature Structural and Molecular Biology*, 21:1091–1096, 2014.

[50] X. Dong, B. Zao, R.E. Iacob, J. Zhu, A.C. Koksal, C. Lu, J.R. Engen, and T.A. Springer. Force interacts with macromolecular structure in activation of TGF-β. *Nature*, 542:55–59, 2017.

[51] X. Dong, B. Zao, F.Y. Lin, C. Lu, B. N. Rogers, and T.A. Springer. High integrin αVβ6 affinity reached by hybrid domain deletion slows ligand-binding on-rate. *Proceedings of the National Academy of Sciences*, 115, 2018.

[52] K.J. Doores, C. Bonomelli, D.J. Harvey, S. Vasiljevic, R.A. Dwek, D.R. Burton, M. Crispin, and C.N. Scanlan. Envelope Glycans of Immunodeficiency Virions Are Almost Entirely Oligomannose Antigens. *Proceeding of the National Academy of Sciences*, 107:13800–13805, 2010.

[53] R. C. Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32:1792–1797, 2004.

[54] F. J. Ehlert. Analysis of Allosterism in Functional Assays. *Journal of Pharmacology and Experimental Therapeutics*, 315:740–754, 2005.

[55] D. Elnatan, M. Betegon, Y. Liu, T. Ramelot, M.A. Kennedy, and D.A. Agard. Symmetry broken and rebroken during the ATP hydrolysis cycle of the mitochondrial Hsp90 TRAP1. *eLife*, 6, 2017.

[56] D. A. Enria, N.J. Fernandez, A.M. Briggiler, S.C. Levis, and J.I. Maiztegui. Importance of dose of neutralising antibodies in treatment of Argentine haemorrhagic fever with immune plasma. *The Lancet*, 324:255–256, 1984.

[57] A. Baum et al. Antibody cocktail to SARS-CoV-2 spike protein prevents rapid mutational escape seen with individual antibodies. *Science*, 369:1014–1018, 2020.

[58] A. Paladino et al. Chemical Perturbation of Oncogenic Protein Folding: from the Prediction of Locally Unstable Structures to the Design of Disruptors of Hsp90-Client Interactions. *Chemistry- A European Journal*, 26:9459–9465, 2020.

[59] A. Rodina et al. The epichaperome is an integrated chaperome network that facilitates tumour survival. *Nature*, 538, 2016.

[60] A.G. Reidenbach et al. Multimodal small-molecule screening for human prion protein binders. *bioRxiv*, 538, 2020.

[61] B. Wang et al. Building a Hybrid Physical-Statistical Classifier for Predicting the Effect of Variants Related to Protein-Drug Interactions. *Structure*, 27, 2019.

[62] C.O. Barnes et al. Structures of Human Antibodies Bound to SARS-CoV-2 Spike Reveal Common Epitopes and Recurrent Features of Antibodies. *Cell*, 182:828–842, 2020.

[63] D. Pinto et al. Cross-neutralization of SARS-CoV-2 by a human monoclonal SARS-CoV antibody. *Nature*, 583:290–295, 2020.

[64] Dan Li et al. A potent human neutralizing antibody Fc-dependently reduces established HBV infections. *eLife*, 6:e26738, 2017.

[65] E. P. Go et al. Comparative Analysis of the Glycosylation Profiles of Membrane-Anchored HIV-1 Envelope Glycoprotein Trimers and Soluble Gp140. *Biochimica et Biophysica Acta - General Subjects*, 89:8245–8257, 2015.

[66] I. Casciuc et al. Pros and cons of virtual screening based on public “ Big Data ”: In silico mining for new bromodomain inhibitors. *European Journal of Medicinal Chemistry*, 165:258–272, 2019.

[67] I. R. Taylor et al. Tryptophan scanning mutagenesis as a way to mimic the compound-bound state and probe the selectivity of allosteric inhibitors in cells. *Chemical Science*, 11:1892–1904, 2020.

[68] J. Lyu et al. Ultra-large library docking for discovering new chemotypes. *Nature*, 566:224–229, 2019.

[69] J.F.W. Chan et al. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *the Lancet*, 395:514–523, 2020.

[70] L.J. Gourlay et al. Exploiting the Burkholderia pseudomallei Acute Phase Antigen BPSL2765 for Structure-Based Epitope Discovery/Design in Structural Vaccinology. *Chemical Biology*, 20:1147–1156, 2013.

[71] M. M. Giuliani et al. A universal vaccine for serogroup B meningococcus. *Proceeding of the National Academy of Sciences*, 103:10834–10839, 2006.

[72] M. Yuan et al. Structural basis of a shared antibody response to SARS-CoV-2. *Science*, 369:1119–1123, 2020.

[73] M.R. Prince et al. Allosteric Modulation of the Cannabinoid CB1 Receptor. *Molecular Pharmacology*, 68:1484–1495, 2005.

[74] P. Acharya et al. A glycan cluster on the SARS-CoV-2 spike ectodomain is recognized by Fab-dimerized glycan-reactive antibodies. *bioRxiv*, 2020.

[75] P. Schneider et al. Rethinking drug design in the artificial intelligence era. *Nature Reviews Drug Discovery*, 19:353–364, 2020.

[76] Q. Wang et al. Structural and Functional Basis of SARS-CoV-2 Entry by Using Human ACE2. *Cell*, 181, 2020.

[77] R.A. Friesner et al. Glide: A new approach for rapid, accurate docking and scoring. Method and assessment of docking accuracy. *Journal of Medicinal Chemistry*, 47:1739–1749, 2004.

[78] S. J. Wodak et al. Allostery in Its Many Disguises: From Theory to Applications. *Structure*, 27:566–578, 2019.

[79] Sattin et al. Activation of Hsp90 Enzymatic Activity and Conformational Dynamics through Rationally Designed Allosteric Ligands. *Chemistry - A European Journal*, 21:13598–13608, 2015.

[80] T. Vreven et al. Integrating Cross-Linking Experiments with Ab Initio Protein-Protein Docking. *Journal of Molecular Biology*, 430:1814–1828, 2018.

[81] X. Chi et al. A neutralizing human antibody binds to the N-terminal domain of the Spike protein of SARS-CoV-2. *Science*, 369:650–655, 2020.

[82] X. Wei et al. Antibody Neutralization and Escape by HIV-1. *Nature*, 422:307–312, 2003.

[83] Y. Cao et al. Potent Neutralizing Antibodies against SARS-CoV-2 Identified by High-Throughput Single-Cell Sequencing of Convalescent Patients' B Cells. *Cell*, 182:73–84, 2020.

[84] Y. Watanabe et al. Vulnerabilities in Coronavirus Glycan Shields despite Extensive Glycosylation. *Nature Communications*, 11, 2020.

[85] Z. Zheng et al. Monoclonal antibodies for the S2 subunit of spike of SARS-CoV-1 cross-react with the newly-emerged SARS-CoV-2. *Eurosurveillance*, 25, 2020.

[86] J.J. Falke. A Moving Story. *Science*, 22:1480–1481, 2002.

[87] K.A. Feenstra, B. Hess, and H.J.C. Berendsen. Improving Efficiency of Large Time-scale Molecular Dynamics Simulations of Hydrogen-rich Systems. *Journal of Chemical Physics*, 20:786–798, 1999.

[88] M. Ferraro, I. D'Annessa, E. Moroni, G. Morra, A. Paladino, S. Rinaldi, F. Compostella, and G. Colombo. Allosteric modulators of Hsp90 and Hsp70: Dynamics meets Function through Structure-Based Drug Design. *Journal Medicinal Chemistry*, 62:60–87, 2019.

[89] A. Finka and P. Goloubinoff. Proteomic data from human cell cultures refine mechanisms of chaperone-mediated protein homeostasis. *Cell Stress Chaperones*, 18:591–605, 2013.

[90] S. Fiorucci and M. Zacharias. Prediction of protein-protein interaction sites using electrostatic desolvation profiles. *Biophysical Journal*, 98:1921–1930, 2010.

[91] E. Fisher. Einfluss der Configuration auf die Wirkung der Enzyme. *Berichte der deutschen chemischen Gesellschaft*, 27:2985–2993, 1894.

[92] J.M. Flynn, P. Mishra, and D.N.A. Bolon. Mechanistic asymmetry in Hsp90 dimers. *Journal of Molecular Biology*, 2015.

[93] H. Frauenfelder, S.G. Sligar, and P.G. Wolynes. The energy landscapes and motions of proteins. *Science*, 13:1598–1603, 1991.

[94] J. Garcia-Garcia, S. Schleker, J. Klein-Seetharaman, and B. Oliva. BIPS: BIANA Interolog Prediction Server. A tool for protein-protein interaction inference. *Nucleic Acids Research*, 40:W147–W151, 2012.

[95] A. Genoni, G. Morra, and G. Colombo. Identification of Domains in Protein Structures from the Analysis of Intramolecular Interactions. *Journal of Physical Chemistry B*, 116:3331–3343, 2012.

[96] O. Grant and R.J. Woods. Glycosylated Swiss-model molecular dynamics trajectory of SARS-CoV-2 spike glycoprotein. *Figshare*, August 5, 2020.

[97] O. C. Grant, D. Montgomery, K. Ito, and R. J. Woods. Analysis of the SARS-CoV-2 spike protein glycan shield: implications for immune recognition. *biorxiv*, 2020.

[98] N. Halabi, O. Rivoire, S. Leibler, and R. Ranganathan. Protein sectors: evolutionary units of three-dimensional structure. *Cell*, 138:774–786, 2009.

[99] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, New York, 2001.

[100] R. Henderson, R.J. Edwards, K. Mansouri, K. Janowska, V. Stalls, M. Kopp, B.F. Haynes, and P. Acharya. Glycans on the SARS-CoV-2 Spike Control the Receptor Binding Domain Conformation. *bioRxiv*, 2020.

[101] S. Henikoff and J.G. Henikoff. A statistical method for evaluating systematic relationship. *Proceeding of National Academy of Sciences*, 89:10915–10919, 1992.

[102] B. Hess. P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation. *Journal of Chemical Theory and Computation*, 4:116–122, 2008.

[103] B. Hess, H. Bekker, J.G.E.M. Fraaije, and H.J.C. Berendsen. A linear constraint solver for molecular simulations. *Journal of Computational Chemistry*, 198:1463–1472, 1997.

[104] M. Hessling, K. Richter, and J. Buchner. Dissection of the ATP-induced conformational cycle of the molecular chaperone Hsp90. *Nature Structural and Molecular Biology*, 16:287–293, 2009.

[105] T.K. Ho. Random decision forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition*, 1995.

[106] A.L. Hopkins. Network pharmacology: the next paradigm in drug discovery. *Nature Chemical Biology*, 4:682–690, 2008.

[107] W. Humphrey, A. Dalke, and K. Schulten. VMD: Visual molecular dynamics. *Journal of Molecular Graphics*, 14:33–38, 1996.

[108] R. Hynes. Integrins: Bidirectional, allosteric signaling machines. *Cell*, 110:673–687, 2002.

[109] Y. Bengio I. Goodfellow and A. Courville. *Deep Learning.* MIT Press, 2016. `http://www.deeplearningbook.org`.

[110] D. E. Koshland Jr. Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proceedings of the National Academy of Sciences*, 44:98–104, 1958.

[111] E. Katchalski-Katzir, I. Shariv, M. Eisenstein, A.A. Friesem, C. Aflalo, and I.A. Vakser. Molecular Surface Recognition: Determination of Geometric Fit Between Proteins and Their Ligands by Correlation Techniques. *Proceedings of the National Academy of Sciences*, 89:2195–2199, 1992.

[112] S. Kerrien, B. Aranda, L. Breuza, A. Bridge, F. Broakes-Carter, C. Chen, M. Duesbury, M. Dumousseau, M. Feuermann, U. Hinz, C. Jandrasits, R.C. Jimenez, J. Khadake, U. Mahadevan, P. Masson, I. Pedruzzi, E. Pfeiffenberger, P. Porras, A. Raghunath, B. Roechert, S. Orchard, and H. Hermjakob. The IntAct molecular interaction database in 2012. *Nucleic Acid Research*, 40:D841–D856, 2012.

[113] O. Keskin, N. Tuncbag, A. Gursoy, and F.L. Gervasio. Predicting Protein-Protein Interactions from the Molecular to the Proteome Level. *Chemical Reviews*, 116:4884–4909, 2016.

[114] K.N. Kirschner, A.B. Yongye, S.M. Tschampel, J. González-Outeiriñ, C.R. Daniels, B.L. Foley, and R.J. Woods. GLYCAM06: A generalizable biomolecular force field. Carbohydrates. *Journal of Computational Chemistry*, 29:622–655, 2008.

[115] A. Koutsoukas, K.J. Monaghan, X. Li, and J. Huan. Deep-learning: investigating deep neural networks hyper-parameters and comparison of performance to shallow methods for modeling bioactivity data. *Journal of Cheminformatics*, 9, 2017.

[116] D.J. Kozuch, F.H. Stillinger, and P.G. Debenedetti. Combined molecular dynamics and neural network method for predicting protein antifreeze activity. *Proceedings of the National Academy of Sciences*, 115:13252–13257, 2018.

[117] S. Kumar, B. Ma, C.J. Tsai, N. Sinha, and R. Nussinov. Folding and binding cascades: Dynamic landscapes and population shifts. *Protein Science*, 9:9–19, 2000.

[118] P.J. Kundrotas, Z. Zhu, J. Janin, and I.A. Vakser. Templates are available to model nearly all complexes of structurally characterized proteins. *Proceedings of the National Academy of Sciences*, 109:9438–9441, 2012.

[119] K. Kupferschmidt and J. Cohen. Race to find COVID-19 treatments accelerates. *Science*, 367, 2020.

[120] Y.Y. Kuttner and S. Engel. Protein Hot Spots: The Islands of Stability. *Journal of Molecular Biology*, 415:419–428, 2012.

[121] Y.Y. Kuttner and S. Engel. Complementarity of stability patches at the interfaces of protein complexes: Implication for the structural organization of energetic hot spots. *Proteins: Structure, Function, Genetics*, 86:229–236, 2018.

[122] B.T. Lai, N.W. Chin, A.E. Stanek, W. Keh, and K.W. Lanks. Quantitation and intracellular localization of the 85K heat shock protein by using monoclonal and polyclonal antibodies. *Molecular and Celluar Biology*, 4:2802–2810, 1984.

[123] O.F. Lange, N.A. Lakomek, C. Farès, G.F. Schröder, K.F.A. Walter, S. Becker, J. Meiler, H. Grubmüller, C. Griesinger, and B.L. de Groot. Recognition Dynamics Up to Microseconds Revealed from an RDC-Derived Ubiquitin Ensemble in Solution. *Science*, 13:1471–1475, 2008.

[124] T. Lazaridis and M. Karplus. "New view" of protein folding reconciled with the old through multiple unfolding simulations. *Protein Engineering, Design and Selection*, 278:1928–1931, 1997.

[125] J. Lee, M. Natarajan, V.C. Nashine, M. Socolich, T. Vo, W.P.Russ, S.J. Benkovic, and R. Ranganathan. Surface sites for engineering allosteric control in proteins. *Science*, 322:438–441, 2008.

[126] P. Leff. The two-state model of receptor activation. *Trends Pharmacological Sciences*, 16:89–97, 1995.

[127] G. Li and E. De Clercq. Therapeutic options for the 2019 novel coronavirus (2019-nCoV). *Nature Review on Drug Discovery*, 19:149–150, 2020.

[128] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J.L. Klepeis, R.O. Dror, and D.E. Shaw. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins*, 78, 2010.

[129] E. Lionta, G. Spyrou, D. K. Vassilatis, and Z. Cournia. Structure-based virtual screening for drug discovery: principles, applications and recent advances. *Current Topics in Medicinal Chemistry*, 14:1923–1938, 2014.

[130] S.W. Lockless and R. Ranganathan. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, 286:285–299, 1999.

[131] R.C: Lua, D.C. Marciano, P. Katsonis, A.K. Adikesavan, A.D. Wilkins, and O. Lichtarge. Prediction and redesign of protein-protein interactions. *Progress in Biophysics and Molecular Biology*, 116:194–202, 2014.

[132] B. Ma, S. Kumar, C.J. Tsai, and R. Nussinov. Folding funnels and binding mechanisms. *Protein Engineering, Design and Selection*, 12:713–720, 1999.

[133] J.A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K.E. Hauser, and C. Simmerling. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *Journal of Chemical Theory and Computation*, 11:3696–3713, 2015.

[134] M.G. Marcu, T.W. Schulte, and L. Neckers. Novobiocin and related coumarins and depletion of heat shock protein 90-dependent signaling proteins. *Journal of the National Cancer Institute*, 92:242–248, 2000.

[135] E. Mashiach, D. Schneidman-Duhovny, A. Peri, Y. Shavit, R. Nussinov, and H.J. Wolfson. An integrated suite of fast docking algorithms. *Protein Structure, Function and Bioinformatics*, 78:3197–3204, 2010.

[136] A. Mayr, G. Klambauer, T. Unterthiner, and S. Hochreiter. DeepTox: Toxicity Prediction using Deep Learning. *Frontiers in Environmental Science*, 2016.

[137] K.L. Meagher, L. T. Redman, and H. A. Carlson. Development of polyphosphate parameters for use with the AMBER force field. *Journal of Computational Chemistry*, 24:1016–1025, 2003.

[138] I. Mihalek, I. Res, and O. Lichtarge. A family of evolution-entropy hybrid methods for ranking protein residues by importance. *Journal of Molecular Biology*, 336:1265–1282, 2004.

[139] L. Mirny, V.I. Abkevich, and E. Shakhnovich. How evolution makes proteins fold quickly. *Proceedings of the National Academy of Sciences*, 95:4976–4981, 1998.

[140] J. Monod. From Enzymatic Adaptation to Allosteric Transitions. *Science*, 28:475–483, 1966.

[141] I.S. Moreira, P.I. Koukos, R. Melo, J.G. Almeida, A.J. Preto, J. Schaarschmidt, M. Trellet, Z.H. Gümüç, J. Costa, and A.M.J.J Bonvin. SpotOn: High Accuracy Identification of Protein-Protein Interface Hot-Spots. *Scientific Reports*, 7:8007, 2017.

[142] E. Moroni, H. Zhao, B.S. Blagg, and G. Colombo. Exploiting Conformational Dynamics in Drug Discovery: Design of C-Terminal Inhibitors of Hsp90 with Improved Activities. *Journal of Chemical Information and Modelling*, 2014.

[143] G. Morra and G. Colombo. Relationship between energy distribution and fold stability: Insights from molecular dynamics simulations of native and mutant proteins. *Proteins*, 72:660–672, 2008.

[144] G. Morra, M.A.C. Neves, C.J. Plescia, S. Tsustsumi, L. Neckers, G. Verkhivker, D.C. Altieri, and G. Colombo. Dynamics-Based Discovery of Allosteric Inhibitors: Selection of New Ligands for the C-terminal Domain of Hsp90. *Journal of Chemical Theory and Computation*, 6:2978–2989, 2010.

[145] J.A. Morrone, J.K. Weber, T. Huynh, H. Luo, and W.D. Cornell. Combining Docking Pose Rank and Structure with Deep Learning Improves Protein–Ligand Binding Mode Prediction over a Baseline Docking Approach. *Journal of Chemical Information and Modeling*, 60:4170–4179, 2020.

[146] R. Mosca, A. Céol, and P. Aloy. Interactome3D: adding structural details to protein networks. *Nature Methods*, 10:47–53, 2013.

[147] M.T. Nelson, W. Humphrey, A. Gursoy, A. Dalke, L.V. Kalé, R.D. Skeel, and K. Schulten. Namd: a parallel, object-oriented molecular dynamics program. *The International Journal of High Performance Computing Applications*, 10, 1996.

[148] R. Nussinov, C.J. Tsai, and H. Jang. Protein ensembles link genotype to phenotype. *PLoS Computational Biology*, 15, 2019.

[149] A. Onufriev, D. Bashford, and D.A. Case. Modification of the Generalized Born Model Suitable for Macromolecules. *Journal of Physical Chemistry B*, 104:3712–3720, 2000.

[150] S. Ovchinnikov, H. Kamisetty, and D. Baker. Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *eLife*, 3, 2014.

[151] J.P. Overington, B. Al-Lazikani, and A. L. Hopkins. How many drug targets are there? *Nature Reviews Drug Discovery*, 5:993–996, 2006.

[152] C. Pallara, B. Jiménez-García, M. Romero, I.H. Moal, and J. Fernández-Recio. pyDock scoring for the new modeling challenges in docking: Protein-peptide, homo-multimers, and domain-domain interactions. *Protein Structure, Function and Bioinformatics*, 85:487–496, 2017.

[153] F. Pazos, M. Helmer-Citterich, G. Ausiello, and A. Valencia. Correlated mutations contain information about protein-protein interaction. *Journal of Molecular Biology*, 271:511–523, 1997.

[154] A. S. Perelson and G. Weisbuch. Immunology for physicists. *Reviews of Modern Physics*, 69:1219–1268., 1997.

[155] C. Peri, P. Gagni, F. Combi, A. Gori, M. Chiari, R. Longhi, M. Cretich, and G. Colombo*. Rational Epitope Design for Protein Targeting. *ACS Chemical Biology*, 8:397–404, 2013.

[156] T. Philipson. *Handbook of Health Economics.* vol 1. Elsevier, 2000.

[157] F. Piazza and Y.H. Sanejouand. Discrete breathers in protein structures. *Physical Biology*, 5, 2008.

[158] F. Piazza and Y.H. Sanejouand. Long-range energy transfer in proteins. *Physical Biology*, 6, 2009.

[159] R. Pricer, J.E. Gestwicki, and A. K. Mapp. From Fuzzy to Function: The New Frontier of Protein-Protein Interactions. *Accounts of Chemical Research*, 50:584–589, 2017.

[160] C. Prodromou. The "active life" of Hsp90 complexes. *Biochimica et Biophysica Acta*, 1823:614–623, 2012.

[161] C. Prodromou, B. Panaretou, S. Chohan, G. Siligardi, R. O'Brien, J.E. Ladbury, S.M. Roe, P.W. Piper, and L.H. Pearl. The ATPase cycle of Hsp90 drives a molecular "clamp" via transient dimerization of the N-terminal domains. *The EMBO Journal*, 19:4383–4392, 2000.

[162] R. Rappuoli. Reverse vaccinology. *Current Opinion in Microbiology*, 3:445–450, 2000.

[163] A. Rasola. HSP90 proteins in the scenario of tumor complexity. *Oncotarget*, 8:20521–20522, 2017.

[164] C. Ratzke, F. Berkemeier, and Thorsten Hugel. Heat shock protein 90's mechanochemical cycle is dominated by thermal fluctuations. *Proceedings of National Academy of Sciences*, 109:161–166, 2012.

[165] S. Rinaldi, V.A. Assimon, Z.T. Young, G. Morra, H. Shao, I.R. Taylor, J.E. Gestwicki, and G. Colombo. A Local Allosteric Network in Heat Shock Protein 70 (Hsp70) Links Inhibitor Binding to Enzyme Activity and Distal Protein-Protein Interactions. *ACS Chemical Biology*, 13:3142–3152, 2018.

[166] I. B. Robertson and D. B. Rifkin. Unchaining the beast; insights from structural and evolutionary studies on TGFβ secretion, sequestration, and activation. *Cytokine and Growth Factor Reviews*, 24:355–372, 2013.

[167] D. Rogers and M. Hahn. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling*, 50:742–754, 2010.

[168] S. Romagnoli, A. Peris, A. R. De Gaudio, and P. Geppetti. SARS-CoV-2 and COVID-19: From the Bench to the Bedside. *Physiological Reviews*, 395:1455–1466, 2020.

[169] N.D. Rubinstein, I.Mayrose, D. Halperin, D. Yekutieli, J.M. Gershoni, and T. Pupko. Computational characterization of B-cell epitopes. *Molecular Immunology*, 45:3477–3489, 2008.

[170] J. Sabelko, J. Ervin, and M. Gruebele. Observation of strange kinetics in protein folding. *Proceedings of the National Academy of Sciences*, 96:6031–6036, 1999.

[171] C. Sanchez-Martin, E. Moroni, M. Ferraro, C. Laquatra, G. Cannino, I. Masgras, A. Negro, P. Quadrelli, A. Rasola, and G. Colombo. Rational Design of Allosteric and Selective Inhibitors of the Molecular Chaperone TRAP1. *Cell Reports*, 31:107531., 2020.

[172] A. V. Savin and L. I. Manevitch. Discrete breathers in a polyethylene chain. *Physical Review B*, 67, 2003.

[173] G. Scarabelli, G. Morra, and G. Colombo. Predicting interaction sited from the energetics of isolated proteins: a new approach to epitope mapping. *Biophysical Journal*, 98:1966–1975, 2010.

[174] C. E. M. Schindler, I.C. de Beauchene, S.J. de Vries, and M. Zacharias. Protein-protein and peptide-protein docking and refinement using ATTRACT in CAPRI. *Proteins: Structure, Function and Genetics*, 85:391–398, 2017.

[175] F.H. Schopf, M.M. Biebl, and J. Buchner. The HSP90 chaperone machinery. *Nature Reviews Molecular Cell Biology*, 18:345–360, 2017.

[176] B. Schuler and W.A. Eaton. Protein folding studied by single molecule FRET. *Current opinion in structural biology*, 18:16–26, 2008.

[177] S. A. Serapian and G. Colombo. Designing Molecular Spanners to Throw in the Protein Networks. *Chemistry - A European Journal*, 26:4656–4670, 2020.

[178] C.E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 3:379–423, 1948.

[179] L. Shrestha, H.J. Patel, and G. Chiosis. Chemical Tools to Investigate Mechanisms Associated with HSP90 and HSP70 in Disease. *Cell Chemical Biology*, 23:158–172, 2016.

[180] T. Siebenmorgen and M. Zacharias. Computational prediction of protein-protein binding affinities. *WIRES Computational Molecular Science*, 10, 2020.

[181] M. Sikora, S. von Bülow, F.E.C. Blanc, M. Gecht, R. Covino, and G. Hummer. Map of SARS-CoV-2 spike epitopes not shielded by glycans. *biorxiv*, 2020.

[182] R.R. Sokal. A statistical method for evaluating systematic relationship. *University of Kansas Science Bulletin*, 28:1409–1438, 1958.

[183] A. Spinello, A. Saltalamacchia, and A. Magistrato. Is the Rigidity of SARS-CoV-2 Spike Receptor-Binding Motif the Hallmark for Its Enhanced Infectivity? Insights from All-Atom Simulations. *Journal of Physical Chemistry Letters*, 11:4785–4790, 2020.

[184] A. Stein, M. Rueda, A. Panjkovich, M. Orozco, and P. Aloy. A Systematic Study of the Energetics Involved in Structural Changes upon Association and Connectivity in Protein Interaction Networks. *Structure*, 19:881–889, 2011.

[185] L. Stella, A.M. Caccuri, N. Rosato, M. Nicotra, M. Lo Bello, F. De Matteis, A.P. Mazzetti, G. Federici, and G. Ricci. Flexibility of Helix 2 in the Human Glutathione Transferase P1-1 Time-Resolved Fluorescence Spectroscopy. *The Journal of Biological Chemistry*, 273:23267–23273, 1998.

[186] J. H. Strauss and E. G. Strauss. *Viruses and human disease.* Academic Press, 2002.

[187] M. M. Sultan, H. K. Wayment-Steele, and V. S. Pande. Transferable Neural Networks for Enhanced Sampling of Protein Dynamics. *Journal of Chemical Theory and Computation*, 14:1887–1894, 2018.

[188] L. Sutto, S. Marsili, A. Valencia, and F.L. Gervasio. From residue coevolution to protein conformational ensembles and functional dynamics. *Proceedings of the National Academy of Sciences*, 112:13576, 2015.

[189] A. Szilagyi, R. Nussinov, and P. Csermely. Allo-Network Drugs: Extension of the Allosteric Drug Concept to Protein-Protein Interaction and Signaling Networks. *Current Topic in Medicinal Chemistry*, 13, 2013.

[190] T. Tang, M. Bidon, J.A. Jaimes, G.R. Whittaker, and S. Daniel. Coronavirus membrane fusion mechanism offers a potential target for antiviral development. *Antiviral Research*, 178, 2020.

[191] G. Thomas, I. Hart, P. Speight, and J. Marshall. Binding of TGF-β1 latency-associated peptide (LAP) to αvβ6 integrin modulates behaviour of squamous carcinoma cells. *British Journal of Cancer*, 24:355–372, 2002.

[192] S. Tian, H. Sun, P. Pan, D. Li, X. Zhen, Y. Li, and T. Hou. Assessing an Ensemble Docking-Based Virtual Screening Strategy for Kinase Targets by Considering Protein Flexibility. *Journal of Chemical Information and Modeling*, 54:2664–2679, 2014.

[193] G. Tiana, F. Simona, G.M.S. De Mori, R.A. Broglia, and G. Colombo. Understanding the determinants of folding and stability of small proteins from their energetics. *Protein Science*, 13:113–124, 2004.

[194] M. A. Tortorici and D. Veesler. Chapter four - structural insights into coronavirus entry. In Félix A. Rey, editor, *Complementary Strategies to Understand Virus Structure and Function*, volume 105 of *Advances in Virus Research*, pages 93 – 116. Academic Press, 2019.

[195] A. Tramontano. The computational prediction of protein assemblies. *Current Opinion in Structural Biology*, 46:170–175, 2017.

[196] C.P. Trémaux. Proc. French Engineer of the Telegraph in Public Conference. *Annals academic - École Polytechnique of Paris (X:1876)*, pages 1859–1882, 2010.

[197] C.J. Tsai, B. Ma, and R. Nussinov. Folding and binding cascades: shifts in energy landscapes. *Proceedings of the National Academy of Sciences*, 31:9970–9972, 1999.

[198] C.J. Tsai and R. Nussinov. A Unified View of "How Allostery Works". *PLOS Computational Biology*, 10, 2014.

[199] J. van der Greef and R.N. McBurney. Rescuing drug discovery: in vivo systems pathology and systems pharmacology. *Nature Reviews Drug Discovery*, 4:961–967, 2005.

[200] A. Vangone and A.M.J.J Bonvin. Contacts-based prediction of binding affinity in protein-protein complexes. *eLife*, 4, 2015.

[201] K.A. Verba, R.Y.R. Wang, A. Arakawa, Y. Liu, M. Shirouzu, S. Yokoyama, and D.A. Agard. Atomic structure of Hsp90-Cdc37-Cdk4 reveals that Hsp90 traps and stabilizes an unfolded kinase. *Science*, 352:1542–1547, 2016.

[202] T. Vreven, I.H. Moal, A. Vangone, B.G. Pierce, P. Kastritis, M. Torchala, R. Chaleil, B. Jiménez-García, P.A. Bates, J. Fernandez-Recio, A.M.J.J. Bonvin, and Z. Weng. Updates to the integrated protein-protein interaction benchmarks: Docking benchmark version 5 and affinity benchmark version 2. *Journal of Molecular Biology*, 427:3031–3041, 2015.

[203] J.R. Wagner, C.T. Lee, J.D. Durrant, R.D. Malmstrom, V.A. Feher, and R.E. Amaro. Emerging Computational Methods for the Rational Discovery of Allosteric Drugs. *Chemical Reviews*, 116:6370–6390, 2016.

[204] A. C. Walls, Y.J. Park, M. A. Tortorici, A. Wall, A. T. McGuire, and D. Veesler. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell*, 181:281–292, 2020.

[205] A. C. Walls, M. A. Tortorici, and B. Bosch et al. Cryo-electron microscopy structure of a coronavirus spike glycoprotein trimer. *Nature*, 531:114–117, 2016.

[206] Y. Wang, M. Liu, and J. Gao. Enhanced receptor binding of SARS-CoV-2 through networks of hydrogen-bonding and hydrophobic interactions. *Proceedings of the National Academy of Sciences*, 117:13967–13974, 2020.

[207] Y. Watanabe, T.A. Bowden, I.A. Wilson, and M. Crispin. Exploitation of Glycosylation in Enveloped Virus Pathobiology. *Biochimica et Biophysica Acta - General Subjects*, 1863:1480–1497, 2014.

[208] T.R. Weikl and F. Paul. Conformational selection in protein binding and function. *Protein Science*, 23:1508–1518, 2014.

[209] J. Weiser, P.S. Shenkin, and W.C. Still. Approximate Atomic Surfaces from Linear Combinations of Pairwise Overlaps (LCPO). *Journal of Computational Chemistry*, 20:217–230, 1999.

[210] S. R. Weiss and J. L. Leibowitz. Chapter 4 - coronavirus pathogenesis. volume 81 of *Advances in Virus Research*, pages 85 – 164. Academic Press, 2011.

[211] J. M. White, S. E. Delos, M. Brecher, and K. Schornberg. Structures and Mechanisms of Viral Membrane Fusion Proteins. *Critical Review in Biochemistry and Molecular Biology*, 43:189–219, 2008.

[212] P. Workman, A. A. Antolin, and B. Al-Lazikani. Transforming cancer drug discovery with Big Data and AI. *Expert Opinion on Drug Discovery*, 14:1089–1095, 2019.

[213] D. Wrapp, N. Wang, K. S. Corbett, J. A. Goldsmith, C.L. Hsieh, O. Abiona, B. S. Graham, and J. S. McLellan. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science*, 367:1260–1263, 2020.

[214] B. Xia, S. Vajda, and D. Kozakov. Accounting for pairwise distance restraints in FFT-based protein-protein docking. *Bioinformatics*, 32:3342–3344, 2016.

[215] R. Yan, Y. Zhang, Y. Li, L. Xia, Y. Guo, and Q. Zhou. Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. *Science*, 367:1444–1448, 2020.

[216] M. Zacharias. Protein-protein docking with a reduced protein model accounting for side-chain flexibility. *Protein Structure, Function and Bioinformatics*, 12:1271–1282, 2003.

[217] M. Zacharias. Accounting for conformational changes during protein-protein docking. *Current Opinion in Structural Biology*, 20:180–186, 2010.

[218] P. Zavodszky, L.B. Abaturov, and Y.M. Varshavsky. Structure of glyceraldehyde-3-phosphate dehydrogenase and its alteration by coenzyme binding. *Acta Biochemistry and Biophysics of the Academy of Sciences of Hungary*, 1:389–402, 1966.

[219] A. Zen, C. Micheletti, O. Keskin, and R. Nussinov. Comparing interfacial dynamics in protein-protein complexes: an elastic network approach. *BMC Structural Biology*, 10:26, 2010.

[220] P. Zhao, J. L. Praissman, and O.C. Grant et al. Virus-Receptor Interactions of Glycosylated SARS-CoV-2 Spike and Human ACE2 Receptor. *Cell Host and Microbe*, 28:586–601, 2020.

[221] A. Zhavoronkov, Q. Vanhaelen, and T.I. Oprea. Will Artificial Intelligence for Drug Discovery Impact Clinical Pharmacology? *Clinical Pharmacology and Therapeutics*, 107:780–785, 2020.

[222] H. Zhou, Y. Chen, S. Zhang, P. Niu, K. Qin, W. Jia, B. Huang, S. Zhang, J. Lan, L. Zhang, W. Tan, and X. Wang. Structural definition of a neutralization epitope on the N-terminal domain of MERS-CoV spike glycoprotein. *Nature Communications*, 10:828–842, 2019.

[223] J.A. Zorn and J. A. Wells. Turning enzymes ON with small molecules. *Nature Chemical Biology*, 6:179–188, 2010.

[224] M. J. Zvelebil, G. J. Barton, W. R. Taylor, and M. J. E. Sternberg. Prediction of protein secondary structure and active sites using the alignment of homologous sequences. *Journal of Molecular Biology*, 195:957–961, 1987.

[225] R. Zwanzig. Nonlinear generalized langevin equations. *Journal of Statistical Physics*, 9:215–220, 1973.