



UNIVERSITÀ DEGLI STUDI DI PALERMO

Dottorato in Scienze molecolari e Biomolecolari (XXIX ciclo).
Dipartimento STEBICEF (Scienze e Tecnologie Biologiche Chimiche e Farmaceutiche)
Settore scientifico disciplinare CHIM/08 (Chimica farmaceutica)

DEVELOPMENT AND OPTIMISATION OF COMPUTATIONAL TOOLS FOR DRUG DISCOVERY

DOCTOR
UGO PERRICONE

PhD COORDINATOR
CH.MO PROF. PATRIZIA DIANA

SUPERVISOR
CH.MO PROF. ANNA MARIA ALMERICO

SUPERVISOR
CH.MO PROF. GIAMPAOLO BARONE

**Development and optimisation of
computational tools for drug discovery**

To my whole family and especially to Anna for supporting me in every moment of this PhD, without their help it would be more and more difficult

To Prof. Thierry Langer for having accepted me as student in his group, for his precious friendship and for everything he taught me

To Marcus Wieder for his priceless friendship and for being the most important guide during my PhD, with the hope to return him everything he taught me

To my supervisors for supporting me in every moment

To Laura and Marta, two angels passed away too early in this life, I will always carry your memory in my hearth!

INDEX

1 INTRODUCTION	1
2 AIMS OF THE WORK.....	15
3 CHEMOMETRICS AND DRUG DESIGN.....	16
3.1 Conf-VLKA: A structure-based revisitatio	24
3.1.1 Introduction.....	25
3.1.2 Material and methods.....	25
3.1.3 Results and discussion	26
3.1.4 Conclusions.....	37
4. DYNAMIC APPROACH TO VIRTUAL SCREENING	38
4.1 Comparing pharmacophore models derived from crystal structures and from molecular dynamics simulations.....	42
4.1.1 Introduction.....	43
4.1.2 Materials and methods	44
4.1.3 Results and discussion	45
4.1.4 Conclusions.....	53
4.2 Evaluating the stability of pharmacophore features using molecular dynamics simulations	55
4.2.1 Introduction.....	56
4.2.2 Materials and methods	57
4.2.3 Results and discussion	58
4.2.4 Conclusions.....	64
4.3 Pharmacophore models derived from molecular dynamics simulations: A case study ...	65
4.3.1 Introduction.....	66
4.3.2 Materials and methods	67
4.3.3 Results and discussion	68
4.3.4 Conclusions.....	75
4.4 A dynamic – shared Pharmacophore approach to improve early enrichment in virtual screening. A case study on PPAR alpha.....	76
4.4.1 Introduction.....	77
4.4.2 Materials and methods	78
4.4.3 Results and discussion	81
4.4.4 Conclusions.....	91
5 COMPUTATIONAL CHEMISTRY IN POLYPHARMACOLOGY AND DRUG REPURPOSING	93
5.1 The repurposing of old drugs or unsuccessful lead compounds by in silico approaches: new advances and perspectives.	93
5.2 Drugs polypharmacology by in silico methods: new opportunities in drug discovery.....	94
BIBLIOGRAPHY	95

‘If we were to name the most powerful assumption of all, which leads one on and on in an attempt to understand life, it is that all things are made of atoms, and that everything that living things do can be understood in terms of the jiggings and wiggings of atoms.’

Richard Feynman, Nobel Prize in Physics 1965

1 INTRODUCTION

The most important issue in Medicinal Chemistry is without any doubt the drug design part, often referred to as rational drug design or simply rational design. It represents the process of finding new drugs based on the knowledge of a biological target or all the biochemical steps in which the target is involved [1]. Most commonly, The aim of a drug discovery process is to find an organic small molecule responsible for modulating the biochemical patterns of a cell process. The activation or inhibition of a biomolecule function, such as of a protein or of a nucleic acid, results in turn in a therapeutic benefit to the patient. In its basic sense, rational drug discovery involves the design of molecules that, showing a highly complementary chemistry to a specific target, can interact with it, starting a cascade of biochemical responses. In addition to organic small molecules new classes of drugs become everyday increasingly important as, for example, biopharmaceuticals and especially therapeutic antibodies. In order to test and validate these protein-based therapeutics, different techniques for improving the affinity, selectivity, and stability of them have also been developed [2].

In the drug design process, prediction of binding affinity is nowadays the most improved task and, at the same time, the most reliable. However, there are many other properties, such as bioavailability, metabolic half-life, and side effects that must be optimized prior to get a safe and efficacious drug. These pharmacokinetic parameters are yet difficult to predict through rational design techniques. Nevertheless, today, more attention has been focused on selecting candidate molecules presenting physicochemical properties that can lead to fewer complications during development and hence can help in the pathway from lead compound to marketed drug [3]. Furthermore, *in silico* methods, used prior to *in vitro* experiments, have shown a huge benefit in predicting possible ADME (Absorption, Distribution, Metabolism, and Excretion) properties for the potential candidates as well as their toxicological profiles [4]. In contrast to traditional methods of drug discovery based on testing candidate drugs through *in vitro* and *in vivo* assays, and connecting the retrieved effects to treatments, rational drug design is based on an initial hypothesis that a desired effect is due to the modulation of a precise biological target, specifically tuned by a structurally complementary molecule. The first issue of rational drug design is the knowledge of the real

involvement of the target in the studied biochemical disease pathway. This can be sometimes confirmed by the association between target mutations and disease states [5]. The second is the druggability of the chosen target. This relies on the target capability of binding to a small molecule for the modulation of its activity [6]. In a rational drug discovery protocol, the research of small molecules potentially capable to bind to a specific target begins with a screening of libraries containing probable drug candidates. This process can be assessed as “wet screening” or may be done through the computational means searching for drug and lead-likeness of compounds [7]. Several methods are available to estimate drug-likeness such as Lipinski's Rule of Five and a range of scoring methods such as lipophilic efficiency [8].

The optimisation process of a drug design protocol is characterised by a huge number of properties that must be simultaneously tuned. For this reason, it is of common use to adopt some multi-object optimization techniques [9]. Finally, despite all the efforts made in the last years to optimise drug discovery protocols, a successful drug design campaign seems to be mostly reliant on serendipity and bounded rationality [10].

In the last years, the application of computational techniques in drug discovery and development process has gained in popularity, implementation, and appreciation. Different terms have been applied to this area, the most common used are computer-aided drug design (CADD), molecular modelling and *in silico* drug design. The success behind CADD application is due to its capability of increasing the hit rate of novel drug compounds when compared to the classical HTS approach. Compared to the latter, *in silico* methods allow the use of combinatorial chemistry and a much more targeted search, thanks to publicly available databases growth. The main scope of molecular modelling is to explain the molecular basis of therapeutic activity of some molecules and predict possible derivatives that would improve activity [11, 12]. In a drug discovery campaign, computational techniques are usually used for three major purposes:

- (1) Filter large compound libraries into smaller sets of predicted active compounds that can be tested experimentally leveraging chemical and biological information about ligands and/or targets to identify and optimize new drugs;

- (2) Guide the optimization of lead compound, whether to increase its affinity or

optimize drug metabolism and pharmacokinetics properties such as absorption, distribution, metabolism, excretion, and the potential for toxicity (ADMET);

(3) Help in the rational design of novel compounds, either by modifying starting molecules or by tying together fragments into novel chemotypes.

Fast expansion in this area has been made possible thanks to advances in computational software and hardware, and increasing database of publicly available ligand molecules and target protein structures. One of the most important advantages in the use of *in silico* methods is the reduction of chemical space size and, thereby, the possibility to focus on more promising candidates for lead discovery and optimization. The main goal of virtual screening is therefore to eliminate compounds with undesirable properties and enrich the set of molecules with desirable properties. In another words, *in silico* modelling is used to significantly minimize time and resource requirements of chemical synthesis and biological testing. As shown in Fig. 1.1, *in silico* methods become nowadays more and more important as a first step of the entire workflow for the drug discovery process, avoiding possible false positive or false negative results in the search of possible hits to develop. In the last years, in fact, there has been a rapid growth of virtual screening usage, as confirmed by the increase in the number of citations matching keywords “virtual screening”. By using the SCOPUS database [13], it is possible to check that the articles explicitly reporting the keyword “virtual screening” steeply increase about the year 2000 reaching a number of articles 20 times higher in 2015.

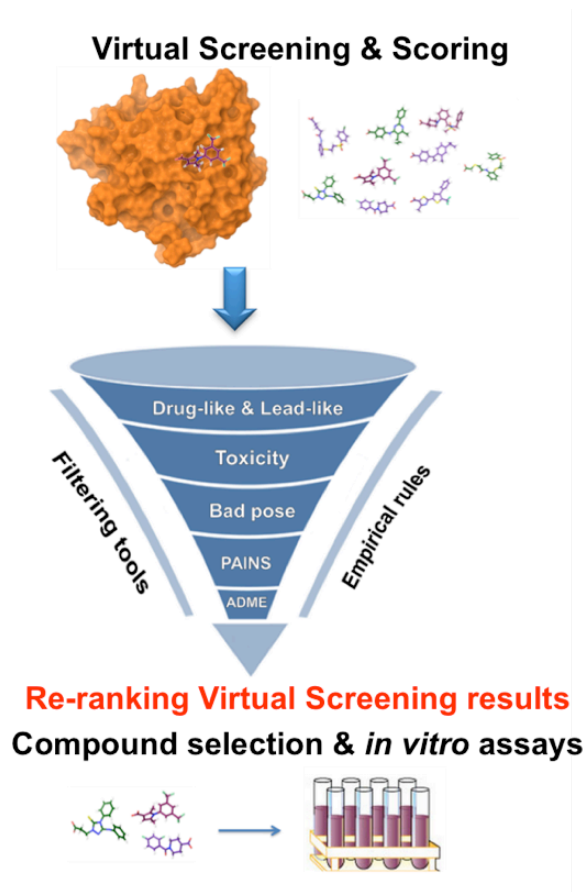


Fig.1.1 *Virtual screening workflow adopted prior to in vitro assays*

D.V. Green of GlaxoSmithKline in a review published in 2003 concluded with: “The future is bright. The future is virtual” [14]. Already in 2003, it was estimated that computer modelling and simulations would account for ~ 10% of pharmaceutical R&D expenditures and that they will have rose to 20% by 2016 [15]. In these days, *PriceWaterhouseCoopers* has published “Pharma2020”, the latest market research about the state of the art and the future of computational chemistry within the pharmaceutical companies [16]. In Fig. 1.2, the comparison between the state of the art and future predictions is reported.

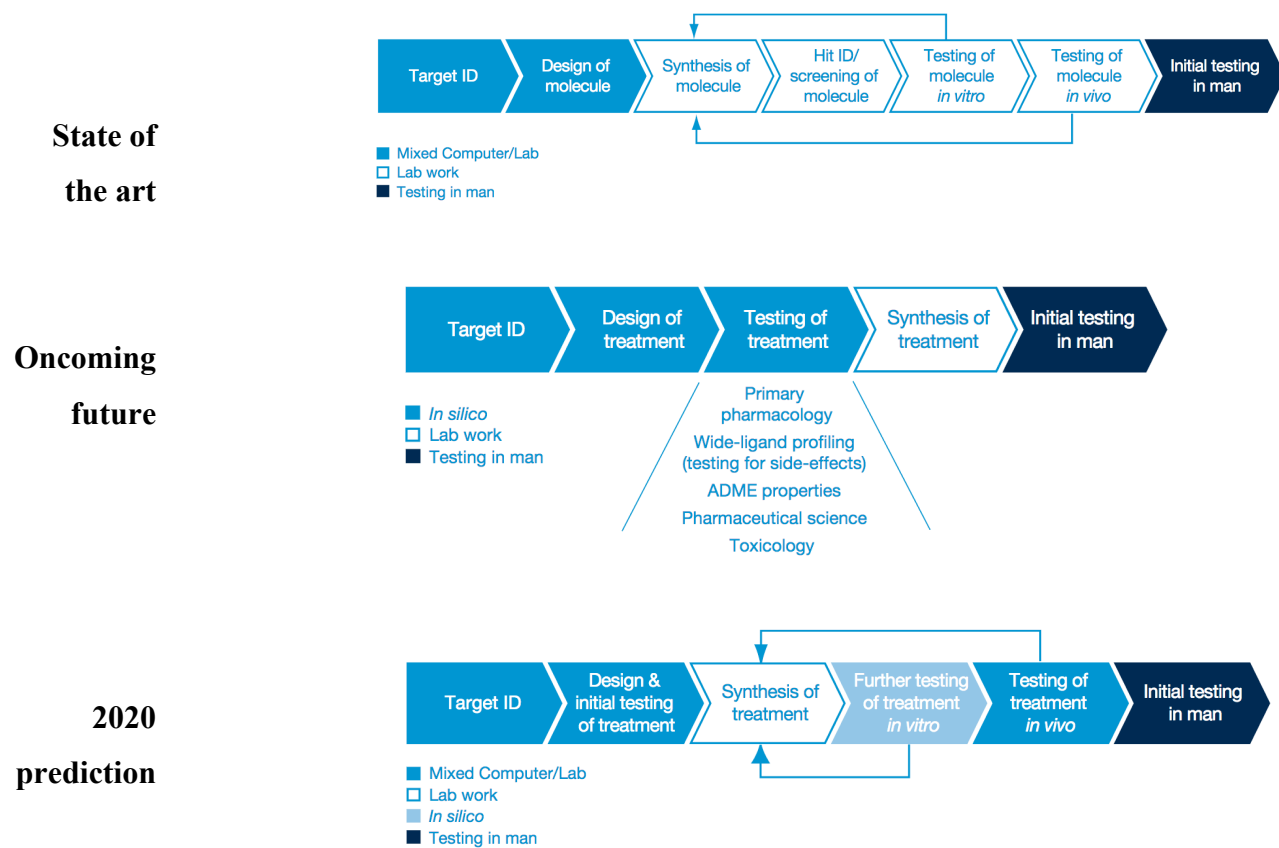


Fig.1.2 comparison between state of the art and future predictions in CADD usage in pharma industries

Referred to CADD there are two major types of drug design. The first is referred to as ligand-based approach [17], and the second, structure-based [18] Fig. 1.3.

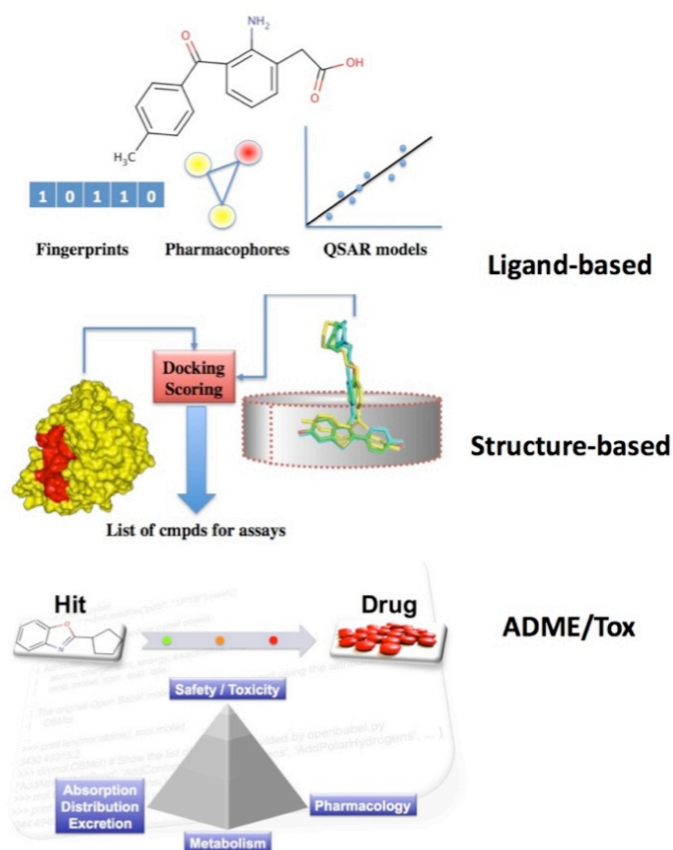


Fig.1.3 *Ligand-based and Structure-based approaches in drug discovery*

Ligand-based drug design is usually adopted when there is no 3D structural knowledge of the target studied. The use of molecules known to be active on the biological target of interest is the starting point used for such an approach. This kind of design strategy is also called indirect drug design because, starting from known active compounds on a specific protein, it tries to find the essential chemical features useful for interacting with that target. Once all the structural information has been collected it is in fact possible to search for chemical similarity between known and new molecules. One of the most applied ligand-based approaches is based on the indirect building of a pseudo receptor derived from a pharmacophore model that defines the minimum necessary structural characteristics a molecule must possess in order to bind to the target. In other words, a model of the biological target binding pocket may be built based on the knowledge of what binds to it, and this model in turn may be used to design new molecular entities that interact with the target [19–

21]. A pharmacophore model can be considered as an abstraction of molecular features necessary for the molecular recognition between a ligand and a biologic macromolecule. The IUPAC defined it as “an ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target and to trigger (or block) its biological response” [22]. The pharmacophore features include hydrophobic centroids, positive or negative ionisable sites, hydrogen bond acceptors or donors and aromatic rings (Fig.1.4). These pharmacophore features may be located on the ligand or may be project points presumed to be located in the receptor [23].

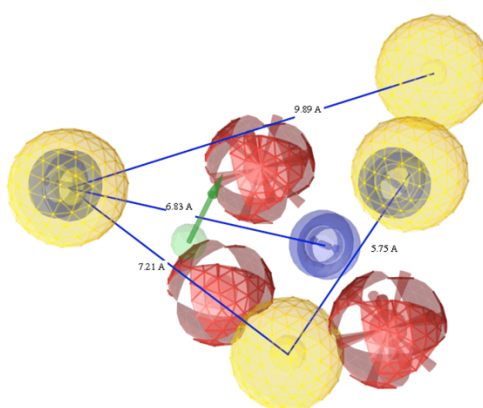


Fig.1.4 *Pharmacophore model generated with LigandScout software. In yellow hydrophobic features are represented, hydrogen-bond acceptors are signed in red and hydrogen-bond donors in green. Blue rings stands for aromatic features*

Another common ligand-based method relies on cheminformatics. In this case, ligand structural information is converted into molecular descriptors, and, through statistical analysis, one can predict possible target for a new molecule. This kind of prediction is based on the structural similarity between the new molecule and a known set of compounds. Such an approach has been developed and applied to the search of new potential drugs [24, 25]. Ligand-based drug design can be also exploited to search for a quantitative structure-activity relationship (QSAR). In this approach, one can determine the statistical correlation between calculated properties of molecules, expressed as molecular descriptors, and their experimental biological activity. Once found the most robust model, the information can be exploited to predict the activity of new analogues [26, 27]. A QSAR model has the form of:

$$\text{Activity} = f(\text{structural properties}) + \text{ERROR} \quad \text{Eq.1}$$

In the last years, a more complete approach has been developed: 3D QSAR. This term refers to the application of force field calculations based on three-dimensional structure of molecules. It exploits the calculation of non-covalent empirical potentials between atom couples, such as the Lennard-Jones potential, rather than using experimental constants to define the interatomic interactions. Some of the parameters analysed are the steric fields (shape of the molecule), the hydrophobic regions (water-soluble surfaces), and the electrostatic fields [28–30].

Structure-based drug design exploits the knowledge of the three dimensional structure of the biological target, obtained through methods such as X-ray crystallography or NMR spectroscopy [31, 32]. The lack of target 3D structure can be overtaken by means of a homology model of the target, using the experimental structures of similar proteins. In this case, the studied protein will be folded according to the amino acid sequence homology with other proteins having known folding structures [33, 34]. In case of low homology levels, it is possible to assess folding prediction through the use of protein threading. In this technique, also known as fold recognition, each amino acid in the target sequence is assigned to a position in a template structure, and an evaluation of how well the target fits the template is done. After the best-fit template is selected, the structural model of the sequence is built [35, 36]. Starting from the knowledge of the biological target structure, candidate drugs can be optimally designed by medicinal chemists, predicting their binding affinity and selectivity. The two main structure based techniques are the 3D pharmacophore modelling [21, 37, 38] and molecular docking [39, 40]. Pharmacophore modelling is more and more preferred to docking for several reasons. First of all, it is more universal. In fact, pharmacophores represent chemical functions, applicable not only to a specific bounded molecule, but also to unknown ones. Secondly, it is very efficient because the computational resources needed for the pharmacophore modelling are really poor. For this reason, it is very suitable for large libraries virtual screening. In the end, it also allows researchers to tune it on the fly adding and removing features or adjusting their tolerance in order to optimise both the sensitivity and selectivity of the screening.

Molecular docking is usually applied to deeply evaluate the interaction between a small molecule and a protein at the atomic level. This helps to study the behaviour of

small molecules in the binding site of target proteins and to deepen biochemical paths. The docking protocol consists of two main parts: firstly, the prediction of the ligand position, conformation and orientation within the binding site (usually referred to as pose) then, on a determined pose, the evaluation of the binding affinity. These two steps rely on what is defined as *searching algorithm* (for the pose research) and *scoring function*, for the binding affinity calculations [39, 41].

Different types of molecular docking have been developed in the last years. The two main approaches exploit ligand flexibility or receptor and ligand flexibility respectively. In the former, ligand conformations may be generated prior to docking or within the receptor binding cavity [42]. To select proper energetically conformations of ligands, knowledge-based [43] or force field-based methods are used [44].

The above mentioned *in silico* approaches used in drug design can be roughly further classified based on the purpose of their application [45]. One of the most used applications of CADD is the virtual screening. It consists in the search of new ligands as potential drugs for a specific target by searching large databases of 3D structures of small molecules that can well fit into the binding pocket of a protein or on a pharmacophore model. A second strategy is the *de novo* design of ligands. In this case, molecules are designed starting from the essential interaction pattern within the binding pocket by assembling molecular fragments that can satisfy those interactions. The strength of such an approach is that molecules created are not present in any database, but are new entities [46]. The last approach consists in the optimisation of already existing molecules to maximise the efficacy or to minimise side effects while maintaining the essential features to interact with the chosen target [47].

In the last years a new way of using CADD has been more and more adopted. It is based on the integration between screening techniques with simulation ones as, for example, Molecular Dynamics (MD).

MD in drug design has demonstrated to give a huge impact in the improvement of drug design strategies. The knowledge of molecular motions can be fundamental for understanding compatibility between two different molecules. Thanks to the modern techniques, the initial idea of a frozen receptor that can accommodate a small molecule without mutating its conformation –also known as “Lock-and key” model [48] - has been largely substituted by a modern idea of dynamic receptor that

undergoes some conformational changes based on the ligand to bind [49, 50]. In Figure 1.5 the general process of molecular dynamics calculations is reported.

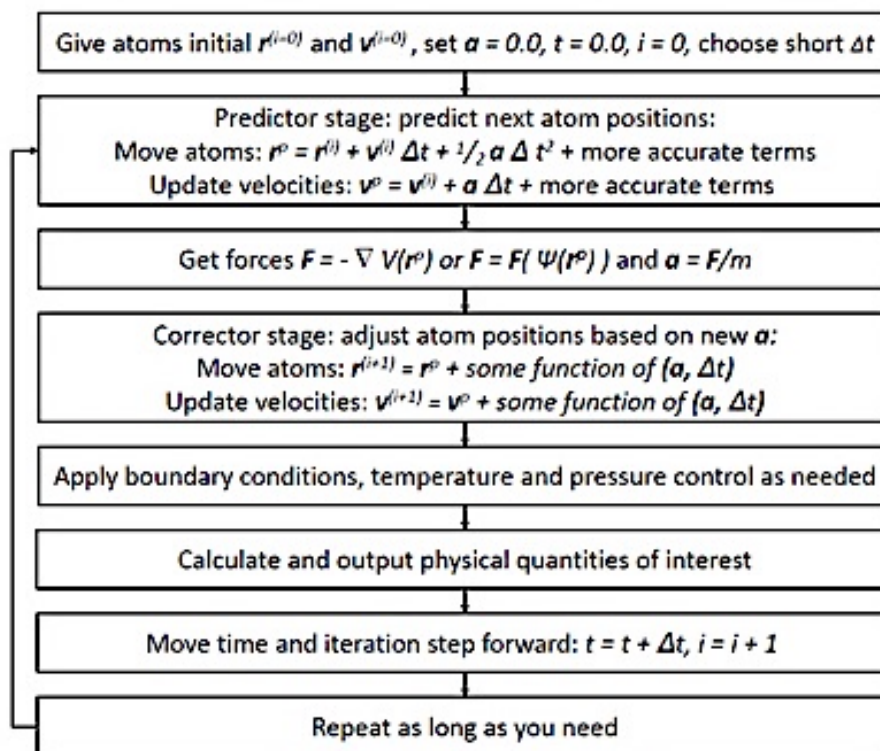


Fig.1.5 Simplified Scheme of molecular dynamics calculations

The first step of MD is the availability of a 3D target structure. This can be obtained throughout X-ray crystallography, Nuclear Magnetic Resonance (NMR), or by homology-modelling. The 3D coordinates of the receptor structure will be used as starting point for the integration of the equation of motion. For this calculations, Energy, expressed as forces between atoms, is calculated exploiting Force Field (FF) parameters according to the formulas reported in Figure 1.6 [51]. The FF contains all the information useful for the calculation of the total energy of the molecules, including bonded and non-bonded terms relatives to atoms within the simulation.

In FF parameters, the bonded part of measurement contains chemical bonds stretching and atomic angles variations modelled as simple virtual springs. Dihedral angles are instead represented by sinusoidal functions that approximate the energy differences between eclipsed and staggered conformations. The non-bonded terms are represented by van der Waals interactions, for the neutral species, Lennard-Jones 6-12 potential, and using Coulomb's law for the charged interactions [52].

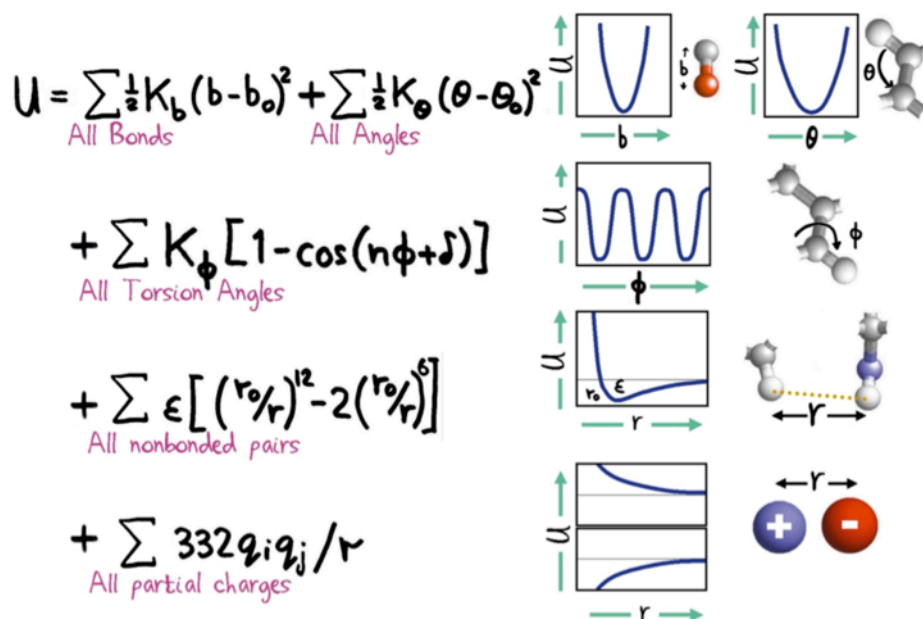


Fig.1.6 Example of empirical Force Field parameters

Even though current force fields present some weaknesses because of several approximations and simplifications, MD simulations play today a very important role in drug discovery because they are the only way to study receptor motions. Just a single protein conformation, for example, tells little about protein dynamics. The static models can be valuable to study the structure of a protein, but drug binding or molecular recognition in general are dynamic processes otherwise not comprehensible if not through the use of MD. The molecular recognition process involves in fact different possible arrangements of both ligand and protein and not their unique and static conformation.

Following the receptor theory, ligands can bind and stabilize only a subset of the different conformations of a receptor and this can cause an induced shift of all the receptor conformations towards the most appropriate to bind the ligands [49]. Moreover, once bound to the protein, the ligand can induce some rearrangements in the binding pocket that are not reproducible in its absence [53]. In the last few years, several approaches have been adopted to simulate the flexibility and dynamicity of the receptors to adopt in virtual screening campaigns. For example, recently, Lexa et al. published a review where it is possible to study all the different approaches adopted in order to take into account protein flexibility for molecular docking [54]. Herein, some of the mentioned methods are presented. One of the most common approaches is the so called “soft docking”. This technique exploits the attenuation of

the Lennard–Jones repulsion term between the receptor and the ligand allowing some minimal backbone movement and side-chain flexibility. Movements are then followed by a rigid-body protein relaxation protocol [55]. In the relaxation methods, the docked complex is taken as a starting point for focusing on protein flexibility by modelling induced-fit effects. The main limitation of this approach is that the dynamic simulation can be only assessed on an all-atom structure: it cannot be performed if the protein is not explicitly represented (e.g. docking grid). Monte Carlo (MC) or MD simulations are actually the most adopted techniques to perform complex relaxation and interactions study. Such a kind of refinement is usually performed after the docking process is finished and the best pose for docking is chosen and it allows other investigations such as solvent effects, examination of the kinetic stability, and prediction of ΔG_{bind} [56, 57]. The last two methods present the limitation that there is not a real view to the conformational modification of the target during the binding process with the ligand. For these reasons other new algorithms have been proposed, for example the induced fit docking method. In the latter, the docking simulation is run considering ligand and protein side chains as flexible to explore new conformational space. The main limitation of such an approach is that its computational requirements are a limitation feature, especially on large-scale virtual screening studies. Furthermore, the only conformational space of the protein is relative to side chains rotamers, it is in fact based on the use of side chain conformation libraries [58, 59]. Most published methods for flexible protein–ligand docking are based on a limited number of receptors and are usually applied to small molecule libraries, that make the evaluation of the methods difficult. The use of a large test set is in fact vital in the performance assessment of a new screening method especially when one wants to measure performance across a range of different targets. Moreover, the use of a dynamical approach to docking is more resource and time-intensive than semi-flexible docking.

The use of multiple receptor conformations for docking however may sometimes decrease the selectivity of the screening process increasing, for example, the false positive rate. The use of multiple conformations may also lead to the creation of a ligand optimal for an average receptor structure that is not experimentally accessible, a so-called ‘paradoxical inhibitor’. To avoid this kind of risks it is possible to take into consideration only receptor conformations that are present in low-energy landscape of the protein. This issue has driven many researchers to focus on the

choice of the optimal method for the selection of the possible receptor conformation to adopt in the screening process.

The dynamic approach has also been adopted in pharmacophore based virtual screening. In these cases, structure-based pharmacophore features are generated starting from protein-ligand complexes taken from molecular dynamics. Recently, a dynamic pharmacophore approach has been proposed by Choudury et al. In their work, some snapshots are extracted from MD and structure-based pharmacophore models are generated within the protein-ligand complexes chosen. The built models are then compared with the docking approach using known active and inactive compounds [60].

Another way to study dynamic pharmacophore, starts from MD to cluster trajectory frames based on the root mean square deviation (RMSD) of the protein-ligand system or the most populated conformations of the receptor [61, 62]. The RMSD for frame x is reported in Equation 2. The procedure is repeated for every frame in the simulation trajectory.

$$RMSD_x = \sqrt{\frac{\sum_{i=1}^N (r'_i(t_x) - r_i(t_{ref}))^2}{N}} \quad \text{Eq.2}$$

*where N refers to the number of atoms in the analysed selection;
 t_{ref} is the reference time, (typically the first frame is adopted as the reference and it refers to time $t=0$);
 r' is the position of the selected atoms in frame x after it has been superimposed on the reference frame, where frame x is recorded at time t_x .*

One of the limitations of these approach is represented by the dismissing of the dynamic information from the MD simulations and the consideration of only some coordinates chosen by the operator. Such a method is in fact strongly biased by the ability of the MD simulation to represent the configurational space and the operator capability to select the most representative frames out of the whole simulation [63, 64]. Moreover, the ligand binding process could be related to a unique receptor conformation, maybe not representative in the dynamic trajectory and this could be missed in the clustering approach. In this case the use of dynamic pharmacophore represent a real thread for the virtual screening campaign [65, 66].

The methods described above, developed for integrating protein flexibility in docking and pharmacophore modelling present several flaws and the output generated could be of ambiguous correctness. Overall, the main, problematic step

always seems to be the correct choice of protein conformation to adopt to generate docking grid or pharmacophore models.

For molecular docking, one should run a number of virtual screening experiments equal to the number of obtained coordinate sets. Unfortunately, the choice of significant structures is no obvious because it is not possible to detect a priori which coordinate set will give good results in virtual screening. From a virtual screening point of view, in fact, every protein conformation that results in a differently ranked molecule list contains potentially important information.

For the pharmacophore approach, the models generated from the MD trajectory (one pharmacophore model for each coordinate set) are equal to the number of screening runs and also in this case every model carrying out new information could be crucial in the realisation of a screening campaign. Comparing the two methods, dynamic structure-based pharmacophore models present less variability compared to the dynamic docking approach based on the coordinate of the amino acid side chains. Pharmacophore feature space is very limited compared to configuration space of the protein side chains coordinates. The geometry tolerance of the pharmacophore features allows to find the same models for slightly different protein configurations.

A possible evident solution for reducing bias in the dynamic approach to virtual screening could be the development of a protocol capable to really explore all the coordinates generated during the MD simulation without having to choose some structures or conformations. Obviously such an approach results to be very time and resources consuming and it is strongly related to the number of atoms to simulate and to the libraries to screen.

2 AIMS OF THE WORK AND OUTLOOKS

The aim of my PhD project was the development, optimisation, and implementation of new *in silico* virtual screening protocols.

Specifically, this thesis manuscript is divided into three main parts, presenting some of the papers published during my doctoral work.

The first one, here named **CHEMOMETRIC PROTOCOLS IN DRUG DISCOVERY**, is about the optimisation and application of an *in house* developed chemometric protocol. This part has been entirely developed at the University of Palermo - STEBICEF Department - under the guide of my supervisors. During the development of this part I have personally worked on the tuning and optimisation of the algorithm and on the docking campaigns to obtain molecule conformations.

The second part, **THE APPLICATION OF MOLECULAR DYNAMICS TO VIRTUAL SCREENING**, presents a new approach to virtual screening, in particular the attention is focused on different approaches to the application of protein flexibility and dynamics to virtual screening.

This part, has been carried out in cooperation with the University of Vienna - Department of Pharmaceutical Chemistry. For these works I have worked in the development of the general workflow, to a lesser extent to the programming (coding) part of the applications used and I mainly focused on the realisation of the screening campaigns and results interpretation.

The third and last part, **COMPUTATIONAL CHEMISTRY IN POLYPHARMACOLOGY AND DRUG REPURPOSING**, concerns the study of the *in silico* methods applied to two main topics of the drug discovery process, such as the drug repurposing and the polypharmacology. In this part I will briefly describe what published in two reviews dealing to the above mentioned topics.

In conclusion during this doctoral project, I have demonstrated how the use of *in silico* tools can be useful in the drug discovery process. The Chemometric protocols developed and optimised represent in fact a helpful strategy to use for target fishing. Whereas, the application of molecular dynamics to virtual screening, especially for pharmacophore modelling, is a new way to deepen crucial features to be adopted in the search of new putative active compounds.

3 CHEMOMETRICS AND DRUG DESIGN

Some of the *in silico* methods such as molecular docking and pharmacophore modelling could be considered as the modern virtual application of the elderly lock-and-key model based on the structural complementarity between a ligand molecule and a receptor [48, 67, 68].

In the recent years, these methods have demonstrated to give an important boost to the pharmaceutical research. On one hand there has been an increase of the computational approaches reliability. On the other hand, however, the putative leads discovered through the computational methods, once synthesized and tested *in vitro* can sometimes disappoint the researchers' expectations. Such a problem causes a waste of a huge amount of time and resources. Moreover, some of the discarded compounds can be instead potentially good candidates to develop. Such a kind of issue is always referred as a false positive and false negative ratio capability of a virtual screening technique. Another interesting aspect is that compounds that sometimes are discarded for a target, can be interesting on others, as suggested in several works [69, 70]. For instance, two main aspects known as "polypharmacology" and "drug repurposing", are known to have shifted researchers' efforts to constantly try to characterize drug-biological target associations [71, 72].

The structural knowledge of targets and ligands has allowed to use chemical and sequence similarities among molecules and receptors to identify putative drugs to be addressed towards different targets in [73, 74]. For this reason, in the first stage of a drug discovery campaign could be useful to test early candidates towards a panel of different biological targets [75]. The possible correlation between ligand and target structures is a well-known issue, but unfortunately today it is still not possible to unambiguously interpret it.

In computational chemistry, the molecular structure can be identified and categorized by molecular descriptors. Molecular descriptors have been successfully adopted by several disciplines, such as chemistry, pharmaceutical sciences, environmental protection policy, and health researches, as well as in quality control. These parameters can be considered as the translation of a chemical property (i.e. chemical structure) into numbers. This kind of conversion, allows treating chemical properties from a mathematical point of view, expanding the exploration panorama that can be applied to molecules. As defined by Roberto Todeschini [76, 77]:

"The molecular descriptor is the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment."

Following this definition, molecular descriptors can be categorized into two main groups: theoretical molecular descriptors, directly connected to the symbolic representation of the molecule, and physico-chemical properties or experimental measurements, such as logP or molar refractivity.

In molecular modelling, theoretical molecular descriptors are usually adopted. This group can be further considered as a collection of smaller groups:

- **0D-descriptors (i.e. constitutional descriptors, count descriptors);**
- **1D-descriptors (i.e. list of structural fragments, fingerprints);**
- **2D-descriptors (i.e. graph invariants);**
- **3D-descriptors (such as, for example, 3D-MoRSE descriptors, WHIM descriptors, GETAWAY descriptors, quantum-chemical descriptors, size, steric, surface and volume descriptors);**
- **4D-descriptors (such as those derived from GRID or CoMFA methods, Volsurf).**

The above classification is taken from the book "The handbook of molecular descriptors" by Roberto Todeschini [77].

The use of a single molecular descriptor is not enough to predict a biological target for a molecule. However, the use of a carefully selected set of molecular descriptors can be a very powerful translator that can reveal important information about necessary structural features of a molecule to interact with a specific receptor. For example, topological descriptors based on a multiple bioactive reference structures have been employed in similarity-based virtual screening, showing to be potentially more effective than fingerprints, scaffold-hopping or ligand topological pharmacophores [78, 79].

Recently, the use of molecular similarity approach, has been more and more adopted for the discovery of some potential lead compounds [80]. Nonetheless, it is important to point out that a chemical similarity between two molecules, expressed as similar molecular descriptors' chemical space, is not always synonym of same biological activity [81, 82].

In the last years, the research group I worked with for my PhD has developed an approach based on the use of the molecular descriptors as the mean through which building biological lock models for different targets in order to identify new putative drug molecules. This indirect approach starts with the calculation of molecular descriptors for known inhibitors of a selected target. Based on the molecular descriptors values, the idea is to build a target profile, here called lock model, which is created on the structural features of its specific binders. The research of new candidates is based on the possibility of finding new molecules responding to the structural requisites for the target profile previously created [24].

All the chemical structures have been collected from the BindingDB, a Public database including chemical structures classified by biological activity [83].

The first key step of this *in house* method, called Virtual Lock-and-Key Approach (VLKA), was the random choice of 47 biological targets, from now indicated as T_n presenting known inhibitors with measured biological activity available in the BindingDB.

Starting from these structures, known inhibitors were chosen from BindingDB and CODESSA PRO software [84] was used in order to calculate a set of molecular descriptors. This software is able to calculate about 1000 molecular descriptors, from 0D to 3D. As mentioned before, the aim of the protocol is to build a lock model for each biological target (T_n) starting from a target profile traced by molecular descriptors value of its known inhibitors. In order to choose a compound selection for the lock model constitution, a biological data cut-off was adopted (K_i , IC_{50} , EC_{50}) [24]. For the creation of the protocol, 173 molecular descriptors were chosen in order to have not blanks for all the selected compounds constituting the lock set. For the calculation of 3D-molecular descriptors, global minimum conformations from *in vacuo* minimisation were selected. Mean (m) and standard deviation (s) of the molecular descriptors values ($X_{i,j}$) for each biological target (T_n) were calculated (Fig. 3.1A). The hypothesis behind the protocol is that the value of each molecular descriptor of a suitable inhibitor should be close to the same molecular descriptor mean (m) calculated for all the inhibitors of the same biological target. Starting from that, every molecular descriptor value [$X_{i,j}(T_n)$] of the compounds, included in the Lock set, was converted into a numerical coefficient in relation the closeness to m (Fig. 3.1B), as reported in Eq. (3.1):

$$\begin{aligned}
& \text{if } X_{i,j}(T_n) > \mu \pm \sigma, \rightarrow \alpha = 0; \\
& \text{if } (\mu - \frac{1}{2}\sigma) < X_{i,j}(T_n) < (\mu + \frac{1}{2}\sigma), \rightarrow \alpha \\
& \quad = 1; \\
& \text{if } -\sigma < X_{i,j}(T_n) < -\frac{1}{2}\sigma, \rightarrow \alpha = 0.5; \\
& \text{if } +\frac{1}{2}\sigma < X_{i,j}(T_n) < +\sigma, \rightarrow \alpha = 0.5.
\end{aligned}
\tag{eq.3.1}$$

where: X represents the molecular descriptor value;
i is related to the structure;
j is related to the molecular descriptor;
Tn represents the biological target.

Basically, each biological target needs specific chemical-physical properties to be activated, so, it is wise to assume that some molecular descriptors could express better than the others the key structural requirements for the specific biological target. Starting from this consideration, the molecular descriptors values were weighed for each Tn on the basis of the α coefficients determined for the lock set, by considering the sum of the α value for each descriptor (Dj) for all compounds, belonging to the specific biological target (Fig. 3.1C). The following step was to normalize these values by defining the ωD_j coefficients (Fig. 3.1D) as reported in Eq. (3.2).

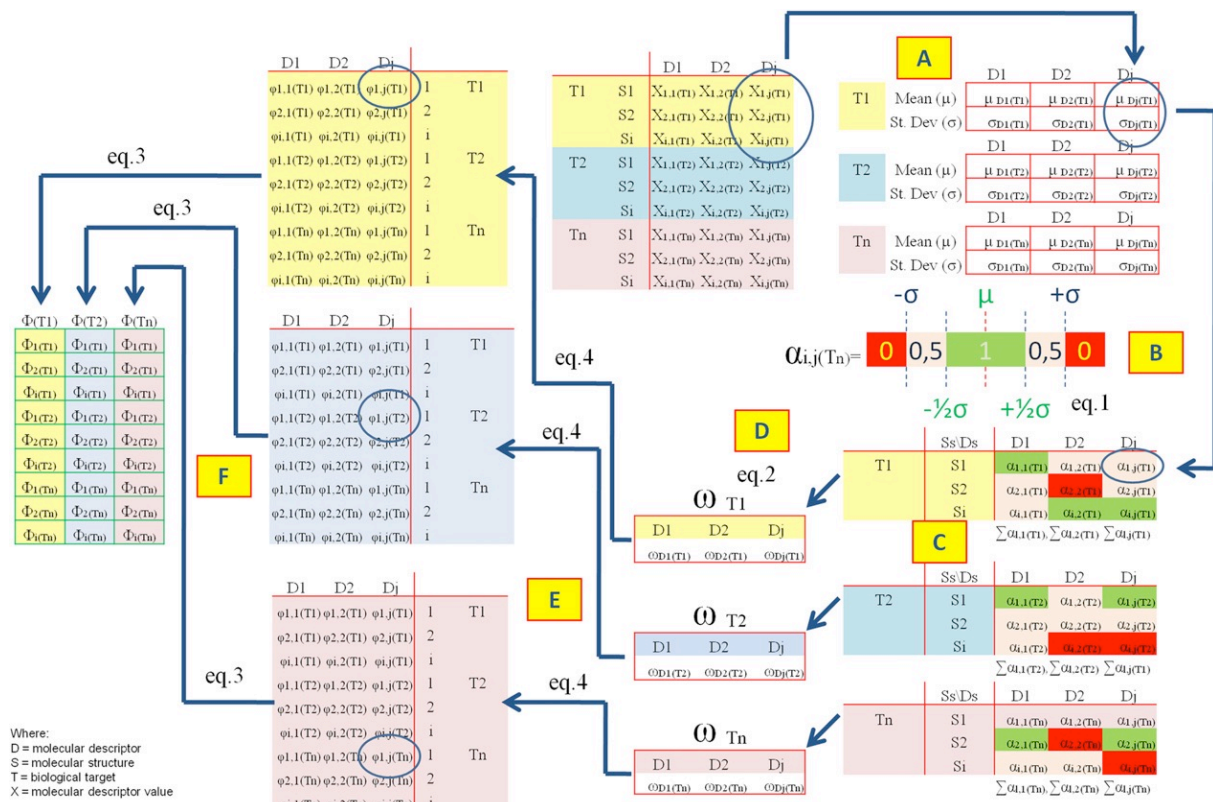


Fig. 3.1 Virtual lock-and-key approach flow chart. A: Calculation of Mean (m) and standard deviation (s) of the molecular descriptors values ($X_{i,j}$) for each biological target (T_n); B: Conversion of each molecular descriptor value [$X_{i,j}(T_n)$] in a coefficient; C: Molecular descriptors weighing by a coefficient for each biological target (T_n); D: Normalization step by defining the ω_{Dj} coefficients; E: Partial scores 4 calculation; F: Total score V calculation

$$\omega_{Dj} = \frac{\sum_{l=1}^i \alpha_{l,j}(T_n)}{\max[\sum_{l=1}^i \alpha_{l,j}(T_n)]} \quad (\text{eq. 3.2})$$

where: i, j , and T_n are defined in Eq. (3.1);

$\max[\sum_{l=1}^i \alpha_{l,j}(T_n)]$ represents the higher α sum of all molecular descriptors belonging to specific biological target.

The $\alpha_{i,j}(T_n)$ and ω_{Dj} coefficients were used to calculate the affinity of all the 7352 compounds under investigation for each biological target. Thus according to Eq. (3.4) the partial score ϕ was calculated, and the total score Φ was defined as sum of ϕ Eq. (3.3) (Fig. 3.1 E-F).

$$\Phi_{i(Tn)} = \sum_{j=1}^{173} \phi_{i,j}(Tn) \quad (\text{eq. 3.3})$$

$$\phi_{i,j} = \alpha_{i,j}(Tn) \omega_{Dj} \quad (\text{eq. 3.4})$$

where: $\phi_{i,j}$ represents the partial score;
 Φ represents the total score;
 i, j , and T_n are defined in Eq. (3.1)

All the calculated scores, for all the structures for each biological target were converted in rankings.

At the end, the Φ scores rank all the 7352 database compounds with respect to the 47 biological target. Inhibitors related to each biological target should occupy the higher rankings. To verify this hypothesis the enrichment score (E%), considered as the percentage of correct classification, was calculated according to eq. (3.5):

$$E\% = \left(\frac{\sum W - \sum P}{\sum W - \sum B} \right) * 100 \quad (\text{eq. 3.5})$$

where: W represents hypothetical lowest rankings;
 B represents hypothetical highest rankings;
 P represents the obtained rankings.

Two different E% scores: E%1 related to the “lock set”, and E%2 for the “total set” of a biological target were calculated. The E%1 reached an average value of 80.4% and for many targets values up to 95%, and the E%2 reached an average value of 79.0% (Fig. 3.2).

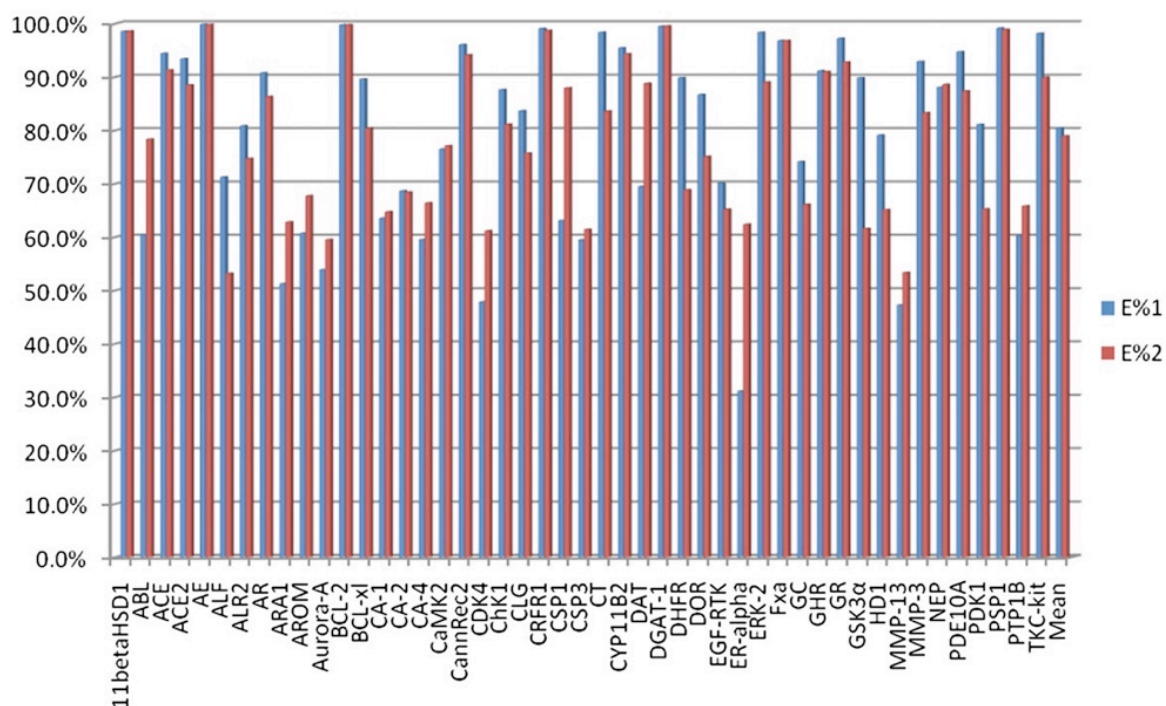


Fig.3.2 E% for the lock and total set

As previously mentioned, the core of VLKA protocol consists in setting-up a “lock model” for a biological target, starting from the respectively known inhibitors. In this scenario, molecular descriptors could be considered as pins of a lock (receptor binding pocket) to be released by a key (molecule) (Fig. 3.3a). Considering this assumption, a new molecule could be considered an inhibitor of a biological target if the values of its molecular descriptors fall in the calculated range values for the set of known inhibitors for the same target.

Briefly, for each structure, the range of molecular descriptors constituting a “lock pin” were defined considering the mean value of them (D mean) and the standard deviation (s) as tolerance (Fig. 3.3b). When the molecular descriptors values of a molecule fall into these defined ranges the lock can be released and the structure can be considered as a potential inhibitor (Fig. 3.3c).

To be released, a real lock needs that all pins must fit the lock structure whereas, in this protocol, the higher is the number of fit pins, the higher will be the affinity to the considered biological target.

In the VLKA, the biological target lock pins are represented by a sequence of 173 molecular descriptors.

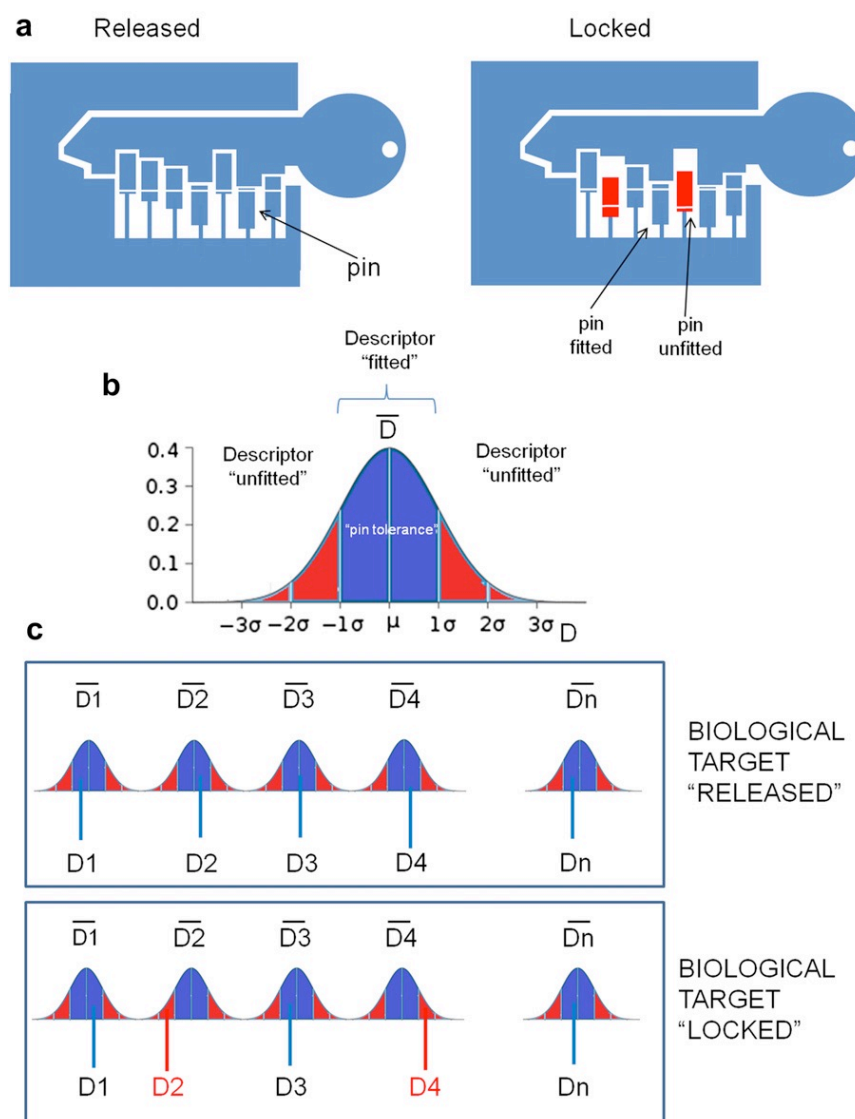


Fig. 3.3 From the lock to biological target. *a)* How a real lock works; *b)* The range of “lock pin” molecular descriptors values (mean μ and σ) can be considered the “pin tolerance”; *c)* When all the molecular descriptors values fall into the “pin tolerances” the biological target “releases”.

The affinity score of a molecule against a specific target is then evaluated as the number of the molecular descriptors “fitted” (Fig. 3.3).

Moreover, as mentioned before, not all the molecular descriptors have the same weight in the lock constitution: some of them are really representative for the lock while some of them can be omitted. So it was necessary to prioritize some descriptors among others. Using this approach, it was possible to rank molecules of the training set based on their affinity against the protein set. What was expected from these assumptions was that inhibitors of a specific biological target should be retrieved in the higher ranking positions for that target.

3.1 Conf-VLKA: A structure-based revisit of the Virtual Lock-and-Key Approach

3.1.1 Introduction

Starting from the *in house* application VLKA, in the attempt to deepen the ligand conformation influence on the protocol, we decided to test the same previous algorithm of scoring and ranking starting from the docked conformation of ligands. Docking calculation was used for two different purposes: to retrieve docking scores, first (in order to test the algorithm for target assignation and possible off target application), then to provide the docked pose of ligands into the relative targets to be used in the VLKA method. The docked ligand poses were in fact used to re-calculate the 142 3D-descriptors (Conf-VLKA), out of the total 173 descriptors originally used in the VLKA. In fact, the remaining 31 descriptors out of the initial 173 set were 1D and 2D, and they did not need to be re-calculated because not influenced by ligands conformation. The original VLKA results, based on molecular descriptors obtained with *in vacuo* optimized structures [24], were then compared to the new approach in the attempt to evaluate the likely influence of 3D ligand conformation on the protocol prediction capability. The comparison between the two methods was also made, by analysing docking results in scoring and ranking molecules, for the different targets [85].

3.1.2 Material and Methods

Target choice

Being the most important issue of the approach the comparison of the new protocol with the previous one, we decided to maintain the same targets and ligands of the original method [24].

VLKA algorithm: scoring and ranking

For the algorithm details please refer to the previous paragraph **VIRTUAL LOCK AND KEY APPLICATION** [24].

Ligand Structure similarity evaluation

In order to check the structural diversity of ligands for each target set, preventing the enrichment of redundant molecular analogues, we set up a topological evaluation of the whole database. For each target, ligand structures were submitted to calculation of radial fingerprint [86], molprint2D fingerprint [87] and MACCS keys [88] and then analysed in terms of Tanimoto distance [89] using similarity matrix on

CANVAS [90, 91]. The Tanimoto similarity cut-off value usually chosen as index of similarity is above 0.75 [92].

3D biological structures selection and optimization

To carry out this comparative approach, the 3D structures of the biological targets included in the VLKA have been downloaded from the RCSB Protein Databank (PDB) [93], complexed with co-crystallized ligands. The selected structures were submitted to the optimization and refinement process using Protein Preparation Wizard utility of Maestro Schrödinger suite. During this process bond orders were assigned, the missing hydrogens were added, the disulfide bonds were eventually assigned, the water molecules were deleted, the protonation of aminoacids were determined. At the end, the hydrogen bonds of the proteins were optimized, and restrained minimization was carried out on heavy atoms converging to RMSD equal to 0.30 Å, and on the hydrogen atoms.

Docking and descriptors calculation

The ligands co-crystallized within PDB structures were extracted and docked using Glide XP high performance docking procedure [94–96], as a test for pose prediction quality of the searching docking algorithm. The 7352 compounds of the VLKA were submitted to the docking and scoring procedure versus the own target, and then versus the entire biological targets dataset. The best ligand pose for each compound was selected according to the XP Glide Score. Once docked, 3D molecular descriptors for the best pose structures were re-calculated as in the original work [24].

3.1.3 Results and Discussion

The aim of this work was to explore the VLKA protocol capability using docked conformation of ligands. The original approach was based on molecular features of known inhibitors expressed as 1D, 2D, and 3D descriptors calculated on *in vacuo* conformation of molecules. The new method was based on the 3D descriptors calculation on the best docked conformations of each compound. This last approach (Conf-VLKA), in our opinion, could give a new interesting point of view due to the fact the original descriptor matrix consisted of only 31 1D/2D descriptors over 173, the total descriptors used [24]. So it is plausible to observe a variation in most of the values due to the change of 3D conformation of molecules. Consequently, the

“locks” and the pin tolerance could result different from the original VLKA. To set up the study, biological targets were taken from RCSB Protein Data Bank. 43 out of 47 biological targets were retrieved into the PDB because of a lack for some 3D structures such as CAMK2 (Calmoduline Kinase 2), CB2 (Cannabinoid Receptor 2), Ghrelin Receptor (GHSR), and Diacylglycerol acyltransferase (DGAT-1). Even though it is common practice to re-build protein structures by means of homology modelling, when these ones are not available in databases, this procedure, starting from the primary sequence, allows obtaining calculated structures and hence less reliable structures respect to experimental ones. So, finally we decided to discard targets for which 3D-structures were unavailable. For many targets, multiple structures were retrieved, and for some targets (CA-4, CDK4, CT), no bound ligand was available. All the 3D biological structures, taken into account, are reported in Table 3.1.

Table 3.1. Target PDB ID and relative crystalised ligands codes

Target	PDB ID	Crystalised Ligand
11-beta-Hydroxysteroid Dehydrogenase (11-betaHSD1)	3EY4	352
ABL Kinase (ABL)	2HYY	IMATINIB STI571
	2QOH	PPY-A
	3CS9	NILOTINIB
	3OXZ	OLI
	3QRK	9DP
	3BKK	KAF
Angiotensin Converting Enzyme (ACE)	1R4L	XX5
Angiotensin Converting Enzyme 2 (ACE2)	1PWP	NSC
Anthrax lethal factor (ALF)	1YQY	915
	1ZXV	MFM
	3EML	ZMA
Adenosine receptor A1 (ARA1)	1EF3	FID
Aldose Reductase (ALR2)	3H4G	FID
	3ZQT	30Z
Androgen Receptor (AR)	3EQM	ASD
Aromatase (AROM)	2X6D	X6D
Aurora Kinase A (Aurora-A)	2XNE	ASH
	3K5U	PFQ
	3M11	AKI
	3P9J	P9J
	3R21	D36
	1YSW	43B
	2YXJ	NRC
BCL-2 (BCL-2)	1BZM	MZM
BCL-xl (BCL-xl)	3K34	SUA
Carbonic Anhydrase 1 (CA-1)	3M04	BE9
Carbonic Anhydrase 2 (CA-2)	3NJ9	TE2
	1ZNC	—
	2W96	—
Carbonic Anhydrase 4 (CA-4)	3NLB	5BE
Cyclin-Dependent kinase (CDK4)	3PA3	C70
Checkpoint kinase (ChK1)	2Y6I	IPI
Collagenase (CLG)	3EHT	MAL
Corticotropin Releasing Factor Rec 1 (CRFR1)	1RWK	158
Caspase-1 (CSP1)	2XYG	XVE
	3KJF	B92
	3PDO	—

Table 3.1. cont.

Target	PDB ID	Crystallised Ligand
Chymotrypsin (CT)	2P8O	BVA
Dopamine Transporter (DAT)	3PBL	ETQ
Dihydrofolate Reductase (DHFR)	3NTZ	3TZ
	3OAF	OAG
Estrogen Receptor (ER-alpha)	1PCG	EST
EGF-R Tyrosin Kinase (EGF-R TK)	2RGP	HYZ
	3POZ	03P
ERK-2 Kinase (ERK-2)	3QYW	6PB
Factor Xa (Fxa)	2WYG	461
	2XBV	XBV
	3KQB	M35
Glutaminy Cyclase (GC)	3SI2	PBD
Glucocorticoid Receptor (GR)	1NHZ	486
Glycogen Synthase Kinase 3 α (GSK3 α)	2OW3	BIM
	3GB2	G3B
Histone Deacetylase 1 (HD1)	2VCG	S17
Matrix Metallo Proteinase 13 (MMP-13)	2YIG	5EL
	3KEJ	3EJ
	3LJZ	LA3
Matrix Metallo Proteinase 3 (MMP-3)	3OHO	Z79
Neutrophil Endo Peptidase (NEP)	2YB9	HA0
Phospho Di Esterase Type 10A (PDE10A)	2Y0J	AXC
PDK1 Kinase (PDK1)	3RCJ	3RC
	3NUN	JMZ
Plasmeprin 1 (PSP1)	3QS1	6
Protein Tyrosin Phosphatase (PTP1B)	2OZ5	7XY
Tyrosin Kinase C-kit (TKC-kit)	1T46	STI

According to the Glide docking procedure, target grids were calculated on the 3D coordinates of the crystallised ligand within the PDB crystal. For those targets not bearing any ligand, we decided to exploit the PDBsum database information to calculate target grids [36], on the residues identified to be part of the binding pocket. Cognate docking was applied to the PDB dataset to test the docking searching algorithm capability. The root mean square deviation (RMSD) between the redocked pose of the ligand and the original co-crystallized one was the accuracy parameter chosen. Generally, the lowest is the RMSD the most accurate is the docking algorithm, this also allowing us to choose the more suitable target structure (Table 3.2).

For systems presenting more than one PDB available, the one presenting the lowest RMSD value was chosen. The cut-off value for choosing a system was set as $< 2.0 \text{ \AA}$. For a few systems we had to choose the lowest value that was anyway quite higher than 2.0 \AA (Table 3.3).

Table 3.2. RMSD values for targets with multiple 3D structures

Target	PDB ID	Bound Ligand	RMSD (Å)
ABL Kinase (ABL)	2HYY	IMATINIB STI571	1.53
	2QOH	PPY-A	5.6
	3CS9	NILOTINIB	2.08
	3OXZ	OLI	1.54
	3QRK	9DP	7.4
Anthrax lethal factor (ALF)	1PWP	NSC	1.7
	1YQY	915	2.3
	1ZXV	MFM	2.8
Aldose Reductase (ALR2)	1EF3	FID	2.9
	3H4G	FID	3.1
Aurora Kinase A (Aurora-A)	2X6D	X6D	3.12
	2XNE	ASH	3.8
	3K5U	PFQ	1.63
	3M11	AKI	2.3
	3P9J	P9J	1.7
Carbonic Anhydrase 2 (CA-2)	3R21	D36	3.8
	3K34	SUA	3.6
	3M04	BE9	1.7
	3NJ9	TE2	4.5
Checkpoint kinase (ChK1)	3NLB	5BE	5.4
	3PA3	C70	1.3
Caspase-3 (CSP3)	2XYG	XVE	3.2
	3KJF	B92	2.19
	3PDO	—	4.92
Dihydrofolate Reductase (DHFR)	3NTZ	3TZ	1.7
	3OAF	OAG	2.09
EGF-R Tyrosin Kinase (EGF-R TK)	2RGP	HYZ	1.9
	3POZ	03P	3.7
Factor Xa (Fxa)	2WYG	461	3
	2XBV	XBV	2.4
	3KQB	M35	2.2
Glycogen Synthase Kinase 3 α (GSK3 α)	2OW3	BIM	3.2
	3GB2	G3B	1.09
Matrix Metallo Proteinase 13 (MMP-13)	2YIG	5EL	6.7
	3KEJ	3EJ	2.6
	3LJZ	LA3	3.5
PDK1 Kinase (PDK1)	3RCJ	3RC	2.9
	3NUN	JMZ	0.6

Table 3.3. 3D PDB structures selected

Target	PDB ID	Target	PDB ID
11-betaHSD1	3EY4	CSP3	3KJF
ABL	2HYY	CT	2P8O
ACE	3BKK	DAT	3PBL
ACE2	1R4L	DHFR	3NTZ
ALF	1PWP	EGF-R TK	2RGP
ALR2	1EF3	ER-alpha	1PCG
AR	3ZQT	ERK-2	3QYW
ARA1	3EML	Fxa	3KQB
AROM	3EQM	GC	3SI2
Aurora-A	3K5U	GR	1NHZ
BCL-2	1YSW	GSK3 α	3GB2
BCL-xl	2YXJ	HD1	2VCG
CA-1	1BZM	MMP-13	3KEJ
CA-2	3M04	MMP-3	3OHO
CA-4	1ZNC	NEP	2YB9
CDK4	2W96	PDE10A	2Y0J
ChK1	3PA3	PDK1	3NUN
CLG	2Y6I	PSP1	3QS1
CRFR1	3EHT	PTP1B	2OZ5
CSP1	1RWK	TKC-kit	1T46

The next step was the application of the docking on the 7352 VLKA compounds. For each target, docking calculations were performed on specific target compounds (lock set and total set), and on the rest of the entire VLKA dataset. For three targets (Asparaginyl Endopeptidase, AE; Aldosterone Synthase, CYP11B2, Delta Opioid Receptor, DOR), no docked pose was generated for the majority of the compounds, so we decided to exclude them from the analysis because not significant. The best pose for each molecule was chosen according the Glide Score and visual inspection in order to avoid atomic clashes. The docked conformation of molecules was then

submitted to the VLKA algorithm. As previously explained, in the VLKA, the structural affinity of a compound towards a specific target is expressed by a score (Φ), calculated on the weighted average values of molecular descriptors. Based on this parameter each compound is ranked versus each receptor creating the E% score for every target analysed; the highest is the score, the highest is the probability that the ligand is correctly assigned to its target. At this step of our study, we replaced the Φ scores with the docking scores, and recalculated the E% scores. The application of docking protocol gave us results for 40 out of 47 targets included in the original VLKA. The simple use of docking algorithm for target assignment of molecules pointed out that E%, both in case of lock set (E%1) and test set (E%2), are lower than the E% scores of the original VLKA, with a mean value of 60.0%. Just for few targets, the E% scores exceed the 80.0% (Figure 3.4). These results reflect that the use of docking scores did not revealed suitable for this kind of approach, maybe because the docking score itself does not take into account the structural features of compounds for target assignation of molecule, but simply evaluate the energetic profile of the ligand-protein interaction. One of the reasons why this approach gave to us a lower capability compared to the original one could be due to the docking scoring function. In fact, it does not work the same on all the targets. For this reason we wanted to use docking only to consider the molecules poses to recalculate 3D descriptors instead of using docking score.

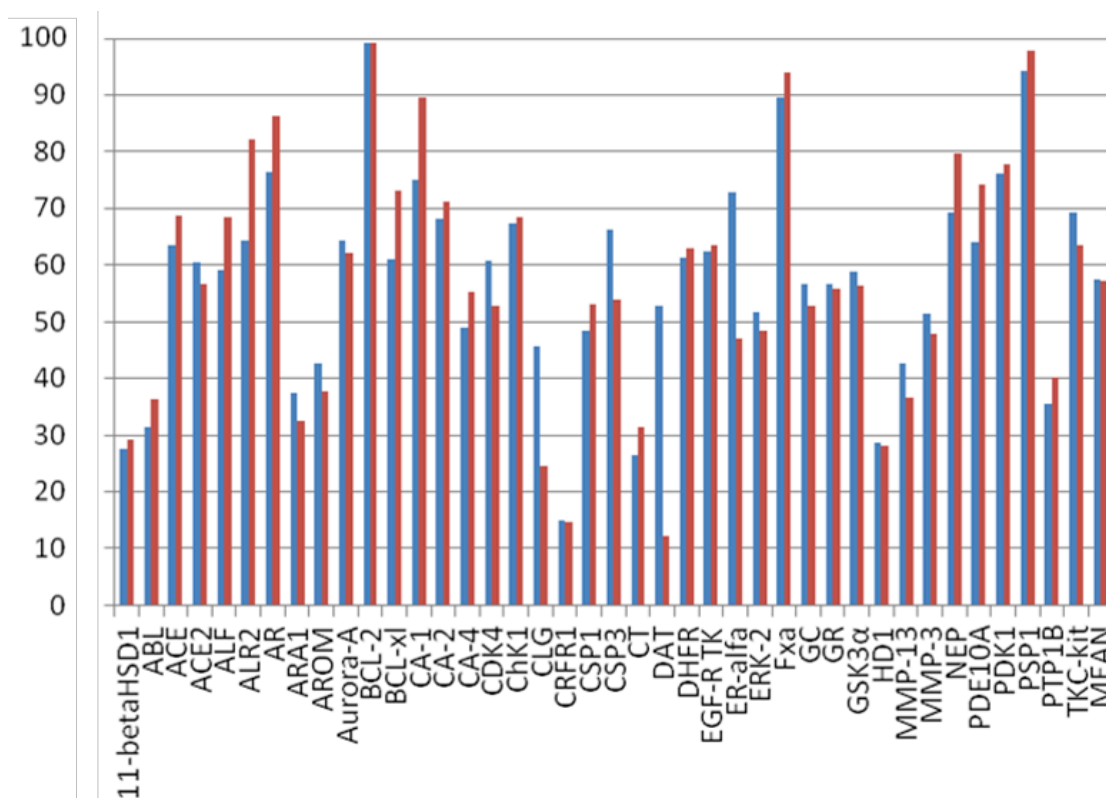


Figure 3.4. *E%1 (blue) and E%2 (red) related to docking scores*

New 3D descriptors values (calculated on the docked conformation of ligands) were inserted into the matrix of the 7352 compounds and the latter was submitted to VLKA algorithm (Conf-VLKA). As last step, the E% scores were calculated again on the new scoring and ranking results. The average E%1 related to the lock set showed a value of 86%, quite greater than the average E%2 related to test set of inhibitors (79.1%). In some cases (11-betaHSD1, BCL2, BCL-x1, CRFR1) E%1 and E%2 showed a more evident rise, while other targets such as ALF, GC, MMP-13, PDK-1, E%2 resulted in a moderate predictive ability (>60%).

In Figure 3.5 we report the E% for the new approach.

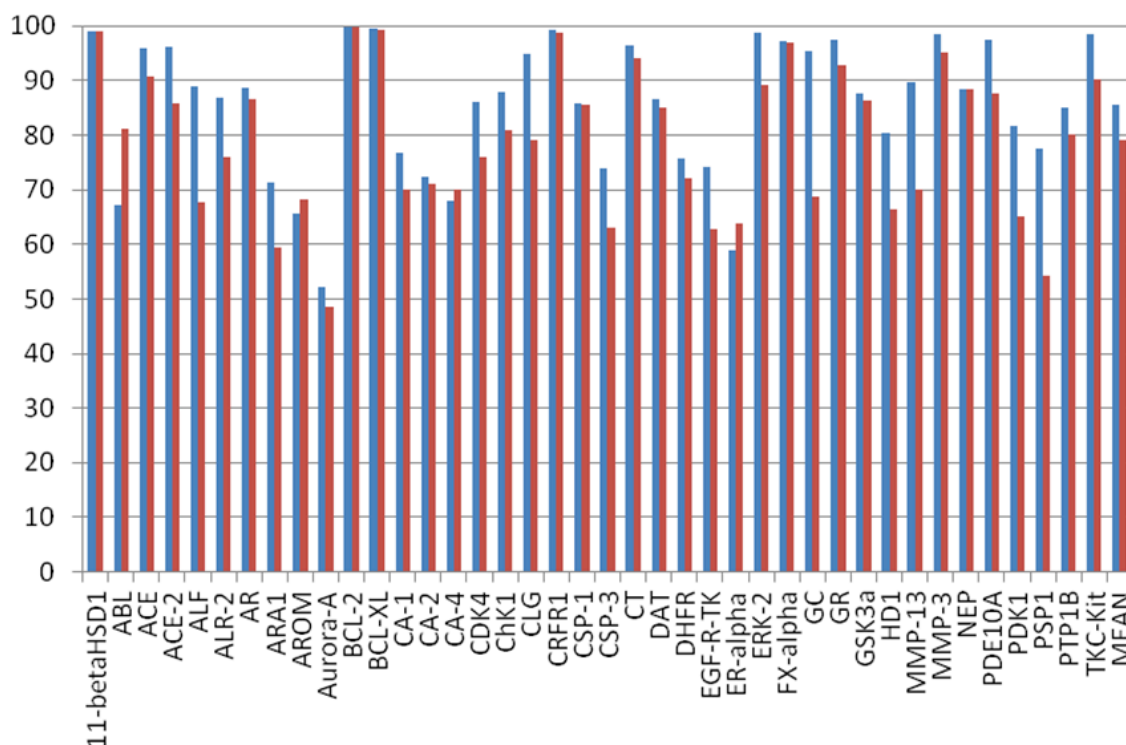


Fig. 3.5 E1% (blue) and E2% (red) related to 3D descriptors calculated from docking poses

In order to avoid analogue redundancies in the ligands set used in this protocol, we wanted to assess structural similarity evaluation of compounds. For each target, ligand structures were submitted to fingerprint calculations as described in the methods section. For the ligand sets analysed, no values higher than 0.75 between the structures belonging to the same target set.

The aim of this paper is to compare three different approaches: the original VLKA, where molecular descriptors are calculated on *in vacuo* optimized structures, with two more approaches, one based on docking scores and the other exploiting docked conformation of ligands for 3D molecular descriptors calculation.

In the original VLKA approach, the average E%1 achieved the 80.4%, and the E%2 hit the 78.9%. In some cases (11betaHSD1, AE, BCL-2, CRFR1, DGAT-1, PSP1) E%2 yielded a high level of predictive capability (98%). For other biological targets (ALF, BCL-xl, CT, DHFR, DOR, GC, GSK3 α , PDK-1) E%2 showed lower values, but despite this, E% values confirmed a quite good predictive capability (>60.0). Only for ALF, this value dropped to 53.1%. In the cases of ABL, ARA1, AROM, AURORA-A, CA-4, CDK4, CSP1, DAT, ER-alpha, the obtained data resulted interesting because E%2 is higher than E%1. In the second approach, the one based

on docking score, the average E%1 and E%2 values were lower than the first above mentioned approach, both near to the 60%. Just few target showed prediction capability >80% (BCL2, CA-1, and PSP1).

In the last approach, the conf-VLKA, the average E%1 was 86%, and for many targets it rose up at 98%. The average E%2 was 79.1%, just greater than E%2 in the original approach. In conclusion, we found that the use of the simple docking score for target fishing is not always reliable, maybe because of a caveat of docking scores which, is known, are not fully related to the protein-ligand binding energy. Docking is much more interesting when used to explore ligand conformations inside the binding pocket. In fact, in the last approach, the use of docked ligand conformations to recalculate the 3D descriptors and the locks, slightly enhanced the E%1 and E%2 compared to the original approach ($\Delta E\%1=+6\%$, $\Delta E\%2=+0.2\%$). Even though the average accuracy of the prediction is similar to the previous one, the most interesting data is that for certain targets there was a rise of the E%. For BCL-xl target an increase of 11% for the E%1 and a 18%. E%2 were observed. For ER-alpha the value of E%1 rose up from 30% to 58% and the variation of the E%2 was only of the 3%. The best results were observed for MMP-13 ($\Delta E\%1=+41\%$, $\Delta E\%2=+18\%$) and CDK-4 ($\Delta E\%1=+9\%$, $\Delta E\%2=+6\%$). Also the PTP-1B target showed a significant variation of the E% values with a $\Delta E\%1=+25\%$, $\Delta E\%2=+14\%$.

In the light of these considerations, the best results and the strongest variation between the old approach and the Conf-VLKA occur for dataset compounds with a high degree of branching considered as number of rotamers. This could be justified by the fact that the most branched is the molecule the most it will be sensible to conformation variations and the best it will be represented by 3D descriptors, as demonstrated by Good et al. in 2004 [97, 98].

3.1.4 Conclusions

In this paper, we modified the previous *in house* developed VLKA protocol in order to analyse the ligand conformational effect on the protocol capability, in particular, calculating 3D molecular descriptors on the docked conformation of ligands. Our VLKA protocol was designed to predict the possible biological target for new molecules starting from the structural information contained in molecular descriptors calculated on a set of known inhibitors. This first protocol was able to correctly predict biological target for the whole dataset with a good degree of reliability (80%) [24], and revealed experimentally useful to optimize the biological activity of some pyrimidine derivatives [99, 100]. Applying the structure based approach to VLKA we observed that, the use of the simple docking scores instead of molecular descriptors, revealed not satisfactory results, instead, the Conf-VLKA showed slightly better results (86%) than the first VLKA protocol for certain target sets, for others no interesting variations were observed. On the light of these considerations, it seems like the conf-VLKA approach works slightly better, compared to the previous protocol, when applied to targets whose ligands present a highly branched structure. According to what already found by Good et al., in a work on the effect of chemical structure complexity on molecular descriptors weight for ligand-based virtual screening [97, 98]. Another issue to be addressed is that probably the performance of the Conf-VLKA is connected to the docking algorithm that works better on some proteins more than others. VLKA and Conf-VLKA revealed different strength points. While VLKA revealed really fast and immediate to apply, Conf-VLKA, although need more computational time, on some proteins revealed a small rise in performance, especially for systems in which compounds have a great number of torsional bonds and branching. Nevertheless, both approaches (VLKA and Conf-VLKA) are totally user-defined, so that it is suitable for the use of *in vacuo* descriptors calculation or the descriptors calculation based on binding conformation of ligands. This work is a first preliminary study on the ligand conformational effect on the VLKA protocol capability. We are now working on the same protocol using induced fit docking in order to take into account the target flexibility induced by ligands.

4. DYNAMIC APPROACH TO VIRTUAL SCREENING

Proteins are constitutionally flexible molecules. They exert their biological function undergoing various conformational changes more or less wide. This aspect covers a huge importance for the exploration of protein – ligand interactions [50, 101]. The receptor and ligand flexibility and the induced conformational changes should be considered to correctly estimate the binding mode and the thermodynamics behind the binding process [102]. Unfortunately, drug design and virtual screening campaigns often neglect these aspects, using a static representation of the protein target. Several approaches have been introduced in computational chemistry software to take into consideration protein flexibility [103, 104]. The most representative are: side-chain flexibility [105], soft docking, induced fit [106, 107] and conformational ensemble-based docking [108, 109]. A correct incorporation of protein dynamics for drug design is still a challenging task. It has been shown in many cases that including protein flexibility leads to higher rates of false positives, since a larger number of putative ligands can be accommodated into different conformations of the binding pocket [110, 111].

Frequently, virtual screening protocols are set up on a conformational ensemble of proteins in order to include protein flexibility. Such an approach is based on behalf of proteins existing as an ensemble of substates of activation represented by different conformations [112, 113]. The main step of this approach is the generation of protein conformational ensembles prior to docking and the subsequent binding simulation of small molecules within the protein binding pocket of different size/shape [114, 115]. However, the approach strongly depends on the sampling quality chosen. One of the biggest limitations in using static X-ray or NMR receptor structures is that the available experimental conformations may not be sufficient to represent suitable conformations of the binding site for correct prediction of accommodation of new ligands [116, 117]. Despite the various methods adopted to sample protein flexibility, it is still difficult to collect suitable receptor conformations to be used prior to virtual screening processes [118, 119].

Often, protein conformations are collected starting from MD simulations [120–123]. One of the recently developed method to use MD prior to virtual screening is presented in the Relaxed Complex Scheme (RCS) approach [124]. Another MD-based approach is based on the sampling of the Receptor Conformation Ensemble,

appropriate for accommodating ligands, which are chemically and structurally diverse and thus unbiased toward a particular class of ligands [125]. Recently, this approach was successfully employed for ligand profiling of drug metabolizing enzymes sulfotransferases. In their work, Martiny et al. adopted docking on RCE generated by MD simulations, combined with hierarchical conformational clustering of different binding site conformations [126].

Another interesting approach, adopted by Rueda et al. is to explore collective movement-based conformational changes [127]. In his work, Rueda exploits cross-docking on the ensemble structures generated by MD. In the last years, receptor flexibility has been also assessed using a potential grid representing the receptor deformed through selected collective movements and global structural changes following ligand binding [128]. However, considering a large number of modelled conformations may sometimes lead to less predictive VS results compared to those obtained by using the best performing crystal or NMR structures, due to the possible generation of non-native protein-ligand conformations [129, 130].

Applying these concepts to structure-based pharmacophore screening, it is important to point out that the pharmacophore model is sensitive to the atomic coordinates of the protein-ligand complex from which it was derived. The first issue is closely linked to the source of the coordinates for the protein-ligand complex, whose coordinates are usually taken from the Protein Data Bank (PDB) [93]. Very often, the protein structures solved by X-ray crystallography may be affected by errors such as crystal contacts and solvent effects; for this reason, the reliability of protein-ligand coordinates has been frequently questioned [131, 132]. Proteins and small molecules are inherently dynamic and undergo a wide range of motions, ranging from the vibrations of individual bonds to collective, large structural movements. The crystal structure of the protein-ligand complex represents only a single snapshot of a dynamic system, providing neither information about the conformational flexibility of the ligand, nor about motions of the residues in and near the binding pocket [133, 134]. Thus, pharmacophore models derived from such structures might include features that are artefacts, caused either by crystal packing effects or by the single set of coordinates of the structure. Moreover, these PDB-derived pharmacophore models could contain too few or too many features resulting in a limited use. Increased number of pharmacophore features are normally accompanied by a loss of sensitivity, and usually pharmacophore models composed of more than seven

chemical features are not suitable for database screening [135]. In this regard, the most important issue becomes the choice of reliable criteria to prioritize them. In the last years, several efforts have been made to integrate the natural dynamic behaviour of proteins in pharmacophore models. One proposed approach was based on the multi-complex pharmacophore models. Here, the models were derived from multiple crystal structures of the same protein in contact with different small molecules. Protein-ligand interaction patterns were extracted from the available structures and merged in pharmacophore maps [135, 136]. This approach, however, is limited to proteins for which multiple crystal structures are available and which have the same binding mode: it does not really consider the dynamics of the ligand-protein complex.

One very general way to avoid dependence on a single set of coordinates is the use of molecular dynamics (MD) simulations to generate multiple sets of coordinates and use these as the basis for pharmacophore models. MD simulations have proven invaluable to understand the dynamics of biomolecules [137–139]. Several approaches have been proposed to generate trajectories of protein-ligand complexes, subsequently clustered to extract representative structures as reliable pharmacophore models. [140, 141]. Most recently, Choudhury et al. presented a new way to build pharmacophore models from MD simulation. For each structure saved during MD simulations, one pharmacophore was generated and then ranked based on docking and screening results [60].

In the next chapters of this PhD thesis, I will present the chronological pathway of the study carried out to explore the use of a new approach to the pharmacophore screening: The Dynamic Pharmacophore.

In Chapter 4.1 I will present the results of a method based on the application of the MD prior to the pharmacophore modelling [142]. In Chapter 4.2 I will describe an approach based on the most frequent pharmacophore features retrieved from MD simulations and then adopted as pharmacophore model [143]. In Chapter 4.3 the application of the “most frequent features” method to the study of the IGF-1R (insulin-like growth factor-1 receptor) kinase domain is reported [144].

Finally, chapter 4.4 will concern an innovative approach to the dynamic pharmacophore model, based on a different starting point compared to those discussed in the previous chapters. Several PDB crystal structures were explored,

containing different ligands, to check the occurrence of a common interaction pattern and maintained during the MD simulations[145].

4.1 Comparing pharmacophore models derived from crystal structures and from molecular dynamics simulations

4.1.1 Introduction

The aim of this study was to compare pharmacophore models obtained from the crystal structure of a ligand-protein complex with the pharmacophore models derived from the last frame of a molecular dynamics (MD) simulation. The pharmacophore model obtained from the crystal structure of a ligand-protein complex was called “initial pharmacophore model”, models created from the last frame of the MD simulation has been defined as “MD pharmacophore models”. Considering the final structure of a given MD simulation is the most basic MD-refined structure refinement protocol. Even though the approach is simple, it can resolve some of the problems connected to protein-ligand structures obtained from X-ray crystallography [146–148]. In our opinion, the comparison between the initial pharmacophore model and the MD-refined models can give some crucial information for constructing a reliable pharmacophore models [142].

In this work we investigated two main issues:

- 1.) Is there any difference in terms of number and type of pharmacophore features comparing the crystal structure derived model with the MD-refined model?
- 2.) Is there a difference in the ability of the initial pharmacophore model and the MD-refined pharmacophore model to distinguish between active and decoy compounds?

The first question was answered by visual inspection of the obtained pharmacophore models. To answer the second question we screened active/decoy databases of the investigated protein-ligand complexes to calculate receiver operating characteristic (ROC) curves and enrichment factors [149, 150].

We analysed 6 different protein ligand systems (PDB CODE: 1J4H, 3BQD, 2HZI, 3L3M, 1UYG and 3EL8). Structures were chosen from the DUD-E database. This database provides datasets of known actives and calculated decoys for protein-ligand complexes. From these complexes using selection criteria that were governed by system size, number of ligands present and kind of ions in complex with the ligand or protein [151].

The virtual screening process was used starting from the pharmacophore model as a query for classification of compounds into decoy and active compounds, assigns score values and constructs a sorted list of these compounds using the score as key. The number of true positive compounds retrieved using a specific pharmacophore

model as opposed to the number hypothetically found if compounds were screened randomly was expressed as Enrichment Factor (EF) [152, 153] as defined in Eq. (4.1).

$$EF_{\text{subset}} = (tp_{\text{hitlist}} / (tp_{\text{hitlist}} + fp_{\text{hitlist}})) / (N_A / (N_A + N_D)) \quad (\text{Eq.4.1})$$

*Where: tp_{hitlist} is the number of true positive in the hitlist;
 fp_{hitlist} corresponds to the number of false positive in the hitlist;
 N_A and N_D are the number of active and decoy compounds in the testset.*

Enrichment factors can range from 1 - molecules are sorted randomly - to > 100, which means that only a small percentage of the order list needs to be screened in-vitro to find a large number of active molecules [149].

4.1.2 Materials and Methods

PDB quality control

The quality and correctness of the PDB structures was audited using the Quality Control server [154]. Modeller 9.15 was used if residues were missing [155]. Subsequently all structures were analysed with PropKa 3.1 in order to check the protonation state of the protein and the ligand [156, 157].

Molecular Dynamics

CHARMM-GUI was used to set up the simulations and the CHARMM software package to run them [158, 159]. The CGenFF and paramchem was used to obtain parameter and topology files for the small molecules [160, 161]. For all the CHARMM/openMM version was used to run molecular dynamic simulations for 6 protein-ligand complexes [162]. The systems were solvated in rectangular water boxes with TIP3P water molecules. Electrostatic interactions were computed by the particle-mesh-Ewald method. From the starting structures we carried constant pressure, constant temperature MD simulations (Berendsen thermostat and barostat). The length of each simulations was 20 ns; the time step was 2 fs and SHAKE was used to keep all bonds involving hydrogen atoms fixed. Before each simulation we equilibrated the protein-ligand-water system for 25ps with a time step length of 1fs.

RMSD calculation

The RMSD was analysed using the python package MDAnalysis [163]. For the protein the RMSD of the C-alpha atoms was calculated and for the ligand the RMSD of all heavy atoms was calculated against the initial PDB structure. Target and reference structures were aligned on C-alpha atoms before the RMSD was calculated, the ligand was not independently aligned.

LigandScout

For generating structure-based pharmacophore models and screening libraries LigandScout 4.09.1 was used. Screening was performed using the command line tool iscreen provided by LigandScout [164]. The screening database for the protein systems were generated using the decoys and actives from the DUD-E database [151].

4.1.3 Results and Discussion

Quality control of protein-ligand structures

For one protein (3EL8) it was necessary to add missing residues. Using the software Modeller 13 residues from residue number 411 to 423 were inserted [155]. The amino acid sequence was obtained from the DNA sequence of the protein from the NCBI database [165]. The protonation state and side chain orientation was set in accord with propka [156, 157] and the Quality Control Check provided by the Joint Center for Structural Genomics [154].

RMSD

For all protein-ligand systems the root mean square deviation (RMSD) for the protein and the ligand was independently calculated and is shown in Fig. 4.1. The ligand and the protein RMSD values were calculated with the aligned C-alpha atoms of the target and reference structure.

The RMSD of the protein and ligand was analysed to detect large scale movements of the protein or the ligand. In addition, we used the deviation of the ligand to determine if the ligand reaches a stable binding state. The RMSD plots of the different ligands show very similar behaviour. The RMSD usually changes in the beginning to an average value from which the ligand deviates only marginal. This transition happens fast, e.g. 2HZI reaches the average value of 1.03 Å in less than 0.1 ns and has a standard deviation of 0.2 Å from the mean. The ligand of 1J4H is the

only exception - it takes nearly 2.5 ns to reach the stable plateau around the average value of 1.58.

For the protein the behaviour of the RMSD was in the range of normal conduct during a MD simulation.

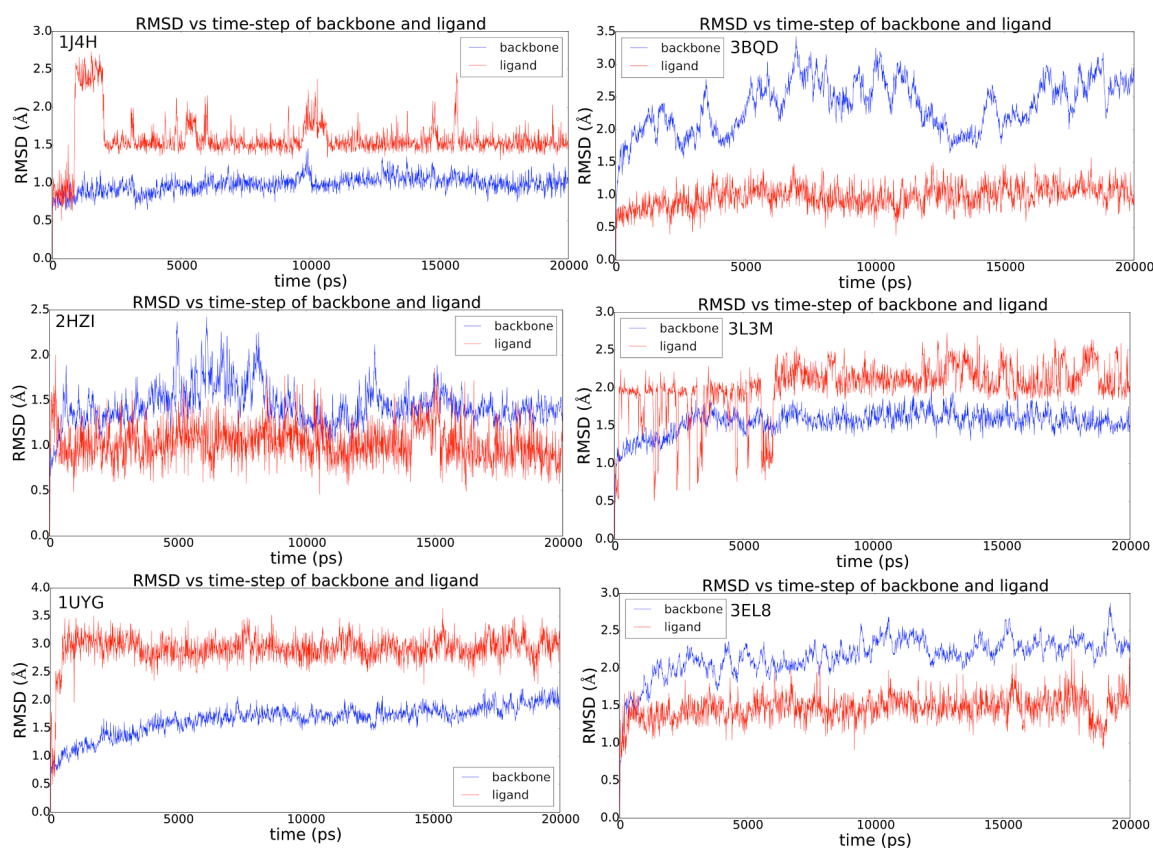


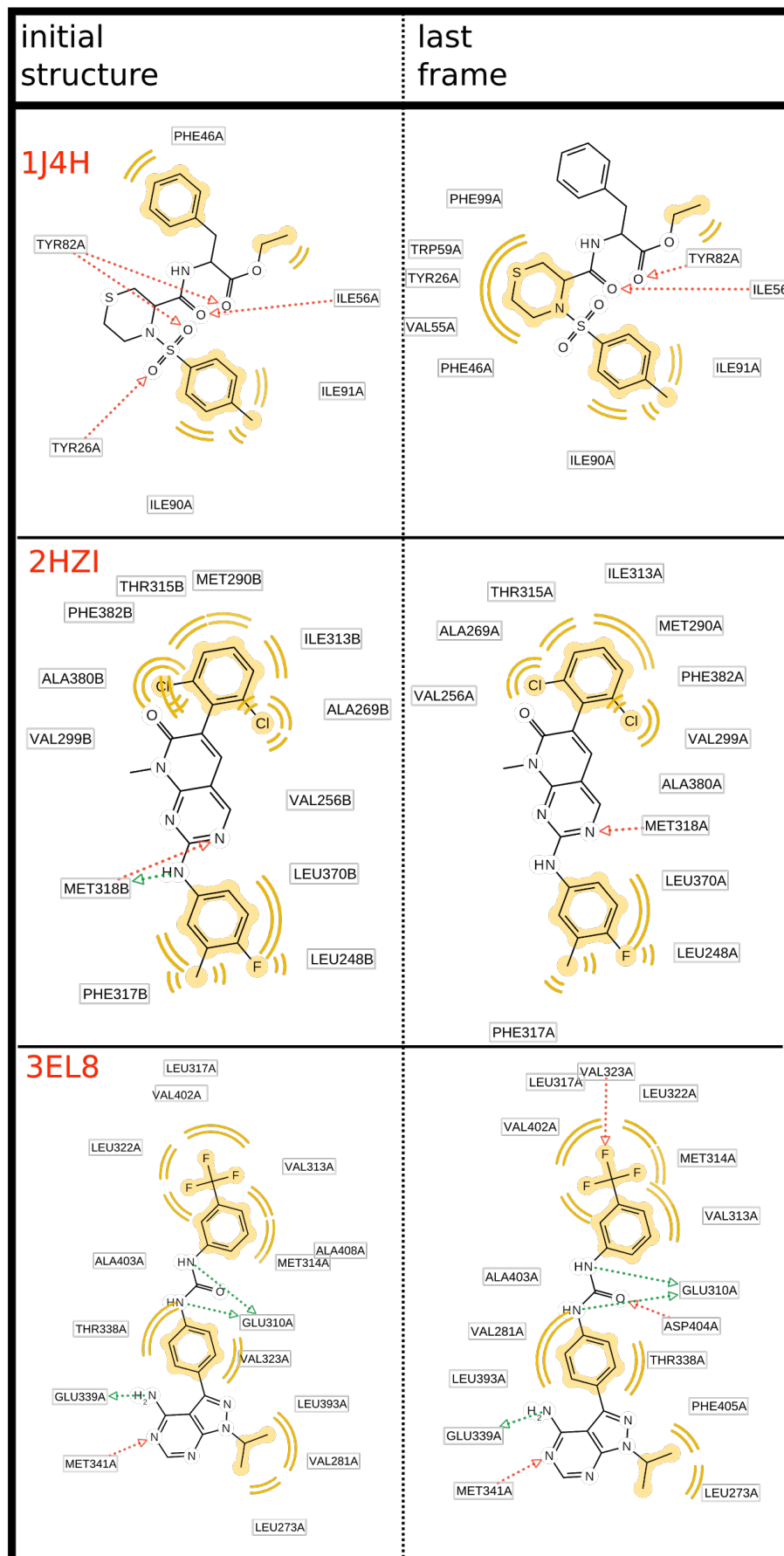
Fig 4.1 The root mean square deviation (RMSD) of the protein (in red) and the ligand (in blue) is provided as a function of time for the six analyzed protein-ligand complexes. The RMSD is calculated against the aligned PDB structure. The protein-ligand complex is aligned based on the protein backbone. For all systems the ligand and the protein experiences a rapid RMSD deviation from the original structure of at least 0.5 Angstrom (Å). The different RMSD ranges on the y-axis should be noted.

Comparing pharmacophore models

In Fig. 4.2 the 2D view of the ligand together with the assigned pharmacophore features is reported. The pharmacophore model obtained from the PDB file and the MD-refined pharmacophore model are shown for every protein-ligand system.

For all analyzed systems the initial pharmacophore model and the MD-refined pharmacophore model differs. Of the six analysed systems the amount of pharmacophore features for the initial model decreased in three cases, in one case the amount of features (but not the kind of features) stayed the same and in two cases the amount of features increased compared to the pharmacophore model obtained with

the MD-refined pharmacophore model. Looking at specific feature types it is interesting to note that hydrophobic features do not change in amount nor in involved atoms. In contrast none of the aromatic features are present in the MD-refined pharmacophore model. Most of the variability in the pharmacophore features was found to be due to hydrogen bond acceptors and donors.



cont.

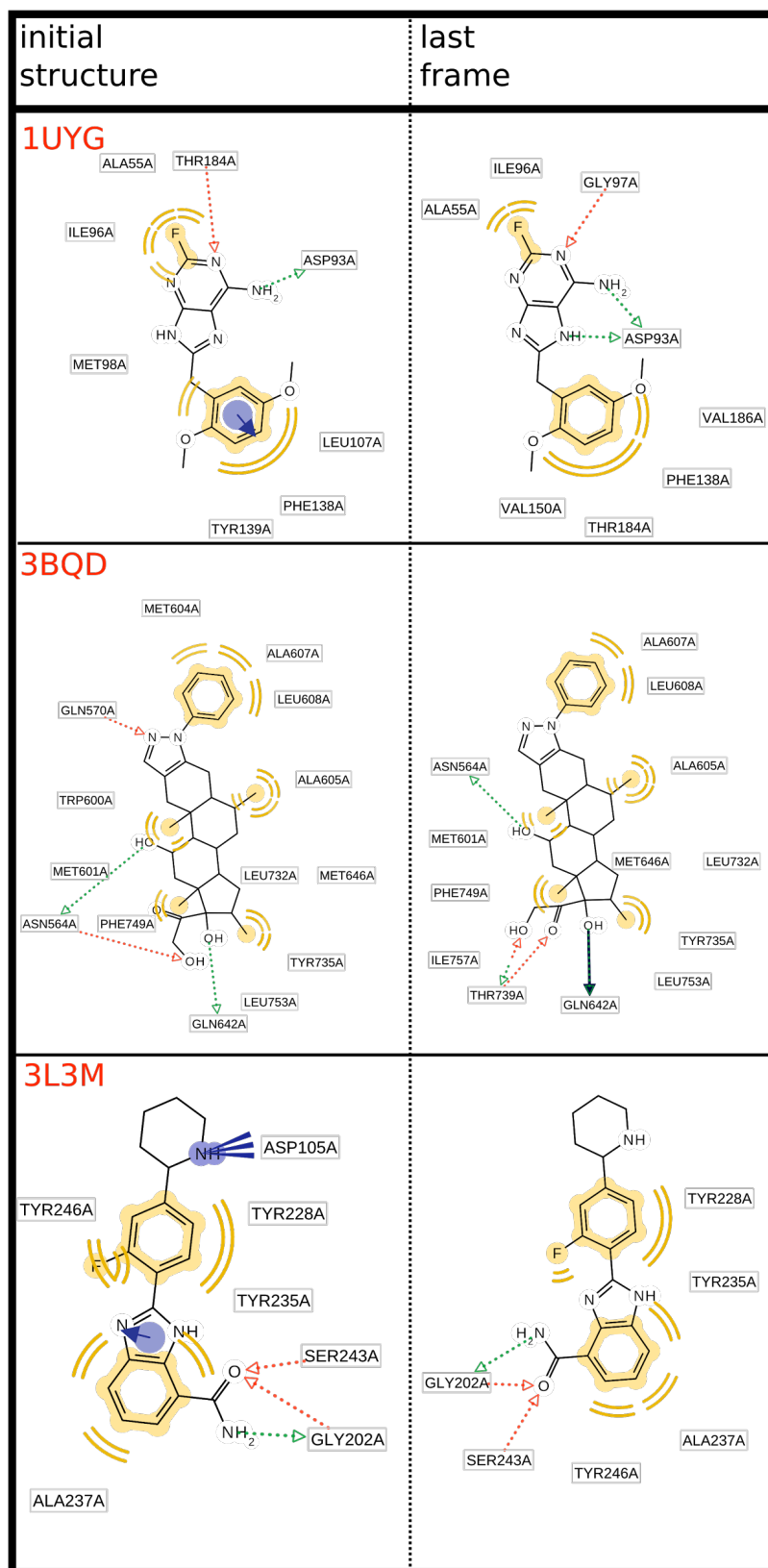


Fig 4.2 Comparing the initial pharmacophore model and the MD-refined pharmacophore model. The features in yellow indicate hydrophobic features, the vector features in red indicate hydrogen bond acceptors, the vector features in green indicate hydrogen bond donors, the feature spheres in blue with associated vectors indicate aromatic features and the features in blue with multiple lines associated indicate salt bridges.

Virtual screening results

In Fig. 4.3 the ROC curves for the different protein-ligand systems are shown.

For 1J4H the initial pharmacophore model and the MD-refined pharmacophore model cannot distinguish between actives and decoys.

For 1UYG the MD-refined pharmacophore model can distinguish between active and decoy compounds. With one omitted feature the overall ability to separate actives and decoys is better than with zero omitted features, but the enrichment factor for the first percent is lower. The initial pharmacophore model with zero omitted features can distinguish between active and decoy compounds for the first percent of the results, but above the 5% mark it favors decoys over actives. The model with one omitted features has among the top ranking results only false positive compounds but after the 1% mark it favors actives over decoys.

The MD-refined pharmacophore model for 2HZI favors active over inactive compounds for zero, one and two omitted features. This is not always visible in the ROC curve but looking at the enrichment factor it becomes clear that even the pharmacophore model with zero omitted features favors actives. The model with one omitted features favors actives only in the highest ranking results, the model with two omitted features favors actives for all results. The initial pharmacophore model, with one and two omitted features, favors actives.

For 3BQD the MD-refined pharmacophore model with one and two omitted features has a high enrichment factors (27.9 and 20.2 for 100%) as well as the initial pharmacophore model with one or two omitted features of 18.6 and 6.6 for 100%.

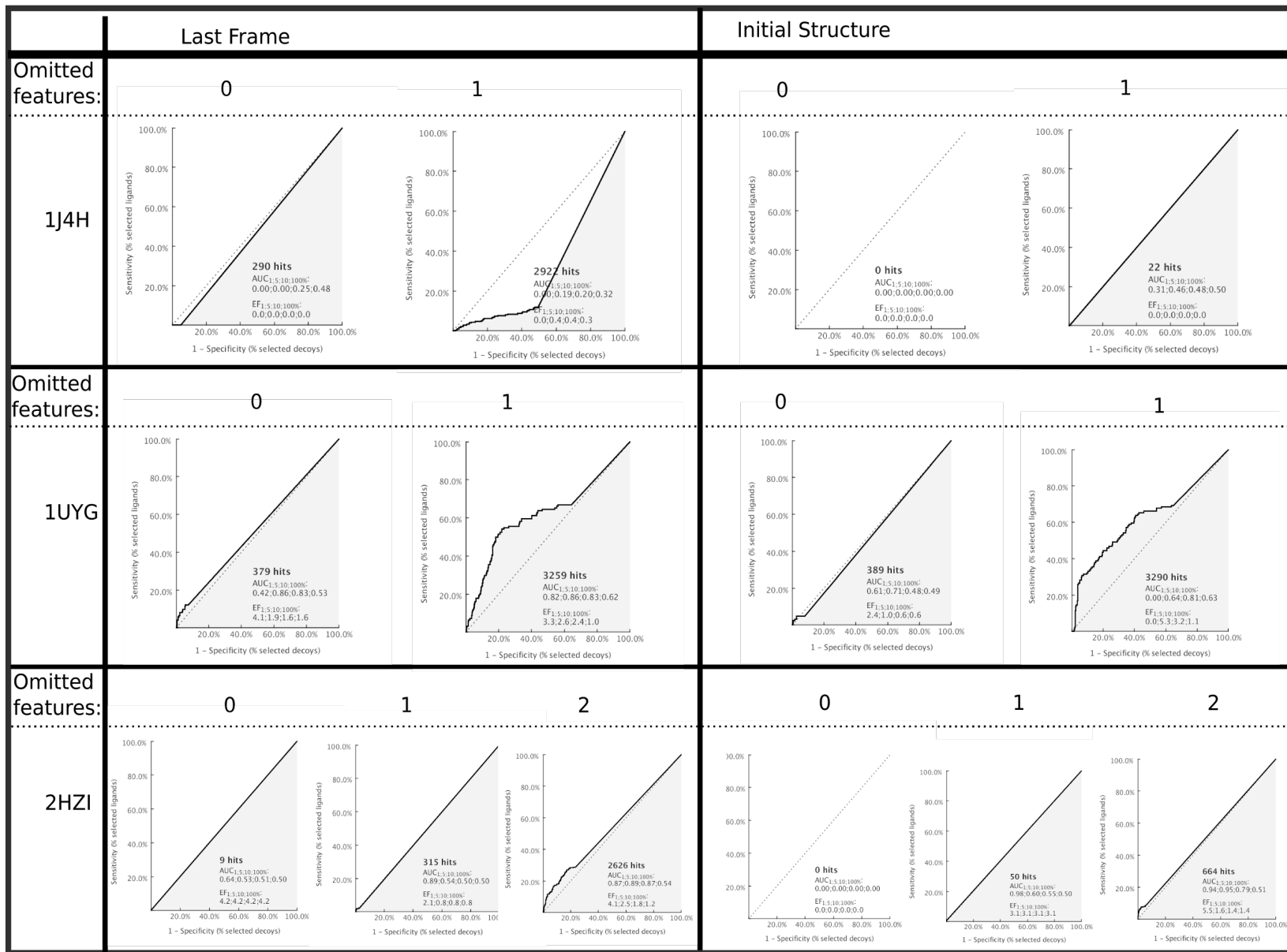


Fig 4.3

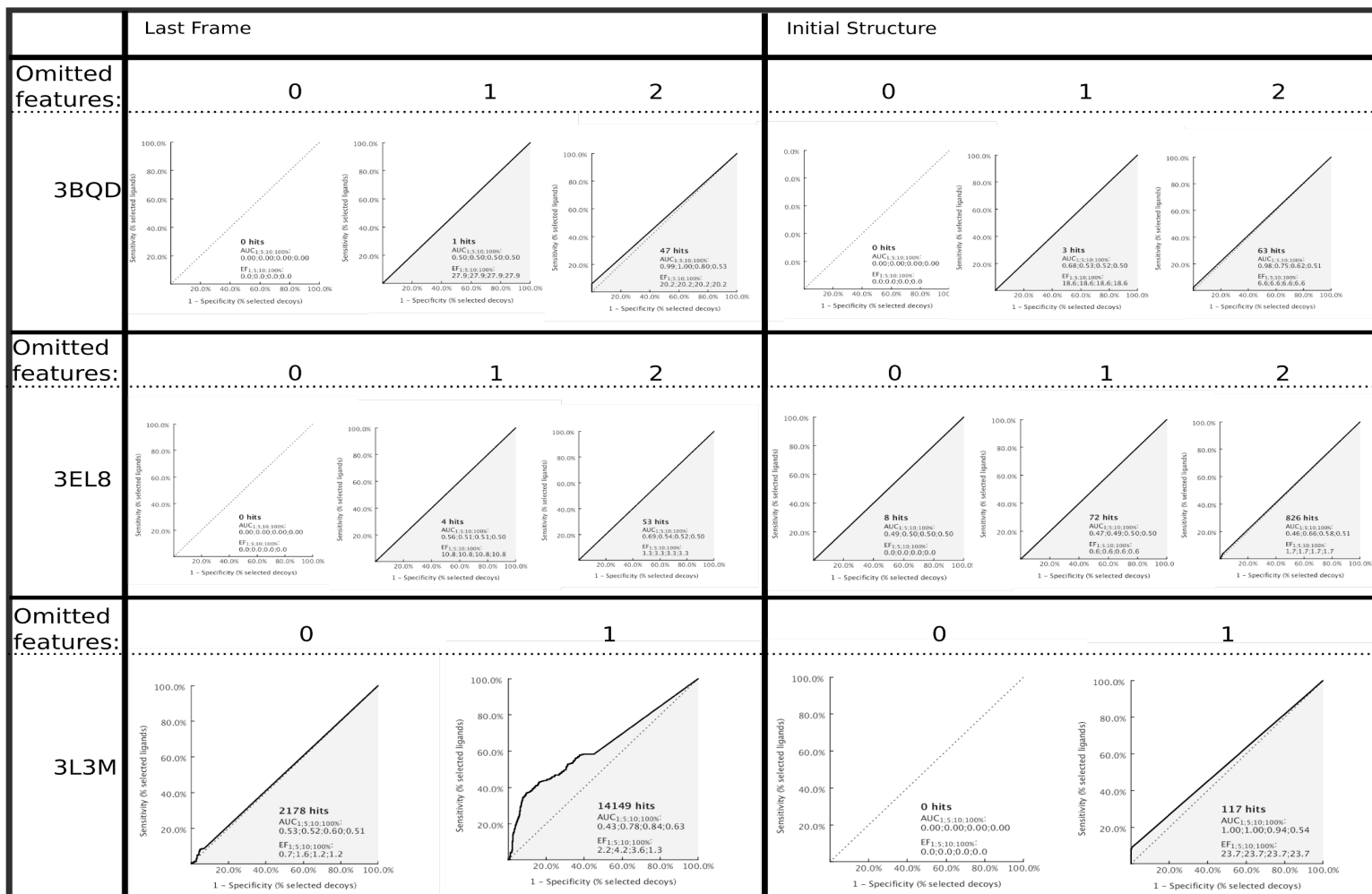


Fig 4.3 The receiver operating characteristic (ROC) curve for the different protein-ligand systems is shown. The true positive rate is seen on the Y axis and the false positive rate on the X axis. The number next to the PDB code indicates the number of omitted features: 0 means that no features were omitted, 1 or 2 means that either one or two features were omitted during the screening. In the plots the number of total hits, the area under the curve (AUC) and the enrichment factor (EF) is shown at 1%, 5%, 10% and 100%. For 3EL8 the MD-refined pharmacophore model with one and two omitted features has high enrichment factors (10.8 and 3.3 for 100%) whereas the initial pharmacophore model with zero or one omitted features has no preference for actives, with two omitted features the model has a slight overall preference for active compounds (ER: 1.7 for 100%)

The MD-refined pharmacophore model for 3L3M with zero omitted features has an overall preference for actives, but this effect is only marginal. The same model with one omitted features has a significant early enrichment but the sensitivity decreases after the 5% mark. The initial pharmacophore model with zero omitted features is not able to return any results, with one omitted feature the model has a good early enrichment (23.7%) with constant sensitivity.

The screening results obtained from the MD-refined pharmacophore model and from the initial pharmacophore model are different. With the exception of 1J4H, for which both pharmacophore models performed badly, either the refined pharmacophore model or the initial pharmacophore model were able to favor active compounds over inactive ones - in some cases e.g. 3BQD both were able to distinguish between the groups. Depending on the preferred result (early enrichment vs overall enrichment factor) the interpretation of the overall performance of the two approaches can vary. Simply looking at early enrichment (considering only the enrichment factor above 1% of the total compounds) the pharmacophore model obtained with the MD-refined pharmacophore model performs better for 1UYG, on average for 2HZI, for 3BQD and for 3EL8. The initial pharmacophore model performs better for the screening on the compounds for 3L3M.

Considering the enrichment factor at 100% of analysed compounds the MD-refined pharmacophore model performs better for 1J4H (even though still badly), on average for 1UYG, on average for 2HZI, for 3BQD. In the analysed cases the overall enrichment factor mirrors the results obtained from the early enrichment results.

4.1.4 Conclusions

The findings reported in this work suggested that the refinement of pharmacophore models using molecular dynamic simulations is an important starting point for better exploring ligand protein interactions. Even very simple structure refinement approaches, like the reported one, lead to pharmacophore models that return on average better screening results.

Additional interaction information can be unveiled analysing the dynamic behaviour of protein and ligand and harvesting these information can lead to better pharmacophore models that can target specific binding sites or interact with transitional conformations. For this reason in the future we aim to deepen the real

dynamics of the pharmacophore feature assessing a study of the time evolution and frequency of the encountered features, this is so far, to our knowledge, an approach still not investigated by the computational chemists.

4.2 Evaluating the stability of pharmacophore features using molecular dynamics simulations

4.2.1 Introduction

In this work we investigated the dynamics and the stability of the structure-based pharmacophore out of 12 protein-ligand complexes. Starting from an MD simulation a pharmacophore model for each structure produced during the simulation was generated. Through the creation of a called *merged* pharmacophore model we took into consideration all features that are seen either in the experimental structure (PDB X-ray structure) or any of the snapshots generated during the MD simulation. Thus, it incorporates the dynamics of the protein—ligand complex. The frequency with which individual features could be a useful way to prioritize the features if needed and to detect which ones only appear rarely.

For each system, a 20ns simulation in aqueous solution was carried out, and a merged pharmacophore model was derived as just outlined. In this proof-of-concept work we focus foremost on exploring three key questions concerning the viability of our approach.

1. What are the differences between the traditional pharmacophore models constructed from the PDB structures (the *PDB* pharmacophores) and the merged pharmacophore model?
2. Are the features arising most frequently during the MD simulation also present in the PDB pharmacophore model?
3. In addition to answering the first two questions for each of the complexes studied, we also explore how the four principal types of pharmacophore features (hydrophobic features, aromatic features, ionizable group features and hydrogen bond features) behave in this respect.

As mentioned before, the information about pharmacophore feature frequencies may aid in prioritizing features. For example features that are present in the pharmacophore model derived from the PDB structure, but occur only rarely during the MD simulation (e.g., less than 5% of the time) might represent artifacts and should possibly be discarded. Conversely, features that are not present in the PDB structure, but appear very frequently during the MD simulation (e.g., more than 90% of the time) should be regarded as important. Even though the frequency information alone may not be enough to rank features, it can help make an informed decision which features to keep/add and which to ignore, particularly if a pharmacophore model has too many features.

4.2.2 Materials and Methods

For this study twelve different protein-ligand complexes were selected from the PDB: 1J4H, 1XL2, 2HZI, 3L3M, 1UYG, 3EL8, 2GTK, 2P54, 3BQD, 2AZR, 2OJ9 and 2OJG. The choice of complexes was somewhat arbitrary, though guided by the following considerations: system size (solvated protein-ligand complex less than 70,000 atoms), only a single ligand, no metal ions involved in the binding. The complexes will be referred to by their PDB code. The following terminology will be used. The pharmacophore model obtained based on the experimental structure is referred to as *PDB* pharmacophore model, features specific to a PDB model as PDB features. In the *merged* pharmacophore model all observed pharmacophore features are mapped on the ligand, the merged model includes features present in the crystal structure, as well as features occurring during the MD. Features not present in the crystal structure, i.e., seen only during the MD simulation, will be referred to as *MD derived* features.

PDB quality control

The quality and correctness of the PDB structures was audited using the Quality Control server [154]. Modeller 9.15 was used if residues were missing [155]. Subsequently all structures were analysed with PropKa 3.1 in order to check the protonation state of the protein and the ligand [156, 157].

Molecular Dynamics

We used CHARMM-GUI to set up the simulations and the CHARMM software package to run them [158, 159]. The CGenFF and paramchem was used to obtain parameter and topology files for the small molecules [160, 161]. For all the CHARMM/openMM version was used to run molecular dynamic simulations for 6 protein-ligand complexes [162]. The systems were solvated in rectangular water boxes with TIP3P water molecules. Electrostatic interactions were computed by the particle-mesh-Ewald method. From the starting structures we carried constant pressure, constant temperature MD simulations (Berendsen thermostat and barostat). The length of each simulations was 20 ns; the time step was 2 fs and SHAKE was used to keep all bonds involving hydrogen atoms fixed. Before each simulation we equilibrated the protein-ligand-water system for 25ps with a time step length of 1fs.

RMSD calculation

The RMSD was analysed using the python package MDAnalysis [163]. For the protein the RMSD of the C-alpha atoms was calculated and for the ligand the RMSD of all heavy atoms was calculated against the initial PDB structure. Target and reference structures were aligned on C-alpha atoms before the RMSD was calculated, the ligand was not independently aligned.

Pharmacophore generation

For generating structure-based pharmacophore models and screening libraries LigandScout 4.09.1 was used to generate a structure based pharmacophore model for each frame saved during the MD simulation (2000 pharmacophore models) and for the PDB structure [164].

The resulting 2001 pharmacophore models for each protein—ligand complex were analyzed as follows. Each pharmacophore feature has two properties: the ligand atoms that are part of the feature and the feature type. If both properties of a pharmacophore feature were present in two models, then this feature was considered identical and the frequency count of this specific feature was incremented. In this manner we obtained statistics how often a certain feature was present during the course of the MD simulation. Separate statistics were made for features not present in the PDB pharmacophore model, i.e., features only seen during the MD simulation. Using this frequency information, the merged pharmacophore model encompassing all features seen during the simulation was constructed by mapping the features on a representative 2D and 3D structure of the ligand.

4.2.3 Results and Discussion

All trajectories were inspected visually to ensure that no large scale movement took place and that the ligand remained within the binding site at all times. For all twelve systems, the RMSD of the C_α-atoms was in an acceptable range. The same was true for most ligands, except for 2OJ9 and 2OJG. In the case of 2OJ9, the imidazole and pyridine moieties of the ligand rotated freely during the MD simulation, causing an average ligand RMSD of 5.6 Å. Similarly, the dimethylamino and phenyl group of the ligand in 2OJG was highly flexible. Nevertheless, even in those two cases the protein-ligand complexes were stable during the simulation.

In Fig. 4.4 the merged pharmacophore models of all twelve systems are shown. The models contain all encountered features, and the frequency (in percent) with which a feature occurs is given (the numbers in the small boxes in Fig. 4.4). This schematic

representation provides an overview of the stability / robustness of the individual features in the twelve complexes.

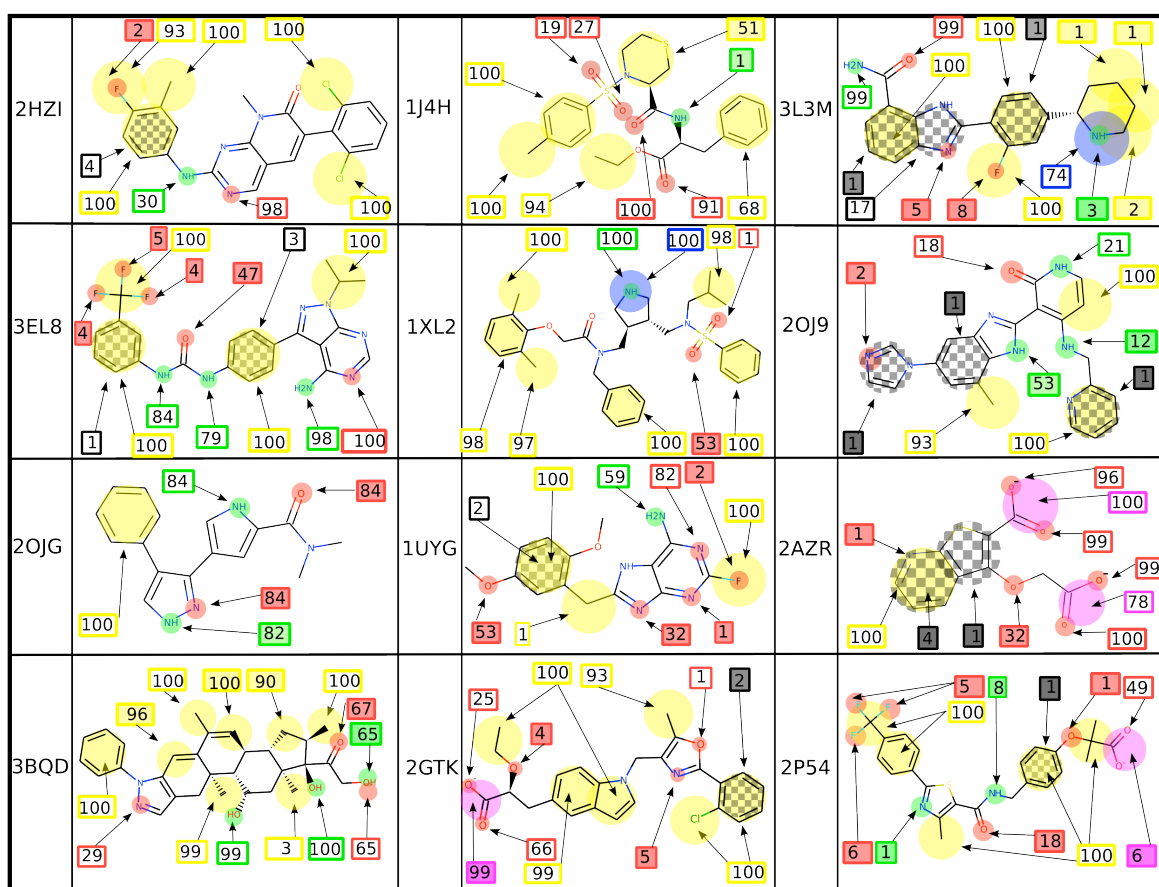


Fig.4.4 Merged pharmacophore models mapped on the 2D representation of the ligand for all twelve protein-ligand systems studied. Individual figures are labeled by the PDB code. Feature types are color-coded as follows: yellow spheres indicate hydrophobic features (H), grey/white chessmate spheres indicate aromatic interactions (AR), small green spheres indicate hydrogen bond donors (HBD), small red spheres indicate hydrogen bond acceptors (HBA), blue spheres indicate positively ionizable groups (PI), and big pink spheres indicate negatively ionizable groups (NI). The numbers in boxes indicate the frequency in percent with which a feature is found in each of the individual pharmacophore models from which the merged model was constructed (cf. Methods). The box color indicates the feature type: H (yellow), AR (black), HBA (light green), HBD (red), PI (blue), NI (pink). Filled boxes indicate features not present in the crystal structure, i.e., which appear only in the course of the MD simulation

Even a cursory inspection of Fig 4.4 shows that in all protein-ligand systems there are MD derived features, i.e., pharmacophore features which appear during the MD simulations; these are indicated by the shaded background of the boxes listing the frequency information. The number of the MD derived features, however, varies considerably, from one in, e.g., 2HZI up to five in, e.g., 2P54. Similarly, MD derived

features can occur very rarely (less than 5% of the time), see, e.g., 1J4H, as well as very frequently (more than 90% of the time), e.g., 2GTK.

The merged pharmacophore models shown in Fig 4.4 can be roughly divided into two groups: (1) pharmacophore models, for which the MD simulation added little new information, i.e. any MD derived features had low frequencies, and (2) pharmacophore models, for which MD simulation revealed information that can be helpful for further work with the models, i.e. the model has PDB features that disappeared completely during the MD simulation or occurred only infrequently and MD-derived features that appeared with high frequencies.

We considered 2HZI, 1J4H, 3L3M, 2AZR and 2P54 to be members of group (1). All of them display high frequencies for most PDB features and low frequencies for MD-derived features. In particular, 2HZI and 1J4H are prototypical members of this first group: only one (2HZI) or two (1J4H) MD derived features were observed, and their frequencies were very low. By contrast, the assignment of the other three complexes to group (1) was a bit more ambiguous. E.g., there were eight MD derived features in 3L3M, and one PDB feature only had a frequency of 17% during the MD simulation. However, all MD derived features were observed very rarely (1-8%), and the “low-frequency” PDB feature is an aromatic feature. As will be discussed below, aromatic features in general tended to occur with low frequencies during the MD simulations; thus, 17% is a relatively high value for an aromatic feature. Because of this, we feel that 3LM1 is best assigned to group (1); in the case of 2AZR and 2P54, the situation is similar.

We considered 2OJG, 3BQD, EL8, 1XL2, 2OJ9, 2GTK and 1UYG to belong to group (2). All of them have either PDB features that occur rather infrequently during the MD simulation (1XL2, 2OJ9), MD-derived features with high frequencies (2OJG, 3BQD) or both (3EL8, 1UYG, 2GTK).

Two examples for group (2) should be considered in detail: the pharmacophore model for 2OJ9 and 2OJG. The pharmacophore model for 2OJ9 has one hydrogen bond acceptor and one hydrogen donor PDB feature with frequencies below 25%. The three other PDB features - all of them hydrophobic features - have frequencies above 90%. There are six MD-derived features. While one of the MD-derived hydrogen bond donor features has is seen more than 50% percent of the time, the other MD-derived features have low frequencies. For this pharmacophore model the

MD simulation shows that some of the initial PDB features are not stable during the MD simulation.

The pharmacophore model for 2OJG has only two PDB features, one hydrophobic and one hydrogen bond donor feature, whereas the merged model contains three additional MD-derived features (one additional hydrogen bond donor feature and two hydrogen bond acceptor features). All three had a high frequency of occurrence (>70%). Their absence in the PDB pharmacophore model may have been caused by an unfavorable pose of the ligand in the binding pocket of the crystallized protein. The members of groups (2) provide good examples how dynamics influences the pharmacophore hypothesis. On the one hand, PDB features which exhibit low frequencies during the MD simulation could be artifacts of the initial protein-ligand coordinates. If these features are kept for virtual screening, they could result in erroneous/misleading hits. On the other hand, MD derived features can be essential to construct a usable pharmacophore model. With the PDB pharmacophore model of 2OJG virtual screening would not be possible since a minimum of three features are normally necessary. By contrast, the merged model with five features would provide a usable starting point for virtual screening. While Fig 4.4 illustrates the effect of dynamics on the individual pharmacophore models of the twelve complexes studied here, Table 4.1 summarizes the behavior of feature *types* during the MD simulations. For each of the six basic pharmacophore feature types (hydrophobic (H), hydrogen bond acceptor (HBA), hydrogen bond donor (HBD), negatively ionizable (NI), positively ionizable (PI), and aromatic (AR)), first the total (=cumulative) number of occurrences in all twelve merged models is listed in Table 4.1 (column 'Merged features, total'). This number is then broken down into features present in the PDB model (columns 'PDB features' in Tab.4.1) and features only seen during the MD (columns 'MD derived features'). For each of these two groups, the total number of occurrences, as well as the numbers of occurrences present >90% and >50% of the simulation time are given.

Table 4.1. Overall occurrence of pharmacophore feature types in the 12 systems studied.

Feature a)	Merged features		PDB features		MD derived features		
	total b)	total b)	>90% c)	>50% d)	total b)	>90% c)	>50% d)
H	53	46	40	41	7	3	4
HBA	44	19	9	12	25	0	4
HBD	19	12	5	10	7	1	3
NI	4	2	1	2	2	1	1
PI	2	2	1	2	-	-	-
AR	13	4	0	0	9	0	0
Sum	135	85	56	67	50	5	12

a) For the meaning of abbreviated feature types see Fig. 4.4

b) Total count how often a feature type occurs.

c) Number of instances of a particular feature type present >90% during the MD simulation.

d) Number of instances of a particular feature type present >50% during the MD simulation.

Overall, PDB features appear more stable than MD derived ones. If one sums up all entries in column 'PDB features, >50%' of TAB 1 and compares to the sum of column 'PDB features, total', one finds that 79% (67 out of 85) of the PDB features are present 50% or more during the MD simulation. The same calculation for MD derived features gives 24% (12 out of 50). This distinction becomes even more pronounced if one repeats this analysis for all features present >90% during the MD simulation, which gives 66% for PDB features vs. 10% for MD derived features.

However, at the same time Table 4.1 illustrates that dynamics affects the various feature types quite differently. Consider e.g., aromatic features (AR). There are significantly more MD derived aromatic features (9) than aromatic PDB features (4). Both of them occur rather infrequently, i.e., for AR all entries in columns >50% and >90% are zero (last line of TABLE 4.1). The reverse situation is found for hydrophobic features (H). Here, there are substantially more PDB than MD-derived features, and, particularly in the PDB case, the robustness of the feature is high (40 out of 43 hydrophobic PDB features are present >90% of the time).

The AR and H features are the most extreme examples, the other four feature types (HBA, HBD, NI, PI) lie between these two. Clearly, aromatic features appear much more sensitive to small changes resulting from the motions of ligand and protein during the MD simulation compared to hydrophobic features. Hydrogen bond acceptor and hydrogen bond donor types follow the hydrophobic type in terms of prevalence. Similarly to the hydrophobic features, the stability of the PDB features is higher than for the MD-derived features. More than 70% of the PDB hydrogen bond features (HBA and HBD together) belong to the >50% group; for the MD-derived

hydrogen bond features this number drops to 22%. However, there are slightly more MD-derived hydrogen bond features (32) than PDB hydrogen bond features (31). In particular, the number of MD derived HBA features (25) is larger than that of the PDB HBA features.

The differences in stability between feature types have much to do with the definitions of / criteria for the various feature types.

The low frequency of the aromatic features is the consequence of the rules used to classify them: The geometric constraints on the aromatic ring plane are easily violated by flexible aromatic rings, e.g., rings that can rotate; furthermore, the rule set that atom groups on the protein side have to fulfil to be regarded as counterpart of an aromatic interaction are rather strict.

The relatively large number of MD-derived HBA features is the result of the chemical nature of the ligands (most ligands have multiple groups that can act as hydrogen bond acceptors) and the definition of hydrogen bond interactions. LigandScout uses angle and length criteria for the classification of hydrogen bond features; i.e., the interaction partner must be within a specified angle range and nearer than a certain distance threshold [164]. Thus, on the one hand, the constraints for hydrogen bond features are more rigorous compared to the hydrophobic feature, and a miniscule change in geometry can toggle whether an acceptor-donor pair is classified as a hydrogen bond or not. Consequently, the dynamical behavior of hydrogen bond features is different than that of hydrophobic features (data not shown). On the other hand, the amino acids acting as potential interaction partners for hydrogen bond acceptor features (threonine, tyrosine, lysine, cysteine, glutamine, serine, histidine, arginine) have mostly small atom groups that can rotate and move during the MD simulation. Thus, while a particular hydrogen bond is broken, the partner on the protein side can be often easily replaced by another amino acid.

In contrast the high stability of the hydrophobic features is also the consequence of the rules governing the classification as a hydrophobic feature. Specifically, in LigandScout, the interaction partner on the macromolecular side can be located anywhere between 1.0 and 5.9 Å away, and the requirements for atom groups on ligand and protein side to be considered part of a hydrophobic interaction are rather unspecific [164]. Finally, there is a small number of ionizable features present in the systems studied (four NI, two PI). Again, the stability of this features can be regarded as a consequence of the feature definition used. For a ionizable group in the

ligand to be considered an ionizable feature, LigandScout requires the presence of an opposite ionizable group on the macromolecule side within a distance of 1.5 to 5.6 Å. Given this generous constraint, this feature will be present during most of the simulation time, provided a counterpart is available at all [164].

4.2.4 Conclusions

We have shown that different pharmacophore feature types display varying stability during the MD simulations. On average PDB features are more stable than MD-derived features, but there are notable exceptions. These exceptions represent pharmacophore features that are not accessible using only the PDB structure and in one example (2OJG) these additional features were necessary for further work with the pharmacophore model.

We believe that the presented results indicate that the frequency information obtained using MD simulation can be used to refine the pharmacophore model (add/remove features) - yet we acknowledge the fact that frequencies between different feature types might not be comparable, at least not on a linear scale. Using these results as a first step we will continue our work on this topic and will investigate the possibility of pharmacophore model refinement using information obtained through MD simulations.

4.3 Pharmacophore models derived from molecular dynamics simulations: A case study

4.3.1 Introduction

In the previous chapter we investigated the possibility of improving pharmacophore model using molecular dynamic simulations to get the most frequent pharmacophore features. In this case study we present preliminary results that extend the analysis to one protein-ligand system for virtual screening. We analysed the variability of the interaction partners of the pharmacophore model and analyse the occurrence of features as a function of time. From this analysis two pharmacophore models are derived based on the frequency of interactions and the time resolved dynamics of the pharmacophore features.

The used protein-ligand complex was the PDB code 2OJ9 and represents the crystal structure of the IGF-1R (insulin-like growth factor-1 receptor) kinase domain in complex with a benzimidazole inhibitor. Overexpression of IGF-1R has been demonstrated in a variety of tumors, including glioma, lung, ovary, breast, carcinomas, sarcomas, and melanoma [166]. This protein-ligand complex was chosen from the analysed complexes in [143] because the pharmacophore model contains a balanced number of the most common features (3 hydrophobic features, 3 hydrogen bond donor features, 2 hydrogen bond acceptor features, 3 aromatic features).

Typically, most ligands of protein kinases bind in the hinge region at the folding cleft of the N- and C-lobes. Common scaffolds that bind this region contain two hydrogen bond features, usually a donor-acceptor pair that interact with the hinge backbone. The PDB pharmacophore model displays the typical hydrogen bond interaction pattern with MET103 and GLU101, as described in the literature [166].

As fully described in our published paper [143], starting from MD simulation, every 10 ps the coordinates were saved resulting in 2000 coordinate sets, we also considered the initial PDB one. A pharmacophore model was derived from each structure that was obtained during the MD simulation. For further analysis a consensus pharmacophore model (a merged pharmacophore model) was generated which consists of all features that are present either in the experimental structure or in any snapshot generated during the multiple MD simulations, thus it incorporates information about the dynamics of the protein-ligand complex. The frequency with which individual features are present permits to rank/prioritize the features if needed and to detect outliers, i.e., features seen only rarely. Additionally, the interaction

partners for each pharmacophore feature were analyzed and an interaction map (interaction matrix) was constructed. An interaction map allows to quantitatively analyze the interaction partners of pharmacophore features. As a final step the frequency of the pharmacophore features was analyzed as a function of time. Combining these different analysis methods for the dynamics of the pharmacophore features make allows to consciously derive pharmacophore models that are different than the corresponding model obtained with the PDB structure. Starting from the *merged* pharmacophore model, we run a virtual screening using an active and decoys database retrieved from DUD-E [151].

4.3.2 Materials and Methods

For the MD simulation methods and Pharmacophore generation I refer to the original papers describing the method [143] or to the previous chapter.

Virtual Screening:

Virtual screening was performed using known active and calculated decoy molecules obtained from the DUD-E database [151]. The database provided 226 actives and 9395 decoys. All molecules were prepared as libraries for the screening using the command line tool `idbgen` provided by LigandScout. Conformers were generated using the `icon` best option in `idbgen`, this option produces a maximum number of 200 conformations for each molecule processed.

Interaction matrix generation:

The columns of the interaction matrix indicate all amino acid residues that are involved in a pharmacophore feature at some point during the MD simulations, the rows designate all pharmacophore features and the values in the matrix indicate how often a specific amino acid was involved in a specific pharmacophore feature. In this way it is possible to analyze the number of interaction partners and also their statistical frequency. The numeric values were coded as a heat map – the colors range from blue (zero interaction) to dark red (interaction at every time step). The numeric values for hydrogen bond and aromatic interactions are given explicitly in the heat map for values below 400. The interaction map was generated using the python package `matplotlib` [167].

Frequency Plot generation:

For the MD simulations a frequency plot is calculated. This plot shows the occurrence of the features as a function of time. This is calculated as follows: The pharmacophore models are chronologically sorted and for every pharmacophore feature an occurrence list is calculated. Every time a pharmacophore model at a specific time step displays a specific feature 1 is inserted at the time step defined position in the list, otherwise 0 gets inserted. This results in a list with 2001 entries for every pharmacophore feature, which contains zeros and ones. In the end these lists are reduced by summing over chunks of 100 entries – resulting in a new list with 20 entries containing numbers between 100 and 0. To obtain a graphical representation, these lists are subsequently plotted using the python package matplotlib.

4.3.3 Results and Discussion

The trajectory of the protein-ligand complex was visually inspected to ensure that no large scale movements took place and that the ligand remained within the binding site at all times. The root mean square deviation (RMSD) values of the C α -atoms of the protein were in an acceptable range (the RMSD plot for the ligand and protein backbone is shown in Fig. 4.5). In contrast, the RMSD values of the ligand are rather high (ranging to a maximum of 9 Angstrom).

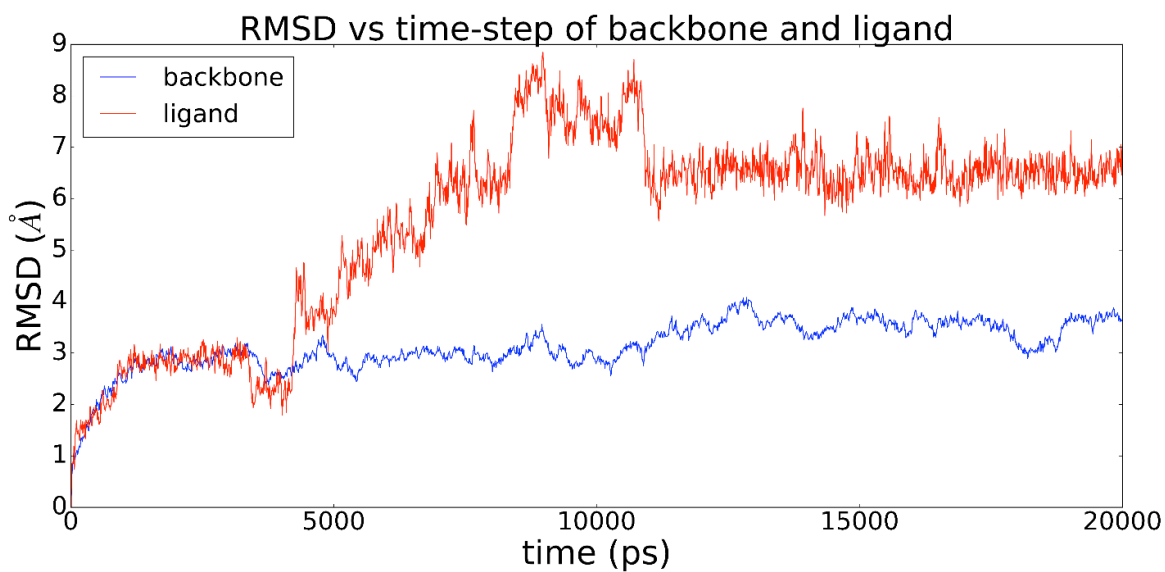


Figure 4.5. *The root mean square deviation (RMSD) in Angstrom (Å) of the protein backbone (in blue) and the ligand (in red) as a function of time for the analysed protein-ligand*

In Fig. 4.6 representative structures for the first (Fig. 4.6 (1)) to the fourth quarter (Fig. 4.6 (4)) of the MD simulation are shown. As can be seen the pyridine moieties of the ligand rotates freely during the MD simulation, but also the translation of the imidazole contributes to the elevated RMSD values. Nevertheless, the protein-ligand complex was stable during the simulation.

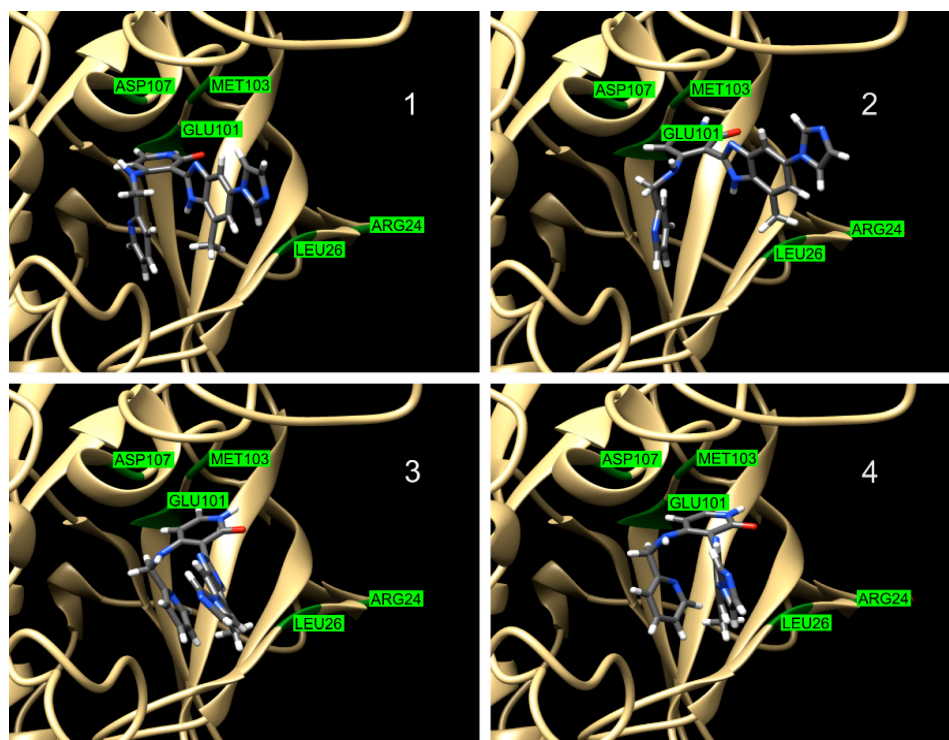


Figure 4.6. *The ligand inside the binding pocket is shown at 4 different timesteps. The length of the MD simulation is divided in 4 equally long parts and clustering is performed based on the RMSD of the ligand. A representative ligand structure is extracted from the most populated cluster and shown from 4.6 (1)(representative structure of most populated cluster from 0 to 5 ns) to (4) (representative structure of most populated cluster from 15 to 20 ns). The amino acids that are most common in hydrogen bond interactions are explicitly labeled*

The frequency of the specific pharmacophore features and the interaction map for 2OJ9 are shown in Fig. 4.7.

The initial pharmacophore hypothesis (shown in Fig. 4.7 A) includes 5 pharmacophore features (which will be called PDB features) and during the MD simulation 6 additional pharmacophore features (3 aromatic features, 1 hydrogen bond acceptor and 2 hydrogen bond donor features) are revealed (which will be called MD derived features).

As can be seen in Fig. 4.7 A, most of the MD derived pharmacophore features have a lower statistical frequency than the PDB features – only the MD derived feature HBD1 occurs more often than HBD3 or HBA2 (both are PDB features). A further observation that can be drawn from Fig. 1A is that hydrophobic features are far more stable during the MD simulation than hydrogen bond features and that aromatic features are the most unstable feature type. This finding is in accordance to our previous findings reported in [143].

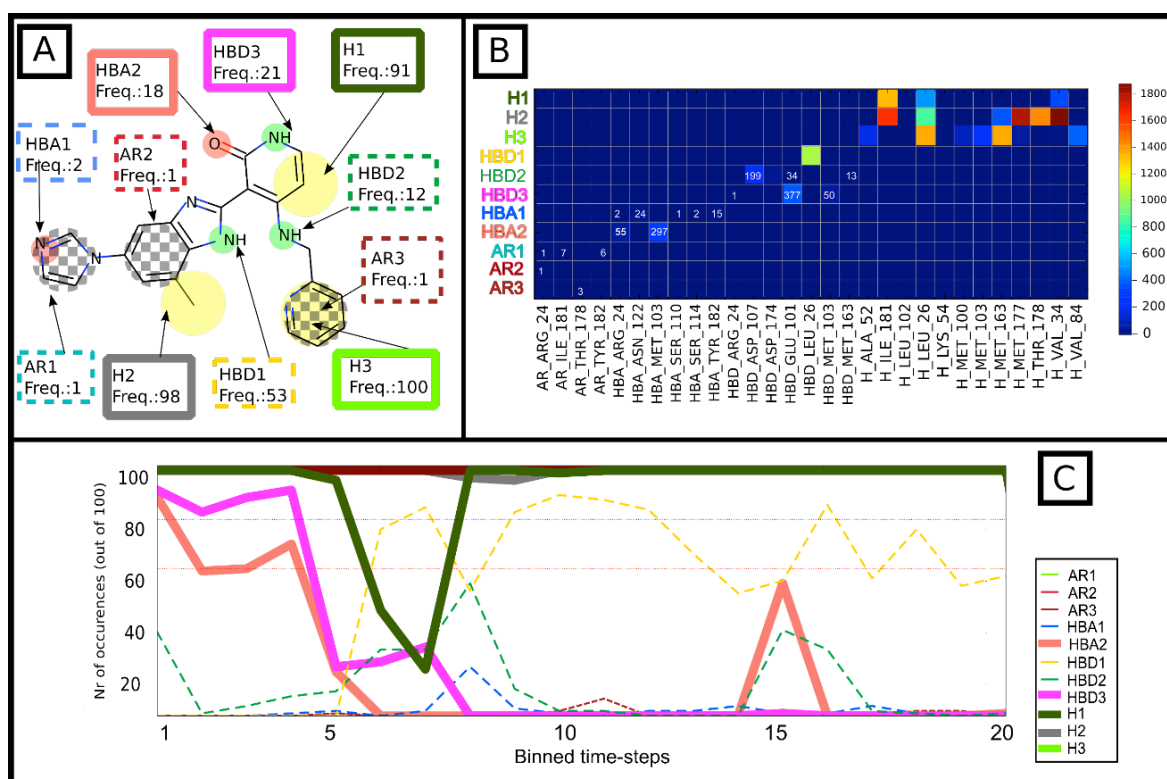


Figure 4.7. Analysis of the dynamics of the pharmacophore features of 2OJ9. (A) shows a 2D representation of the ligand with the pharmacophore features mapped on the structure. Yellow spheres represent hydrophobic (H) interactions, small green circles indicate hydrogen bond donor (HBD) features and small red circles hydrogen bond acceptor (HBA) features. Black and white chess-fields represent aromatic features (AR). For every feature a box is shown, providing the feature name that is used in part (B) and (C) of the figure and information about the statistical frequency (given in percent and rounded to integers) of the specific feature. Dashed outlined boxes indicate features that are not present in the PDB pharmacophore, continuous lined boxes indicate features that are present in the PDB pharmacophore. The color of the boxes are consistent with the colored row labels in part (B) and the colored lines in part (C) of the figure. (B) shows the interaction matrix as heat-map. The row names indicate the pharmacophore features and the column names the interaction partners of the pharmacophore features. The column names consist of three parts, separated by underscores: the first part indicates the feature type, the second part the 3-letter amino acid code and the third part the residue number of the amino acid. The entries in the interaction map are color coded, ranging from dark blue to dark red (as shown in the legend of Fig. B.). The absolute values of the cells in the interaction map are written as numbers for all feature types other than hydrophobic features if the number of interaction is below 400. (C) shows the statistical frequency of the features as a function of time. Thick enclosing lines indicate pharmacophore features that were present in the pharmacophore model obtained with the PDB structure, whereas thin, dashed lines indicate features that are not present in the PDB structure. The Y-axis corresponds to the number of occurrences of the specific feature per binned time-step and the X-axis corresponds to the binned time-steps. For a detailed description of this plot see the Methods Section, 'Frequency Plot generation'.

A closer look at Fig. 4.7 B reveals why most hydrophobic features have such stable and high frequencies – they interact with multiple interaction partners at the same time, thus preserving the interaction even in the case that one interaction partner leaves the range of influence of the ligand. It should be noted that the presented hydrogen bond features also have multiple interaction partners, but, indicated by the numbers for the different interaction partners for the hydrogen bond features, this feature type changes rarely between them, and if so, the change is slow and infrequent.

Fig. 4.7 C shows additional time resolved information about the frequencies of the pharmacophore features. As can be seen, the three hydrophobic features occur steadily above 95% for all binned time-steps – with the exception of H1 between the binned time-step 5 and 8. Around the same binned-time steps the frequency of HBA2 and HBD3 (both PDB features) drops and HBD1 (MD derived feature) appears with a subsequent frequency of around 70%. The high frequency of HBD1 is only partly represented in the total frequency (as seen in Fig. 4.7 A), since the feature was not present for the first quarter of the simulation. Our analysis provides an explanation for what happens at these binned time-steps. The RMSD value of the ligand starts to rise (shown in Fig. 4.6) and the movement of the ligand results in a change of the presented interaction partner, therefore we observe the drop in the frequencies for HBD3 and HBA2. The data presented in Fig. 4.7, especially in Fig. 4.7C, suggest, that the pharmacophore model which is appropriate for the first quarter of the simulation (based on the frequencies of the features) does not represent the second half of the simulation. Considering the different frequencies of the pharmacophore features two pharmacophore models are proposed: The first model contains the three hydrophobic features and HBD3 and HBA2 – this is the pharmacophore model derived from the PDB structure (and will be called subsequently PDB pharmacophore model). The second pharmacophore model contains the three hydrophobic features but HBA2 and HBD3 are exchanged in favor of HBD1 and HBD2 (and this model will be called MD derived pharmacophore model). These two pharmacophore models represent the pharmacophore features with different frequencies in the beginning and at the end of the MD simulation. Especially in the light of the work in [166], the reported findings are interesting. Although in the presented study a different tautomer was used than in [166], the typical hydrogen pattern with MET103 and GLU101 is present. But it appears as if the interaction

with LEU26 and ASP107 (as shown in Fig. 4.7 B) can also play an important role. In the following section the virtual screening results with these two pharmacophore models against a library of known actives and calculated decoys will be shown and discussed in detail.

The screening results (the receiver operator curve, enrichment factor and number of total hits) for the different pharmacophore models are shown in Fig. 4.8 Both pharmacophore models are able to discriminate between actives and decoys, and thus both provide good early enrichment.

The PDB pharmacophore model gives rise to 81 hits and the enrichment factor at 1.5% is 24.7. The pharmacophore hypothesis is able to retrieve 45 of the 226 active compounds.

The MD derived pharmacophore model leads to 530 hits, the early enrichment factor at 1.5% is 6.2. The pharmacophore model retrieves 66 of the 226 active compounds in the library.

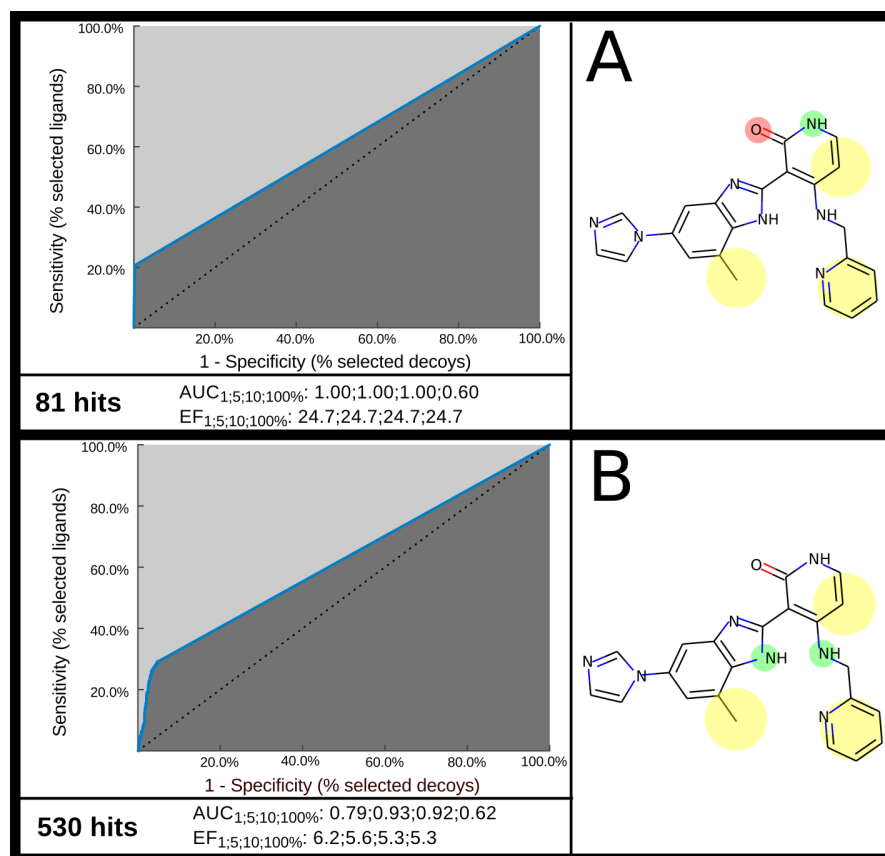


Figure 4.8. The receiver operating characteristic (ROC) curves and the two pharmacophore models are shown. In the ROC curve the true positive rate on the Y axis is plotted against the false positive rate on the X axis. The number of total hits and the enrichment factor (EF) are shown at 1%, 5%, 10% and 100%, respectively. In (A) the pharmacophore model obtained with the PDB structure and the virtual screening results are shown. In (B) the pharmacophore models with two MD derived hydrogen bond donor features and the virtual screening results are shown. For the description of the graphical 2D representation of the pharmacophore features see legend of Fig. 4.7.

A closer look at the hit-list obtained with both pharmacophore models reveals that the PDB pharmacophore model retrieves 33 active molecules that are not present in the hit-list obtained with the MD derived pharmacophore model. The MD derived pharmacophore model retrieves 54 unique hits – the hit-list of both pharmacophore models share only 12 active molecules. This is not surprising since the pharmacophore models are different and represent two distinct interaction modes. These two pharmacophore models can be used together – the PDB pharmacophore model is more likely able to distinguish between active and decoy models, but the MD derived pharmacophore model correctly identifies a higher number of active molecules. Since both models share only 12 active molecules in the resulting hit list, combining the results of these models results in a higher number of active candidates than only using the PDB pharmacophore.

4.3.4 Conclusions

In conclusion, MD simulations can reveal otherwise hidden pharmacophore features that are not present in the pharmacophore model derived from the experimental crystal structure. Using additional information obtained from MD simulation, i.e. time resolved frequency information and interaction plots, it is possible to construct pharmacophore models that integrate the dynamic of the ligand inside of the binding pocket. Furthermore, this approach provides an objective way to add MD derived pharmacophore features to PDB derived pharmacophore models or, on the other side, remove PDB features that are less important based on the observed frequencies.

4.4 A dynamic–shared Pharmacophore approach to improve early enrichment in virtual screening.

A case study on PPAR alpha

4.4.1 Introduction

Analysing the interaction pattern of ligands with a specific protein target has a strong bias towards the molecular structure of the ligand. Proteins can interact with active ligands of diverse shape, size and composition – which is not surprising since binding is a dynamic equilibrium process and the conformation of the binding site can be strongly influenced by the shape of the molecular binder [49]. But these findings raise doubts about the usefulness of the interaction pattern of one particular active binder as sole starting point for further drug discovery approaches. This issue can be avoided if the interaction patterns of multiple ligands are regarded for model development and refinement. In two recent papers [142, 143] we have shown how the information gained in the course of MD simulation can be combined with pharmacophore modelling. However, in this chapter we followed a different approach: to develop a workflow that addresses the arising issues of molecular docking and pharmacophore modelling when using (1) a single set of coordinates and (2) a single active ligand.

The starting point of this new approach are the crystal structures of three different ligands co-crystallised with the same protein (PDB CODE: 2P54, 4CI4, 3VI8) [93]. MD simulations are carried out for each of the structures, ligand-target interactions are analysed and finally modelled as pharmacophore features. A pharmacophore model is constructed using only the common pharmacophoric feature patterns that all three ligands exhibit during MD simulations. This ‘Molecular dYnamics SHARED PharmacophorE’ (MYSHAPE) is subsequently used for virtual screening using active and inactive molecules. The virtual screening performance of molecular docking is improved by adding constraints to the docking grid. These constraints reflect the pharmacophore feature interaction pattern obtained from the analysis of the MD simulation data. A consensus score based on the docking score and the pharmacophore alignment score is adopted at the end to maximize the virtual screening performance. In order to validate the approach, the virtual screening results of the different pharmacophore models and molecular docking runs are analysed. In other words, the screening results of the MYSHAPE model and the pharmacophore models obtained using the crystal structures are compared as well as the docking results using the crystal coordinate set with or without constraints. For the validation of the screening results Receiver-Operator Characteristics (ROC) graphs were

generated and then the Area Under the ROC Curve (ROC-AUC) is calculated. In particular, the focus is on the early enrichment of the resulting hit-list. The screening database contained molecules obtained from the DUD-E database for this target and was further enriched with ligands from the ChEMBL [168] and ZINC [169] databases.

4.4.2 Materials and Methods

System quality assessment and protein preparation

The selection of the investigated systems was based on the following criteria: high resolution of the crystal structure (below 2.5 Å), no metal ions in the binding pocket and only one bound ligand in the PDB structure. The electron density (ED) was evaluated using VHELIBS [33]. The protein preparation wizard [170] provided in MAESTRO 10.2 was used to add bond orders and hydrogens to the crystal structure. Subsequently Prime 4.0 [171, 172] was used to fix missing residues or atoms in the protein and to remove co-crystallised water. The protonation state of the protein and the ligand were evaluated using PropKa 3.1 [156, 157].

Molecular Dynamics

For every protein-ligand complex three 20 ns molecular dynamics simulations were performed. For each of the three simulations the same coordinate sets but different initial velocities were used. The MD simulations were performed with DESMOND 4.2 using the OPLS3 force field [173–175]. The complexes were solvated in orthorhombic boxes using the TIP3P water model. Ions were added to neutralize charges. The systems were minimised and equilibrated at a temperature of 303.15 K and at 1.013 bar pressure. The system was simulated as NPT ensemble, using a Nose-Hover thermostat and a Martyna-Tobia-Klein barostat. The integration time step was chosen to be 2 fs. In order to keep the hydrogen - heavy atom bonds rigid, the SHAKE algorithm was utilized. A 9 Å cut-off radius was set for the short range Coulomb interactions and smooth particle mesh Ewald was used for the long range interactions. The stability of systems was evaluated using the root mean square deviation (RMSD) of the aligned protein and ligand coordinate set calculated from the initial frame.

Shared features evaluation

The presence of pharmacophore features and their frequency during the MD simulation was investigated using the Interaction diagram tool provided by Maestro 10.2. A pharmacophore model based on the MD derived common features was then created considering only the common interactions that were found for the ligands in all MD simulations. In the following, the term MYSHAPE will be used to distinguish this type of model from the ‘default’ pharmacophore models generated using the crystal structure of the ligand-protein complex.

MYSHAPE model pharmacophore generation

In order to construct the MYSHAPE, the following workflow was applied: the different PDB protein-ligand structures were imported into LigandScout [20, 176] and for each complex a structure-based pharmacophore model was generated. Subsequently a shared pharmacophore model was generated using the structure-based pharmacophore models. Pharmacophore features, occurring during the MD simulation but not present in the original shared pharmacophore model, were added to the shared model. For the newly added pharmacophore features the tolerance radius was increased by 0.15 Å in order to compensate for small deviations in the 3D coordinates.

Docking Grid generation

Using the pharmacophore interaction pattern obtained from the MD simulations constraints were set on the docking grid. Specifically, positional constraints were imposed considering aromatic interactions, hydrogen bonds and hydrophobic interactions with the ligand according to the GLIDE grid constraints panel workflow. For each analysed protein-ligand system a docking grid with and without constraints was generated.

Ligand selection and preparation

In order to validate the virtual screening performance of the pharmacophore models and the docking grids, databases with known active and inactive compounds were generated. Using compounds with known activity makes it possible to test the capability of pharmacophore models and docking approaches to differentiate between active compounds and inactive molecules - which is the ultimate goal of both methods. To generate the screening library, active and decoy compounds were retrieved from the DUD-E database [151] and filtered to remove duplicates. The final data set contained 373 active and 5810 decoy molecules. The active and decoy

molecules were then optimised utilizing the Ligprep plugin provided by the MAESTRO software. The OPLS3 [175] force field was chosen and the protonation state of ligand set in accordance to a pH value equal to 7.

Pharmacophore screening

LigandScout [176] was used to perform the virtual screening analysis for the generated pharmacophore models. Standard settings were used as described in the user manual [176]. Receiver Operating Characteristics (ROC) graphs were generated and the Area Under the Curve (AUC) as well as the enrichment factor (EF) was calculated to validate the virtual screening performance of the pharmacophore models. For the 3VI8 PDB system we interpolated the four pyrene hydrophobic features into two and increase the tolerance of 0.30 Å, to avoid screening results with 0 hits. This resulted in a more sensitive model that also presented the same number of features than the others, allowing an easier comparison.

Molecular Docking Screening

Standard Precision (SP) and Extra Precision (XP) molecular docking with and without constraints was performed using GLIDE [94–96]. Ligands were considered flexible and EpiK state penalties were added to the docking score. ROC graphs and Robust Initial Enrichment (RIE) [177–179] were used to evaluate the virtual screening capability of the docking runs. In contrast to pharmacophore modelling the AUC was not calculated for different fractions of the screening database but a numeric representation of the Receiver Operator Characteristic area underneath the curve was obtained. This ROC value can be interpreted as the probability that an active will appear before an inactive compound and is calculated as follows (eq. 4.2):

$$ROC = \left(\frac{AUAC}{R_i} - \frac{R_a}{2R_i} \right) \quad eq.4.2$$

*AUAC is the area under the accumulation curve,
R_i is the ratio of inactive molecules s to the total number of compounds
in the screening library
R_a is the ratio of active compounds to the total number of entries in the
screening library.*

The RIE and the ROC value were generated using the “enrichment calculator” python script provided by Schrodinger. The EF was not calculated because the different docking runs produce ranked lists with different lengths and the EF is affected by the length of the datasets.

Pharmacophore alignment and ranking of docked conformations of ligands

For the protein-ligand system that performed best in molecular docking the resulting molecule list was imported into LigandScout [20, 176] and pharmacophore scores were calculated for each compound. This was done in order to compare the ranking of molecular docking with the one of pharmacophore modelling. The different rankings were evaluated using the EF at different percentages of the screening dataset with particular attention to the early enrichment [177, 178].

Post-processing Consensus Score

Furthermore, for the best performing system a consensus score was calculated that combined the pharmacophore and docking score as shown in Equation (4.3).

$$\text{Consensus Score} = \left(\frac{\text{docking score}}{\text{max docking score}} + \frac{\text{pharmacophore score}}{\text{max pharmacophore score}} \right) \quad \text{eq.4.3}$$

Evaluation of the chemotype enrichment

For checking the similarity of the compounds in the hit-lists of the pharmacophore models a KNIME workflow [180] was created that used the Morgan fingerprinting functionality of RDKit [86]. The chemotype similarity was evaluated using a 2D fingerprints (Morgan/Circular) [181]. Tanimoto distances of the fingerprints were calculated and K-medoids clustering was performed [182] (setting K =5) in order to analyze the distribution of chemotypes in the actives library. For the K-medoids search and clustering, no constraints were applied to the iterations. This means that the search for medoids and then the population of the clusters was stopped only when the best result was reached or no better solution was available. The protocol was applied on the whole active set and on the first 100 actives retrieved by the screening runs, to check if the model was capable to discriminate only one type of molecules or if the heterogeneity of the actives molecules was maintained in the early recognised ones.

4.4.3 Results and Discussion

Molecular dynamics

For the MD simulations the RMSD of the protein and ligand was calculated as described in the Methods Section. The protein and ligand showed normal RMSD

values over the course of the 20 ns of simulation and no large scale movement could be observed.

Shared features evaluation

As described in the Methods section, the interactions between the ligand and protein were investigated and a common interaction pattern was compiled. Figure 4.9 shows the protein-ligand interactions of the three different ligands. In such figure, the interactions that were used to construct the MYSHAPE model are highlighted. (shown in Figure 4.9).

One of the most interesting interaction event is the aromatic feature that only appears during the MD simulation. Remarkably, no pharmacophore model constructed from the crystal structure presents such pharmacophore feature. Moreover, without MD simulations this feature would have remained hidden. The common interaction pattern derived from the MD simulation of the different protein-ligand complexes consists of two hydrogen-bond acceptors, three hydrophobic interactions and one aromatic interaction (as shown in Figure 1). Four of these pharmacophore features are in all ligands associated with the same chemical group - the two hydrogen-bond acceptors are always located at a carboxyl group and the aromatic and one hydrophobic feature are always located at a central phenyl ring. The remaining two hydrophobic features are located at similar chemical groups but not at the same in all three ligands.

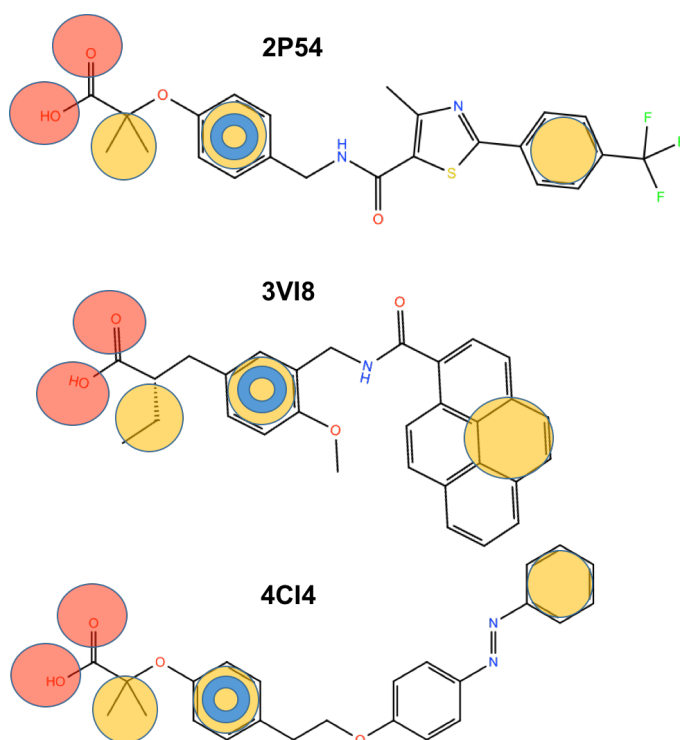


Fig.4.9 The three different ligands of 2P54, 3VI8 and 4CI4 are shown. The common features retrieved from the MD simulations are depicted on the ligands as coloured spheres. Red spheres indicate hydrogen-bond acceptor features, yellow spheres represent hydrophobic interactions and the blue ring represents an aromatic interaction. This common interaction pattern was used to generate a shared interaction.

Pharmacophore screening and Pharmacophore alignment score

The virtual screening performance of the four different pharmacophore models was investigated using a screening library composed of molecules with known activity as described above. Figure4.10 shows the three pharmacophore models that were created using the crystal structure, the generated MD derived common feature pharmacophore model and their virtual screening performance.

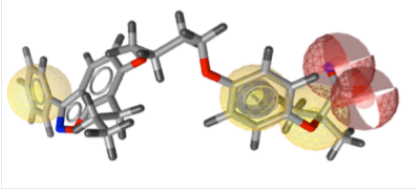
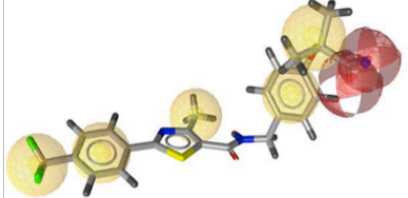
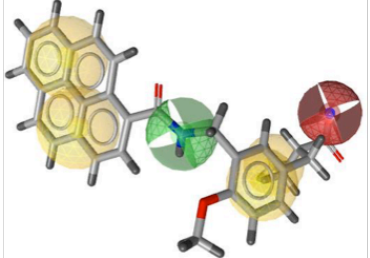
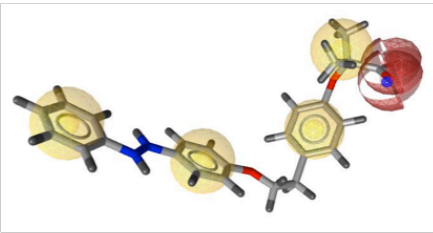
<p style="text-align: center;">MYSHAPE MODEL</p> 	<p style="text-align: center;">Nr. of Hits retrieved = 61</p> <p>AUC 1% = 1 EF 1% = 9.1 AUC 5% = 0.99 EF 5% = 9.1 AUC 10% = 0.93 EF 10% = 9.1 AUC 100% = 0.54 EF 100% = 9.1</p>
<p style="text-align: center;">2P54 MODEL</p> 	<p style="text-align: center;">Nr. of Hits retrieved = 5</p> <p>AUC 1% = 0.96 EF 1% = 11.6 AUC 5% = 0.59 EF 5% = 11.6 AUC 10% = 0.55 EF 10% = 11.6 AUC 100% = 0.50 EF 100% = 11.6</p>
<p style="text-align: center;">3VI8 MODEL</p> 	<p style="text-align: center;">Nr. of Hits retrieved = 26</p> <p>AUC 1% = 0.95 EF 1% = 3.1 AUC 5% = 0.60 EF 5% = 3.1 AUC 10% = 0.55 EF 10% = 3.1 AUC 100% = 0.50 EF 100% = 3.1</p>
<p style="text-align: center;">4CI4 MODEL</p> 	<p style="text-align: center;">Nr. of Hits retrieved = 14</p> <p>AUC 1% = 1 EF 1% = 9.9 AUC 5% = 0.72 EF 5% = 9.9 AUC 10% = 0.61 EF 10% = 9.9 AUC 100% = 0.51 EF 100% = 9.9</p>

Fig.4.10 Pharmacophore models with ligands and related ROC curve derived metrics for the pharmacophore screening of the three PDB structures and the MD-MRC derived one. For the MD-MRC model, best ranked molecule has been used. AUC is refers to the entire screened dataset; the EF is calculated for the hit-list with retrieved molecules. Red spheres = hydrogen-bond acceptor, Green spheres = hydrogen-bond donor, Yellow spheres = hydrophobic feature, Blue ring = aromatic feature.

In Figure 4.11, the ROC graphs are reported for each system. Virtual screening results were evaluated using the ROC-AUC values for 1%, 5%, 10% and 100% of the screening database and calculating the enrichment factors for 1%, 5%, 10% and 100%. A closer look at the ROC-AUC values for the four models (MYSHAPE model, 2P54 pharmacophore model, 3VI8 pharmacophore model and 4CI4 pharmacophore model) shows that all models perform well for 1% and 5%. The three

pharmacophore models retrieve hit lists with a relatively low number of hits (5, 14 and 26) compared to the results of the MYSHAPE model (61 hits).

The number of hits heavily influences the EF calculation. For this reason, the ROC-AUC value is a better metric to judge the performance of virtual screening runs that return hit-lists with different number of compounds.

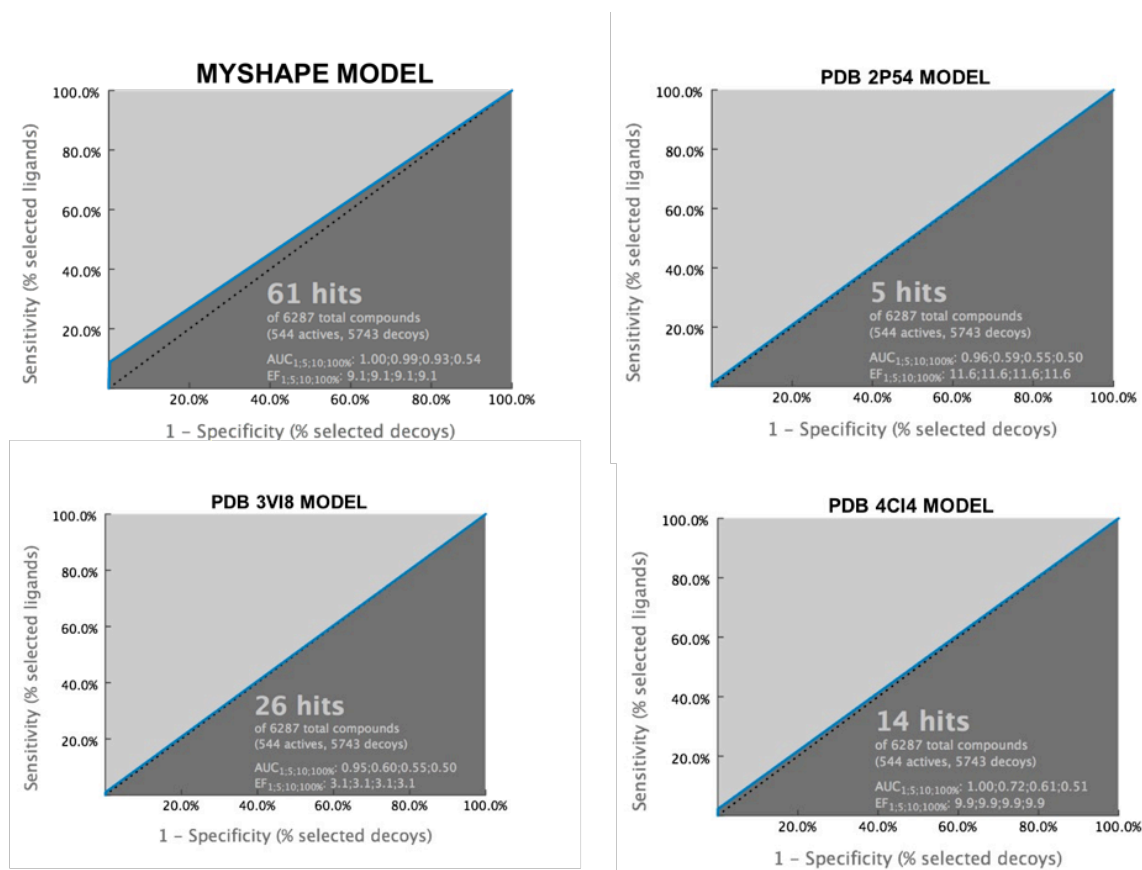


Fig.4.11 The ROC graphs for each model –pharmacophore screening

The calculated ROC-AUC values from the hit-list of the MYSHAPE model are 1 (1%), 0.99 (5%), 0.93 (10%) and 0.54 (100%). Comparing these values with the screening performance of the default pharmacophore models obtained from the crystal structure show that the MD derived common feature pharmacophore model has a better enrichment with active compounds than all other models. 4CI4 is the only model that retrieves for 1% of the screening database the same number of active compounds, but for 5% the ROC-AUC value is lower than the ROC-AUC calculated from the hit-list obtained with the MYSHAPE model.

Docking Process

As described in the Methods section docking grids were constructed for the three different investigated protein-ligand systems - either with constraints derived from

the MD simulations or without constraints. As a first approach SP docking was used to analyze the change in virtual screening capability of using grids with constraints. As a metric ROC graph values and RIE were used. The results are shown in Figure 4.12.

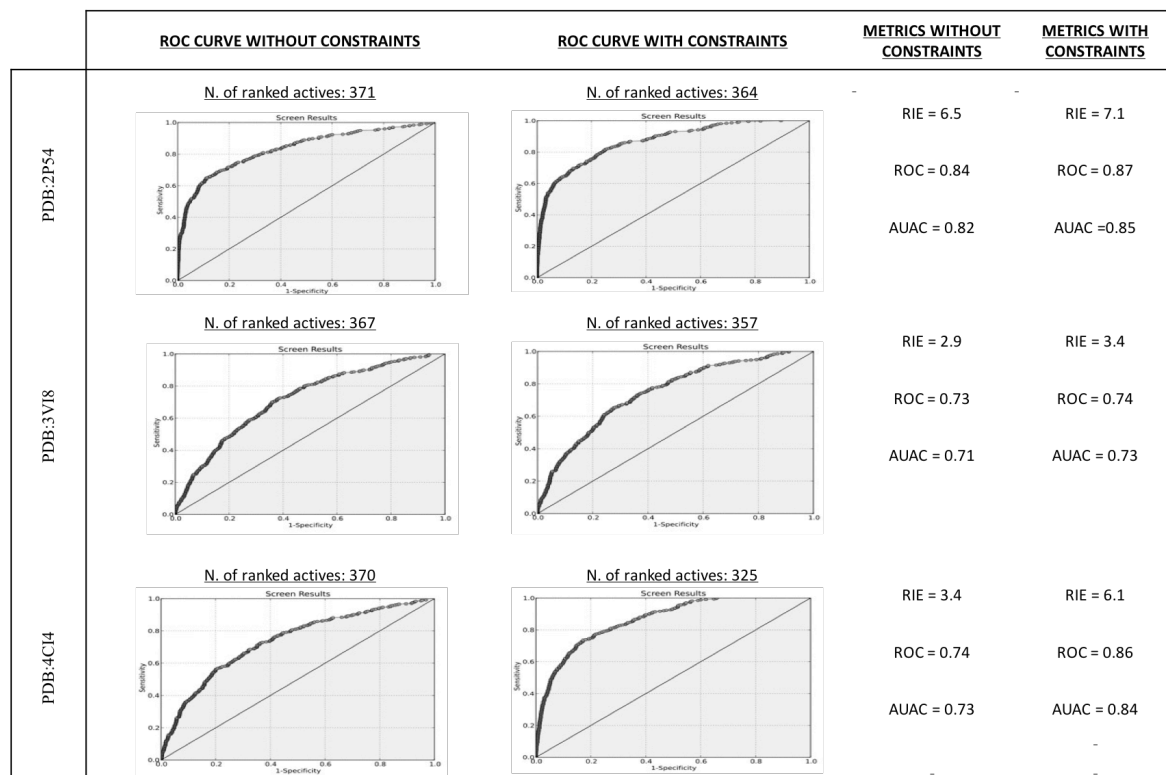


Fig.4.12 ROC curves and metrics for the SP docking on the three analysed system

The SP docking performs well for all investigated systems. As can be seen from the ROC graphs for all three systems the SP docking favours active molecules over inactive compounds. The ROC graph for 2P54 shows the best virtual screening performance of the systems. For all the studied systems one can conclude that the usage of constraints in the docking grid improves the virtual screening capability.

After the SP docking analysis we decided to proceed only with the 2P54 and submit it to the XP docking process trying to further improve the screening capability of our protocol.

In Figure 4.13 the results obtained from the XP docking runs with and without grid constraints are reported. As demonstrated by both XP curves and metrics, the application of grid constraints has not a high impact on the XP docking protocol. Moreover, comparing the ROC curve of the XP docking run with the one of the SP run with constraints, it seems that there is only a small increase in the capability of

the model. On the other hand, the application of constraints seems to reduce a little bit the sensitivity of the model, possibly because the combination of the XP algorithm with the constraints is too strict for the retrieval of molecules.

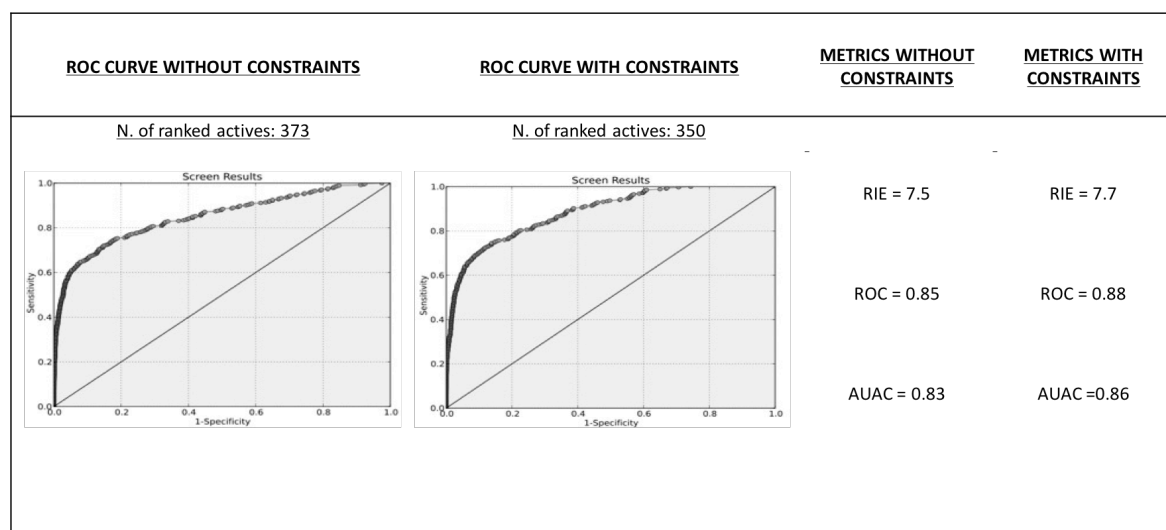


Fig.4.13 ROC curves and metrics for the 2P54 system after XP docking

Pharmacophore alignment and score of the docked conformations

We wanted then to test the capability of the MD derived common feature pharmacophore model to correctly rank actives and decoys. In the following tables we compare the EF of the hit-list molecules ranked by XP docking score (with and without constraints) with the EF of the hit-list ranked by the pharmacophore alignment score of the docked conformations.

Tab.4.3. EF for the two XP docking processes compared with the Pharmacophore score applied to docked conformations.

	EF of docking ranking without constraints	EF of pharmacophore alignment score on the docked conformation without constraints	EF of docking ranking with constraints	EF of pharmacophore alignment score on the docked conformation with constraints
EF1%	14.75	14.21	12.77	12.2
EF2%	13.52	9.7	11.21	9.65
EF5%	9.65	4.93	8	4.99
EF10%	6.36	3.03	5.9	3.38
EF20%	3.71	1.93	3.59	2
	MAX POSSIBLE EF =16.67		MAX POSSIBLE EF =13.33	

As the table 4.3 shows, it seems that the ranking of the molecules is better using the docking score than the pharmacophore score, showing a slight decrease of the screening capability in early enrichment.

However, it was interesting to see that the ranking produced by the two methods is quite different. The early enrichment produced by the two scoring models thus leads to different sets of molecules. For this reason, we decided to merge the two scoring functions into a consensus score.

Consensus Score

Starting from docking and pharmacophore alignment scores, we ranked the dataset with the consensus score described earlier in the Materials and Methods section (Eq. 2). We observed an interesting increase in early enrichment, with and without constraints. The consensus score was able to maintain the maximum EF at the 1% of the ranked list, and improved the general trend of the screening process until the 2% for the docked conformations without constraints and to the 5% for the ones docked with constraints. (table 4.4).

Tab.4.4. EF for the two XP docking processes compared with the consensus score applied to docked conformations.

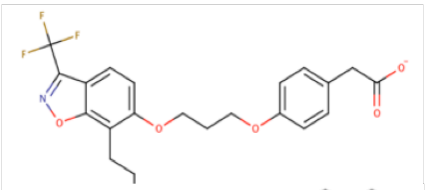
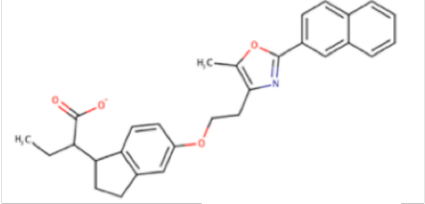
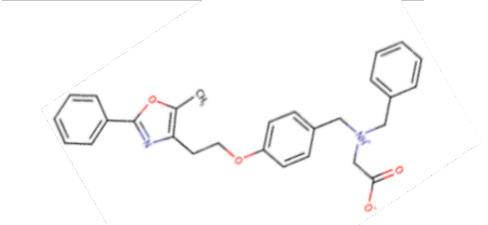
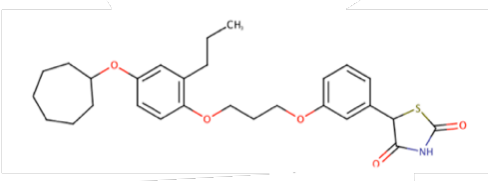
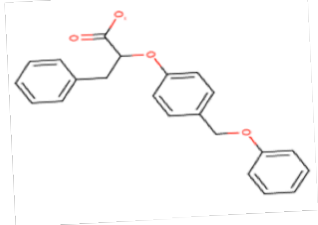
	EF of consensus score on the docked conformation without constraints	EF of consensus score on the docked conformation with constraints
EF _{1%}	16.67	13.33
EF _{2%}	14.62	12.2
EF _{5%}	8.83	8.69
EF _{10%}	5.35	5.65
EF _{20%}	3.25	3.45
	MAX POSSIBLE EF =16.67	MAX POSSIBLE EF =13.33

Evaluation of the chemotype enrichment

Finally, we checked the presence of possible biases in the whole protocol towards a unique chemotype of active molecules. As described in the Methods section, the calculation of distance matrix and k-medoids clustering based on Morgan fingerprints was used to check the chemotype distribution of active compounds. In Table 4.5 we report the medoid molecules of each cluster formed by the K-medoids algorithm, whereas in the table 4.6 are reported the distribution of active molecules for each cluster found. Values are expressed as percentage.

As shown by the results in table 4.5, the model adopted is able to rank all the different chemotypes found in the actives set.

Tab.4.5. Medoid molecules for each cluster formed

CLUSTER	MOLECULE ID	MOLECULE STRUCTURE
1	CHEMBL 210973	
2	CHEMBL 371041	
3	CHEMBL 371120	
4	CHEMBL 118753	
5	CHEMBL 572566	

Tab.4.6. Distribution of active molecules in the 5 cluster created

CLUSTER	DISTRIBUTION % OF ACTIVES DATASET	DISTRIBUTION % OF FIRST ACTIVES RETRIEVED
1	24	30
2	17	13
3	32	13
4	11	27
5	16	17

4.4.4 Conclusions

We have presented a new virtual screening workflow that addresses the arising issues of molecular docking and pharmacophore modelling when using a single set of coordinates and a single active ligand. The starting point of our study were three crystal structures of the PPAR α receptor containing different ligands co-crystallised with the same protein (PDB CODE: 2P54, 4CI4, 3VI8) [30]. For each structure, MD simulations were carried out and ligand-protein interactions were analysed and collected together with their appearance frequency. A pharmacophore model was then created using only the common feature patterns that all three ligands exhibited during MD simulations. This ‘Molecular dYnamics SHARed PharmacophorE’ (MYSHAPE) was then used for virtual screening on active and inactive molecules library. SHAPE was also used as constraints for the creation of the docking grid. This approach contributed to a rise in the molecular docking virtual screening performance. Finally, a consensus score based on the docking score and the pharmacophore alignment score was adopted to maximise the virtual screening performance. In order to validate the approach, we compared the virtual screening results of the different pharmacophore models and molecular docking runs. The aim of the work was the comparison between the screening performance of the shared feature pharmacophore and the pharmacophore models obtained using the crystal structures as well as the docking results using the crystal coordinate set with or without constraints.

The application of the MYSHAPE model showed an interesting increase of the screening capability both in terms of sensitivity of the model and specificity when compared to the three PDB models. The use of these interaction patterns to create the docking grid showed an improvement in the early recognition of actives compounds, especially for one of the three systems (4CI4) where the Robust Initial Enrichment (RIE) passed from 3.45 to 6.1. At first sight, the application of constraints at the XP docking protocol does not seem to strongly influence the docking protocol capability, probably because of the high precision of the XP docking algorithm itself that avoids high false positives rate. Nevertheless, when MYSHAPE pharmacophore model was used for the alignment of molecules in the docked conformation, the screening capability did not increase in both cases (with and without constraints). Anyway, adopting only the docking score or the pharmacophore score, the Enrichment Factor

(EF) of the protocol was good, but improvable especially for the early enrichment. When the two scoring methods were then combined in a consensus score there was an interesting boost of the virtual screening capability rising the value of the EF to be maximised in both docking methods. The early recognition was however improved until the 2% of the ranked list for the docked conformation without constraints whereas it was boosted until the 5% of the screening list for the docked conformation of molecules with constraints. The results obtained using the consensus score on the XP without constraints is to consider as the best compromise of speediness and accuracy in the virtual screening process.

This work is a first assay for a workflow that should be applied to different proteins. The strength behind the protocol is the ease of use related to the improvement of results. It also could represent a valid alternative to the use of very time consuming techniques such as XP docking with constraints. The increase of prediction reliability could be in fact reached through the use of pharmacophores, a fast and effective tool combined with no-constraints docking. This approach also represents a possible guide to consider or to discard some of the pharmacophore features retrieved from the static PDB crystal structures. Moreover, the MD simulations using more crystal structures of the same protein but with different ligands, is an interesting approach to retrieve some crucial interaction features that could be missed by the use of a single crystal structure. In the next months this approach will be applied to other receptors (ER, FXR, RXR, MapKinase and others) in order to test the application of the approach to other proteins.

5 COMPUTATIONAL CHEMISTRY IN POLYPHARMACOLOGY AND DRUG REPURPOSING

During my PhD, I had the opportunity to deeply study the possible applications of computational methods to drug repurposing and polypharmacology. From this applied study, two reviews have been published in the last year: Here I just report the abstract of the two published reviews [183, 184].

5.1 The repurposing of old drugs or unsuccessful lead compounds by in silico approaches: new advances and perspectives.

Abstract

Have you a compound in your lab, which was not successful against the designed target, or a drug that is no more attractive? The drug repurposing represents the right way to reconsider them. It can be defined as the modern and rationale approach of the traditional methods adopted in drug discovery, based on the knowledge, insight and luck, alias known as serendipity. This repurposing approach can be applied both in silico and in wet. In this review we report the molecular modeling facilities that can be of huge support in the repurposing of drugs and/or unsuccessful lead compounds. In the last decades, different methods were proposed to help the scientists in drug design and in drug repurposing. The steps strongly depend on the approach applied. It could be a ligand or a structure based method, correlated to the use of specific means. These processes, starting from a compound with potential therapeutic properties and a sizeable number of toxicity passed tests, can successfully speed up the very slow development of a molecule from bench to market. Herein, we discuss the facilities available to date, classifying them by methods and types. We have reported a series of databases, ligand and structure stand-alone software, and of web-based tools, which are free accessible to scientific community. This review does not claim to be exhaustive, but can be of interest to help in drug repurposing through in silico methods, as a valuable tool for the medicinal chemistry community.

5.2 Drugs polypharmacology by in silico methods: new opportunities in drug discovery.

Abstract

Polypharmacology, defined as the modulation of multiple proteins rather than a single target to achieve a desired therapeutic effect, has been gaining increasing attention since 1990s, when industries had to withdraw several drugs due to their adverse effects, leading to permanent injuries or death, with multi-billiondollar legal damages. Therefore, if up to then the "one drug one target" paradigm had seen many researchers interest focused on the identification of selective drugs, with the strong expectation to avoid adverse drug reactions (ADRs), very recently new research strategies resulted more appealing even as attempts to overcome the decline in productivity of the drug discovery industry.

Polypharmacology consists of two different approaches: the former, concerning a single drug interacting with multiple targets related to only one disease pathway; the latter, foresees a single drug's action on multiple targets involved in multiple disease pathways. Both new approaches are strictly connected to the discovery of new feasible off targets for approved drugs.

In this review, we describe how the in silico facilities can be a crucial support in the design of polypharmacological drug. The traditional computational protocols (ligand based and structure based) can be used in the search and optimization of drugs, by using specific filters to address them against the polypharmacology (fingerprints, similarity, etc.). Moreover, we dedicated a paragraph to biological and chemical databases, due to their crucial role in polypharmacology. Multitarget activities provide the basis for drug repurposing, a slightly different issue of high interest as well, which is mostly applied on a single target involved in more than one diseases. In this contest, computational methods have raised high interest due to the reached power of hardware and software in the manipulation of data.

BIBLIOGRAPHY

- [1] Mandal, S., Moudgil, M., Mandal, S.K., Rational drug design. *Eur. J. Pharmacol.* 2009, 625, 90–100.
- [2] Shirai, H., Prades, C., Vita, R., Marcatili, P., et al., Antibody informatics for drug discovery. *Biochim. Biophys. Acta - Proteins Proteomics* 2014, 1844, 2002–2015.
- [3] Waring, M.J., Arrowsmith, J., Leach, A.R., Leeson, P.D., et al., (11) An analysis of the attrition of drug candidates from four major pharmaceutical companies. *Nat Rev Drug Discov* 2015, 14, 475–486.
- [4] Yu, H., Adedoyin, A., ADME-Tox in drug discovery: Integration of experimental and computational technologies. *Drug Discov. Today* 2003, 8, 852–861.
- [5] Ganellin, R., Roberts, S., Jefferis, R., Stocks, M., The small molecule drug discovery process – from target selection to candidate selection, in: *Introduction to Biological and Small Molecule Drug Research and Development*, 2013, pp. 81–126.
- [6] Yuan, Y., Pei, J., Lai, L., Binding site detection and druggability prediction of protein targets for structure-based drug design. *Curr. Pharm. Des.* 2013, 19, 2326–33.
- [7] Rishton, G.M., Nonleadlikeness and leadlikeness in biochemical screening. *Drug Discov. Today* 2003, 8, 86–96.
- [8] Kuentz, M., Imanidis, G., In silico prediction of the solubility advantage for amorphous drugs - Are there property-based rules for drug discovery and early pharmaceutical development? *Eur. J. Pharm. Sci.* 2013, 48, 554–562.
- [9] Nicolaou, C.A., Brown, N., Multi-objective optimization methods in drug design. *Drug Discov. Today Technol.* 2013, 10.
- [10] Ban, T.A., The role of serendipity in drug discovery. *Dialogues Clin. Neurosci.* 2006, 8, 335–344.
- [11] Ou-Yang, S.-S., Lu, J.-Y., Kong, X.-Q., Liang, Z.-J., et al., Computational drug discovery. *Acta Pharmacol. Sin.* 2012, 33, 1131–40.
- [12] Chang, C. a, Ai, R., Gutierrez, M., Marsella, M.J., Computational Drug Discovery and Design. *Methods* 2012, 819, 3–12.
- [13] Ballew, B.S., Elsevier's Scopus® Database. *J. Electron. Resour. Med. Libr.* 2009, 6, 245–252.
- [14] Green, D.V.S., Virtual screening of virtual libraries. *Prog. Med. Chem.* 2003, 41, 61–97.
- [15] van de Waterbeemd, H., Gifford, E., ADMET in silico modelling: towards prediction paradise? *Nat. Rev. Drug Discov.* 2003, 2, 192–204.
- [16] Pharma 2020: Virtual R&D Which path will you take? n.d.
- [17] Acharya, C., Coop, A., Polli, J.E., Mackerell, A.D., Recent advances in ligand-based drug design: relevance and utility of the conformationally sampled pharmacophore approach. *Curr. Comput. Aided. Drug Des.* 2011, 7, 10–22.
- [18] Kalyaanamoorthy, S., Chen, Y.P.P., Structure-based drug design to augment hit discovery. *Drug Discov. Today* 2011, 16, 831–839.
- [19] Wolber, G., Sippl, W., Pharmacophore Identification and Pseudo-Receptor Modeling, in: *The Practice of Medicinal Chemistry: Fourth Edition*, 2015, pp. 489–510.
- [20] Wolber, G., Seidel, T., Bendix, F., Langer, T., Molecule-pharmacophore superpositioning and pattern matching in computational drug design. *Drug Discov. Today* 2008, 13, 23–29.
- [21] Langer, T., Wolber, G., Pharmacophore definition and 3D searches. *Drug Discov. Today Technol.* 2004, 1, 203–207.
- [22] Wermuth, C.G., Ganellin, C.R., Lindberg, P., Mitscher, L.A., Glossary of terms used in

- medicinal chemistry (IUPAC Recommendations 1998). *Pure Appl. Chem.* 1998, 70, 1129–1143.
- [23] Agrawal, R., Jain, P., Dikshit, S.N., Bahare, R.S., et al., Ligand-based pharmacophore detection, screening of potential pharmacophore and docking studies, to get effective glycogen synthase kinase inhibitors. *Med. Chem. Res.* 2013, 22, 5504–5535.
- [24] Lauria, A., Tutone, M., Almerico, A.M., Virtual lock-and-key approach: The in silico revival of Fischer model by means of molecular descriptors. *Eur. J. Med. Chem.* 2011, 46, 4274–4280.
- [25] Lauria, A., Tutone, M., Barone, G., Almerico, A.M., Multivariate analysis in the identification of biological targets for designed molecular structures: The BIOTA protocol. *Eur. J. Med. Chem.* 2014, 75, 106–110.
- [26] Cramer, R.D., The inevitable QSAR renaissance. *J. Comput. Aided. Mol. Des.* 2012, 26, 35–38.
- [27] Tropsha, A., Best practices for QSAR model development, validation, and exploitation. *Mol. Inform.* 2010, 29, 476–488.
- [28] Kuninyi, H., QSAR and 3D QSAR in drug design. *Drug Discov. Today* 1997, 2, 457–467.
- [29] Martinek, T.A., Ötvös, F., Dervarics, M., Tóth, G., et al., Ligand-based prediction of active conformation by 3D-QSAR flexibility descriptors and their application in 3+3D-QSAR models. *J. Med. Chem.* 2005, 48, 3239–3250.
- [30] Verma, J., Khedkar, V.M., Coutinho, E.C., 3D-QSAR in Drug Design - A Review. *Curr. Top. Med. Chem.* 2010, 10, 95–115.
- [31] Jhoti, H., Leach, A.R., Structure-based drug discovery, 2007.
- [32] Schneider, M., Fu, X., Keating, A.E., X-ray vs. NMR structures as templates for computational protein design. *Proteins* 2009, 77, 97–110.
- [33] Krieger, E., Nabuurs, S.B., Vriend, G., Homology Modeling. *Homol. Model. Methods Protoc.* 2012, 857, 419.
- [34] Cavasotto, C.N., Phatak, S.S., Homology modeling in drug discovery: current trends and applications. *Drug Discov. Today* 2009, 14, 676–683.
- [35] Peng, J., Xu, J., Low-homology protein threading. *Bioinformatics* 2010, 26.
- [36] Torda, A.E., Protein Threading, in: *The Proteomics Protocols Handbook SE - 70*, 2005, pp. 921–938.
- [37] Langer, T., Pharmacophores for medicinal chemists: a personal view. <http://dx.doi.org/10.4155/fmc.11.34> 2011.
- [38] Gerhard Wolber*, † and, Langer‡, T., LigandScout: 3-D Pharmacophores Derived from Protein-Bound Ligands and Their Use as Virtual Screening Filters. 2004.
- [39] Dias, R., de Azevedo, W.F., Molecular docking algorithms. *Curr. Drug Targets* 2008, 9, 1040–1047.
- [40] Morris, G.M., Lim-Wilby, M., Molecular docking. *Methods Mol. Biol.* 2008, 443, 365–382.
- [41] Meng, X.-Y., Zhang, H.-X., Mezei, M., Cui, M., Molecular docking: a powerful approach for structure-based drug discovery. *Curr. Comput. Aided. Drug Des.* 2011, 7, 146–57.
- [42] Kearsley, S.K., Underwood, D.J., Sheridan, R.P., Miller, M.D., Flexibases: A way to enhance the use of molecular docking methods. *J. Comput. Aided. Mol. Des.* 1994, 8, 565–582.
- [43] Klebe, G., Mietzner, T., A fast and efficient method to generate biologically relevant conformations. *J. Comput. Aided. Mol. Des.* 1994, 8, 583–606.
- [44] Wang, Q., Pang, Y.P., Preference of small molecules for local minimum conformations when binding to proteins. *PLoS One* 2007, 2.
- [45] Klebe, G., Recent developments in structure-based drug design. *J. Mol. Med. (Berl)*. 2000, 78, 269–81.
- [46] Schneider, G., Fechner, U., Computer-based de novo design of drug-like molecules. *Nat. Rev. Drug Discov.* 2005, 4, 649–63.

- [47] Jorgensen, W.L., The many roles of computation in drug discovery. *Science* 2004, 303, 1813–8.
- [48] Fischer, E., Einfluss der Configuration auf die Wirkung der Enzyme. *Berichte Der Dtsch. Chem. Gesellschaft* 1894, 27, 2985–2993.
- [49] Ma, B., Shatsky, M., Wolfson, H.J., Nussinov, R., Multiple diverse ligands binding at a single protein site: a matter of pre-existing populations. *Protein Sci.* 2002, 11, 184–97.
- [50] Teague, S.J., Implications of protein flexibility for drug discovery. *Nat. Rev. Drug Discov.* 2003, 2, 527–541.
- [51] Cornell, W.D., Cieplak, P., Bayly, C.I., Gould, I.R., et al., A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* 1995, 117, 5179–5197.
- [52] Jones, J.E., On the Determination of Molecular Fields. I. From the Variation of the Viscosity of a Gas with Temperature. *Proc. R. Soc. A Math. Phys. Eng. Sci.* 1924, 106, 441–462.
- [53] Koshland, D.E., Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proc. Natl. Acad. Sci. U. S. A.* 1958, 44, 98–104.
- [54] Lexa, K.W., Carlson, H.A., Protein flexibility in docking and surface mapping. *Q. Rev. Biophys.* 2012, 45, 301–43.
- [55] Ferrari, A.M., Wei, B.Q., Costantino, L., Shoichet, B.K., Soft docking and multiple receptor conformations in virtual screening. *J. Med. Chem.* 2004, 47, 5076–5084.
- [56] Mizutani, M.Y., Takamatsu, Y., Ichinose, T., Nakamura, K., et al., Effective handling of induced-fit motion in flexible docking. *Proteins Struct. Funct. Genet.* 2006, 63, 878–891.
- [57] Pencheva, T., Lagorce, D., Pajeva, I., Villoutreix, B.O., et al., AMMOS: Automated Molecular Mechanics Optimization tool for in silico Screening. *BMC Bioinformatics* 2008, 9, 438.
- [58] Nabuurs, S.B., Wagener, M., De Vlieg, J., A flexible approach to induced fit docking. *J. Med. Chem.* 2007, 50, 6507–6518.
- [59] Dunbrack, R.L., Karplus, M., Backbone-dependent rotamer library for proteins. Application to side-chain prediction. *J. Mol. Biol.* 1993, 230, 543–574.
- [60] Choudhury, C., Priyakumar, U.D., Sastry, G.N., Dynamics based pharmacophore models for screening potential inhibitors of mycobacterial cyclopropane synthase. *J. Chem. Inf. Model.* 2015, 55, 848–860.
- [61] Sohn, Y.S., Park, C., Lee, Y., Kim, S., et al., Multi-conformation dynamic pharmacophore modeling of the peroxisome proliferator-activated receptor γ for the discovery of novel agonists. *J. Mol. Graph. Model.* 2013, 46, 1–9.
- [62] Spyraakis, F., Benedetti, P., Decherchi, S., Rocchia, W., et al., A Pipeline to Enhance Ligand Virtual Screening: Integrating Molecular Dynamics and Fingerprints for Ligand and Proteins. *J. Chem. Inf. Model.* 2015, 55, 2256–2274.
- [63] De Paris, R., Quevedo, C. V., Ruiz, D.D.A., De Souza, O.N., An effective approach for clustering InhA molecular dynamics trajectory using substrate-binding cavity features. *PLoS One* 2015, 10.
- [64] Sperandio, O., Mouawad, L., Pinto, E., Villoutreix, B.O., et al., How to choose relevant multiple receptor conformations for virtual screening: A test case of Cdk2 and normal mode analysis. *Eur. Biophys. J.* 2010, 39, 1365–1372.
- [65] Sinko, W., Lindert, S., Mccammon, J.A., Accounting for Receptor Flexibility and Enhanced Sampling Methods in Computer-Aided Drug Design. *Chem. Biol. Drug Des.* 2013, 81, 41–49.
- [66] Plattner, N., Noé, F., Protein conformational plasticity and complex ligand-binding kinetics explored by atomistic simulations and Markov models. *Nat. Commun.* 2015, 6, 7653.
- [67] Fischer, E., Influence of configuration on the action of enzymes. *Ber.* 1894, 27, 2985–2993.
- [68] Landsberg, P.T., Nobel Lectures in Physics, 1901-1921. *Phys. Bull.* 1967, 18, 151.
- [69] Favia, A.D., Thornton, J.M., Nobeli, I., Protein promiscuity and its implications for

- biotechnology. *Nat. Biotechnol.* 2009, 27, 157–67.
- [70] Peterson, R.T., Chemical biology and the limits of reductionism. *Nat. Chem. Biol.* 2008, 4, 635–638.
- [71] Oprea, T.I., Tropsha, A., Faulon, J., Rintoul, M.D., Systems chemical biology. *Nat. Chem. Biol.* 2007, 3, 447–50.
- [72] Bajorath, J., Computational analysis of ligand relationships within target families. *Curr. Opin. Chem. Biol.* 2008, 12, 352–358.
- [73] Wagner, B.K., Kitami, T., Gilbert, T.J., Peck, D., et al., Large-scale chemical dissection of mitochondrial function. *Nat. Biotechnol.* 2008, 26, 343–51.
- [74] Young, D.W., Bender, A., Hoyt, J., McWhinnie, E., et al., Integrating high-content screening and ligand-target prediction to identify mechanism of action. *Nat. Chem. Biol.* 2008, 4, 59–68.
- [75] Krejsa, C.M., Horvath, D., Rogalski, S.L., Penzotti, J.E., et al., Predicting ADME properties and side effects: the BioPrint approach. *Curr. Opin. Drug Discov. Devel.* 2003, 6, 470–480.
- [76] Todeschini, R., Consonni, V., Molecular Descriptors for Chemoinformatics, 2010.
- [77] Todeschini, R., Consonni, V., Handbook of Molecular Descriptors. *New York* 2000, 11, 688.
- [78] Hert, J., Willett, P., Wilton, D.J., Acklin, P., et al., Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures. *Org. Biomol. Chem.* 2004, 2, 3256.
- [79] Willett, P., Similarity-Based Virtual Screening Using 2D Fingerprints. *Drug Discov. Today* 2006, 11, 1046–1053.
- [80] Keiser, M.J., Setola, V., Irwin, J.J., Laggner, C., et al., Predicting new molecular targets for known drugs. *Nature* 2009, 462, 175–181.
- [81] Martin, Y.C., Kofron, J.L., Traphagen, L.M., Do structurally similar molecules have similar biological activity? *J. Med. Chem.* 2002, 45, 4350–4358.
- [82] Kubinyi, H., Chemical similarity and biological activities, in: *Journal of the Brazilian Chemical Society*, 2002, pp. 717–726.
- [83] Liu, T., Lin, Y., Wen, X., Jorissen, R.N., et al., BindingDB: A web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.* 2007, 35.
- [84] Karelson, M., Lobanov, V.S., Katritzky, A.R., Quantum-Chemical Descriptors in QSAR/QSPR Studies. *Chem. Rev.* 1996, 96, 1027–1044.
- [85] Tutone, M., Perricone, U., Almerico, A.M., Conf-VLKA: A structure-based revisit of the Virtual Lock-and-Key Approach. *J. Mol. Graph. Model.* 2016, 71, 50–57.
- [86] Rogers, D., Hahn, M., Extended-connectivity fingerprints. *J. Chem. Inf. Model.* 2010, 50, 742–754.
- [87] Bender, A., Mussa, H.Y., Glen, R.C., Reiling, S., Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): Evaluation of performance. *J. Chem. Inf. Comput. Sci.* 2004, 44, 1708–1718.
- [88] Christie, B.D., Henry, D.R., Wipke, W.T., Mook, T.E., Database structure and searching in MACCS-3D. *Tetrahedron Comput. Methodol.* 1990, 3, 653–664.
- [89] LEVANDOWSKY, M., WINTER, D., Distance between Sets. *Nature* 1971, 234, 34–35.
- [90] Sastry, M., Lowrie, J.F., Dixon, S.L., Sherman, W., Large-scale systematic analysis of 2D fingerprint methods and parameters to improve virtual screening enrichments. *J. Chem. Inf. Model.* 2010, 50, 771–784.
- [91] Duan, J., Dixon, S.L., Lowrie, J.F., Sherman, W., Analysis and comparison of 2D fingerprints: Insights into database screening performance using eight fingerprint methods. *J. Mol. Graph. Model.* 2010, 29, 157–170.
- [92] Kruger, F.A., Overington, J.P., Global analysis of small molecule binding to related protein targets. *PLoS Comput. Biol.* 2012, 8.
- [93] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., et al., The Protein Data Bank. *Nucleic Acids Res.* 2000, 28, 235–242.

- [94] Halgren, T.A., Murphy, R.B., Friesner, R.A., Beard, H.S., et al., Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening. *J. Med. Chem.* 2004, *47*, 1750–1759.
- [95] Friesner, R.A., Murphy, R.B., Repasky, M.P., Frye, L.L., et al., Extra precision glide: Docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *J. Med. Chem.* 2006, *49*, 6177–6196.
- [96] Friesner, R.A., Banks, J.L., Murphy, R.B., Halgren, T.A., et al., Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* 2004, *47*, 1739–1749.
- [97] Good, A.C., Hermsmeider, M.A., Measuring CAMD technique performance. 2. How “druglike” are drugs? Implications of random test set selection exemplified using druglikeness classification models. *J. Chem. Inf. Model.* 2007, *47*, 110–114.
- [98] Good, A.C., Hermsmeider, M.A., Hindle, S.A., Measuring CAMD technique performance: A virtual screening case study in the design of validation experiments. *J. Comput. Aided. Mol. Des.* 2004, *18*, 529–536.
- [99] Lauria, A., Abbate, I., Patella, C., Martorana, A., et al., New annelated thieno[2,3-e][1,2,3]triazolo[1,5-a]pyrimidines, with potent anticancer activity, designed through VLAK protocol. *Eur. J. Med. Chem.* 2013, *62*, 416–424.
- [100] Lauria, A., Patella, C., Abbate, I., Martorana, A., et al., Lead optimization through VLAK protocol: New annelated pyrrolo-pyrimidine derivatives as antitumor agents. *Eur. J. Med. Chem.* 2012, *55*, 375–383.
- [101] B-Rao, C., Subramanian, J., Sharma, S.D., Managing protein flexibility in docking and its applications. *Drug Discov. Today* 2009, *14*, 394–400.
- [102] Gallicchio, E., Levy, R.M., Advances in all atom sampling methods for modeling protein-ligand binding affinities. *Curr. Opin. Struct. Biol.* 2011, *21*, 161–166.
- [103] Chen, Y.C., Beware of docking! *Trends Pharmacol. Sci.* 2015, *36*, 78–95.
- [104] Cerqueira, N.M.F.S.A., Gesto, D., Oliveira, E.F., Santos-Martins, D., et al., Receptor-based virtual screening protocol for drug discovery. *Arch. Biochem. Biophys.* 2015, *582*, 56–67.
- [105] Shin, W.H., Kim, J.K., Kim, D.S., Seok, C., GalaxyDock2: Protein-ligand docking using beta-complex and global optimization. *J. Comput. Chem.* 2013, *34*, 2647–2656.
- [106] Koska, J., Spassov, V.Z., Maynard, A.J., Yan, L., et al., Fully automated molecular mechanics based induced fit protein-ligand docking method. *J. Chem. Inf. Model.* 2008, *48*, 1965–1973.
- [107] Sherman, W., Day, T., Jacobson, M.P., Friesner, R.A., et al., Novel procedure for modeling ligand/receptor induced fit effects. *J. Med. Chem.* 2006, *49*, 534–553.
- [108] Bolia, A., Gerek, Z.N., Ozkan, S.B., BP-dock: A flexible docking scheme for exploring protein-ligand interactions based on unbound structures. *J. Chem. Inf. Model.* 2014, *54*, 913–925.
- [109] Ivetac, A., McCammon, J.A., Molecular recognition in the case of flexible targets. *Curr. Pharm. Des.* 2011, *17*, 1663–1671.
- [110] Barril, X., Morley, S.D., Unveiling the full potential of flexible receptor docking using multiple crystallographic structures. *J. Med. Chem.* 2005, *48*, 4432–4443.
- [111] Bolstad, E.S.D., Anderson, A.C., In pursuit of virtual lead optimization: Pruning ensembles of receptor structures for increased efficiency and accuracy during docking. *Proteins Struct. Funct. Bioinforma.* 2009, *75*, 62–74.
- [112] Forman-Kay, J.D., The “dynamics” in the thermodynamics of binding. *Nat. Struct. Biol.* 1999, *6*, 1086–1087.
- [113] Verkhivker, G.M., Bouzida, D., Gehlhaar, D.K., Rejto, P.A., et al., Complexity and simplicity of ligand-macromolecule interactions: The energy landscape perspective. *Curr. Opin. Struct. Biol.* 2002, *12*, 197–203.
- [114] Nichols, S.E., Baron, R., McCammon, J.A., On the use of molecular dynamics receptor

- conformations for virtual screening. *Methods Mol. Biol.* 2012, 819, 93–103.
- [115] Totrov, M., Abagyan, R., Flexible ligand docking to multiple receptor conformations: a practical alternative. *Curr. Opin. Struct. Biol.* 2008, 18, 178–184.
- [116] Abagyan, R., Rueda, M., Bottegoni, G., Recipes for the selection of experimental protein conformations for virtual screening. *J. Chem. Inf. Model.* 2010, 50, 186–193.
- [117] Isvoran, A., Badel, A., Craescu, C.T., Miron, S., et al., Exploring NMR ensembles of calcium binding proteins: perspectives to design inhibitors of protein-protein interactions. *BMC Struct. Biol.* 2011, 11, 24.
- [118] Miteva, M.A., Robert, C.H., Maréchal, J.D., Perahia, D., Miteva_11 Receptor Flexibility in Ligand Docking and Virtual Screening. *Silico Lead Discov.* 2011, 99–117.
- [119] Osguthorpe, D.J., Sherman, W., Hagler, A.T., Generation of Receptor Structural Ensembles for Virtual Screening Using Binding Site Shape Analysis and Clustering. *Chem. Biol. Drug Des.* 2012, 80, 182–193.
- [120] Asses, Y., Venkatraman, V., Leroux, V., Ritchie, D.W., et al., Exploring c-Met kinase flexibility by sampling and clustering its conformational space. *Proteins Struct. Funct. Bioinforma.* 2012, 80, 1227–1238.
- [121] Degliesposti, G., Portioli, C., Parenti, M.D., Rastelli, G., BEAR, a novel virtual screening methodology for drug discovery. *J. Biomol. Screen. Off. J. Soc. Biomol. Screen.* 2011, 16, 129–133.
- [122] Hou, T., Wang, J., Li, Y., Wang, W., et al., Assessing the performance of the MM/PBSA and MM/GBSA methods: I. The accuracy of binding free energy calculations based on molecular dynamics simulations. *J. Chem. Inf. Comput. Sci.* 2011, 51, 69–82.
- [123] Proctor, E.A., Yin, S., Tropsha, A., Dokholyan, N. V., Discrete molecular dynamics distinguishes natively-like binding poses from decoys in difficult targets. *Biophys. J.* 2012, 102, 144–151.
- [124] Amaro, R.E., Baron, R., McCammon, J.A., An improved relaxed complex scheme for receptor flexibility in computer-aided drug design. *J. Comput. Aided. Mol. Des.* 2008, 22, 693–705.
- [125] Xu, M., Lill, M.A., Utilizing experimental data for reducing ensemble size in flexible-protein docking. *J. Chem. Inf. Model.* 2012, 52, 187–198.
- [126] Martiny, V.Y., Carbonell, P., Lagorce, D., Villoutreix, B.O., et al., In Silico Mechanistic Profiling to Probe Small Molecule Binding to Sulfotransferases. *PLoS One* 2013, 8.
- [127] Rueda, M., Bottegoni, G., Abagyan, R., Consistent improvement of cross-docking results using binding site ensembles generated with elastic network normal modes. *J. Chem. Inf. Model.* 2009, 49, 716–725.
- [128] Leis, S., Zacharias, M., Efficient inclusion of receptor flexibility in grid-based protein-ligand docking. *J. Comput. Chem.* 2011, 32, 3433–3439.
- [129] Korb, O., Olsson, T.S.G., Bowden, S.J., Hall, R.J., et al., Potential and limitations of ensemble docking. *J. Chem. Inf. Model.* 2012, 52, 1262–1274.
- [130] Sgobba, M., Caporuscio, F., Anighoro, A., Portioli, C., et al., Application of a post-docking procedure based on MM-PBSA and MM-GBSA on single and multiple protein conformations. *Eur. J. Med. Chem.* 2012, 58, 431–440.
- [131] Liebeschuetz, J., Hennemann, J., Olsson, T., Groom, C.R., The good, the bad and the twisted: A survey of ligand geometry in protein crystal structures. *J. Comput. Aided. Mol. Des.* 2012, 26, 169–183.
- [132] Reynolds, C.H., Protein-ligand cocrystal structures: We can do better. *ACS Med. Chem. Lett.* 2014, 5, 727–729.
- [133] Mirjalili, V., Noyes, K., Feig, M., Physics-based protein structure refinement through multiple molecular dynamics trajectories and structure averaging. *Proteins Struct. Funct. Bioinforma.* 2014, 82, 196–207.
- [134] Whitesides, G.M., Krishnamurthy, V.M., Designing ligands to bind proteins. *Q. Rev.*

- Biophys.* 2005, 38, 385–95.
- [135] Yang, S.-Y., Pharmacophore modeling and applications in drug discovery: challenges and recent advances. *Drug Discov. Today* 2010, 15, 444–450.
- [136] Wu, F., Xu, T., He, G., Ouyang, L., et al., Discovery of novel focal adhesion kinase inhibitors using a hybrid protocol of virtual screening approach based on multicomplex-based pharmacophore and molecular docking. *Int. J. Mol. Sci.* 2012, 13, 15668–15678.
- [137] Borhani, D.W., Shaw, D.E., The future of molecular dynamics simulations in drug discovery. *J. Comput. Aided. Mol. Des.* 2012, 26, 15–26.
- [138] De Vivo, M., Masetti, M., Bottegoni, G., Cavalli, A., Role of Molecular Dynamics and Related Methods in Drug Discovery. *J. Med. Chem.* 2016, 59, 4035–4061.
- [139] Karplus, M., McCammon, J.A., Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.* 2002, 9, 646–652.
- [140] Deng, J., Lee, K.W., Sanchez, T., Cui, M., et al., Dynamic receptor-based pharmacophore model development and its application in designing novel HIV-1 integrase inhibitors. *J. Med. Chem.* 2005, 48, 1496–1505.
- [141] Thangapandian, S., John, S., Lee, Y., Kim, S., et al., Dynamic structure-based pharmacophore model development: A new and effective addition in the Histone deacetylase 8 (HDAC8) inhibitor discovery. *Int. J. Mol. Sci.* 2011, 12, 9440–9462.
- [142] Wieder, M., Perricone, U., Seidel, T., Boresch, S., et al., Comparing pharmacophore models derived from crystal structures and from molecular dynamics simulations. *Monatshefte Fur Chemie* 2016, 147, 553–563.
- [143] Wieder, M., Perricone, U., Boresch, S., Seidel, T., et al., Evaluating the stability of pharmacophore features using molecular dynamics simulations. *Biochem. Biophys. Res. Commun.* 2016, 470, 685–689.
- [144] Gaillard, T., Panel, N., Simonson, T., Protein side chain conformation predictions with an MMGBSA energy function. *Proteins Struct. Funct. Bioinforma.* 2016, 84, 803–819.
- [145] Perricone, U., Wieder, M., Seidel, T., Langer, T., et al., A MOLECULAR DYNAMICS – SHARED PHARMACOPHORE APPROACH TO BOOST EARLY ENRICHMENT VIRTUAL SCREENING. A CASE STUDY on PPAR alpha
ChemMedChem. DOI 10.1002/cmdc.201600526R1.
- [146] Mirjalili, V., Feig, M., Protein structure refinement through structure selection and averaging from molecular dynamics ensembles. *J. Chem. Theory Comput.* 2013, 9, 1294–1303.
- [147] Raval, A., Piana, S., Eastwood, M.P., Dror, R.O., et al., Refinement of protein structure homology models via long, all-atom molecular dynamics simulations. *Proteins Struct. Funct. Bioinforma.* 2012, 80, 2071–2079.
- [148] Terada, T., Kidera, A., Comparative molecular dynamics simulation study of crystal environment effect on protein structure. *J. Phys. Chem. B* 2012, 116, 6810–6818.
- [149] Dror, O., Schneidman-Duhovny, D., Inbar, Y., Nussinov, R., et al., Novel approach for efficient pharmacophore-based virtual screening: Method and applications. *J. Chem. Inf. Model.* 2009, 49, 2333–2343.
- [150] Triballeau, N., Acher, F., Brabet, I., Pin, J.P., et al., Virtual screening workflow development guided by the “receiver operating characteristic” curve approach. Application to high-throughput docking on metabotropic glutamate receptor subtype 4. *J. Med. Chem.* 2005, 48, 2534–2547.
- [151] Mysinger, M.M., Carchia, M., Irwin, J.J., Shoichet, B.K., Directory of useful decoys, enhanced (DUD-E): Better ligands and decoys for better benchmarking. *J. Med. Chem.* 2012, 55, 6582–6594.
- [152] Jain, A.N., Nicholls, A., Recommendations for evaluation of computational methods. *J. Comput. Aided. Mol. Des.* 2008, 22, 133–139.
- [153] Chen, H., Lyne, P.D., Giordanetto, F., Lovell, T., et al., On evaluating molecular-docking methods for pose prediction and enrichment factors, in: *Journal of Chemical Information and*

Modeling, 2006, pp. 401–415.

- [154] Gabanyi, M.J., Adams, P.D., Arnold, K., Bordoli, L., et al., The Structural Biology Knowledgebase: A portal to protein structures, sequences, functions, and methods. *J. Struct. Funct. Genomics* 2011, 12, 45–54.
- [155] Eswar, N., Webb, B., Marti-Renom, M. a, Madhusudhan, M.S., et al., Comparative protein structure modeling using MODELLER. *Curr. Protoc. Protein Sci.* 2007, Chapter 2, Unit 2.9.
- [156] Olsson, M.H.M., SØndergaard, C.R., Rostkowski, M., Jensen, J.H., PROPKA3: Consistent treatment of internal and surface residues in empirical p K a predictions. *J. Chem. Theory Comput.* 2011, 7, 525–537.
- [157] SØndergaard, C.R., Olsson, M.H.M., Rostkowski, M., Jensen, J.H., Improved treatment of ligands and coupling effects in empirical calculation and rationalization of p K a values. *J. Chem. Theory Comput.* 2011, 7, 2284–2295.
- [158] Brooks, B.R., Brooks, C.L., Mackerell, A.D., Nilsson, L., et al., CHARMM: The biomolecular simulation program. *J. Comput. Chem.* 2009, 30, 1545–1614.
- [159] Jo, S., Kim, T., Iyer, V.G., Im, W., CHARMM-GUI: A web-based graphical user interface for CHARMM. *J. Comput. Chem.* 2008, 29, 1859–1865.
- [160] Vanommeslaeghe, K., Hatcher, E., Acharya, C., Kundu, S., et al., CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comput. Chem.* 2010, 31, 671–690.
- [161] Vanommeslaeghe, K., MacKerell, A.D., Automation of the CHARMM general force field (CGenFF) I: Bond perception and atom typing. *J. Chem. Inf. Model.* 2012, 52, 3144–3154.
- [162] Eastman, P., Friedrichs, M.S., Chodera, J.D., Radmer, R.J., et al., OpenMM 4: A reusable, extensible, hardware independent library for high performance molecular simulation. *J. Chem. Theory Comput.* 2013, 9, 461–469.
- [163] Michaud-Agrawal, N., Denning, E.J., Woolf, T.B., Beckstein, O., MDAAnalysis: A toolkit for the analysis of molecular dynamics simulations. *J. Comput. Chem.* 2011, 32, 2319–2327.
- [164] Wolber, G., Langer, T., LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. *J. Chem. Inf. Model.* 2005, 45, 160–169.
- [165] Geer, L.Y., Marchler-Bauer, A., Geer, R.C., Han, L., et al., The NCBI BioSystems database. *Nucleic Acids Res.* 2009, 38.
- [166] Velaparthy, U., Wittman, M., Liu, P., Stoffan, K., et al., Discovery and initial SAR of 3-(1H-benzo[d]imidazol-2-yl)pyridin-2(1H)-ones as inhibitors of insulin-like growth factor 1-receptor (IGF-1R). *Bioorganic Med. Chem. Lett.* 2007, 17, 2317–2321.
- [167] Hunter, J., Droettboom, M., matplotlib, in: *The Architecture of Open Source Applications*, 2014, pp. 1–14.
- [168] Gaulton, A., Bellis, L.J., Bento, A.P., Chambers, J., et al., ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* 2012, 40.
- [169] Irwin, J.J., Shoichet, B.K., ZINC - A free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* 2005, 45, 177–182.
- [170] Madhavi Sastry, G., Adzhigirey, M., Day, T., Annabhimoju, R., et al., Protein and ligand preparation: Parameters, protocols, and influence on virtual screening enrichments. *J. Comput. Aided. Mol. Des.* 2013, 27, 221–234.
- [171] Jacobson, M.P., Friesner, R.A., Xiang, Z., Honig, B., On the role of the crystal environment in determining protein side-chain conformations. *J. Mol. Biol.* 2002, 320, 597–608.
- [172] Jacobson, M.P., Pincus, D.L., Rapp, C.S., Day, T.J.F., et al., A Hierarchical Approach to All-Atom Protein Loop Prediction. *Proteins Struct. Funct. Genet.* 2004, 55, 351–367.
- [173] Guo, Z., Mohanty, U., Noehre, J., Sawyer, T.K., et al., Probing the alpha-helical structural stability of stapled p53 peptides: molecular dynamics simulations and analysis. *Chem. Biol. Drug Des.* 2010, 75, 348–359.
- [174] Shivakumar, D., Williams, J., Wu, Y.J., Damm, W., et al., Prediction of Absolute Solvation Free Energies using Molecular Dynamics Free Energy Perturbation and the OPLS Force

- Field. *J. Chem. Theory Comput.* 2010, 6, 1509–1519.
- [175] Harder, E., Damm, W., Maple, J., Wu, C., et al., OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins. *J. Chem. Theory Comput.* 2015, Ahead of Print.
- [176] Wolber, G., Langer, T., LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. *J. Chem. Inf. Model.* 2005, 45, 160–169.
- [177] Zhao, W., Hevener, K.E., White, S.W., Lee, R.E., et al., A statistical framework to evaluate virtual screening. *BMC Bioinformatics* 2009, 10, 225.
- [178] Truchon, J.F., Bayly, C.I., Evaluating virtual screening methods: Good and bad metrics for the “early recognition” problem. *J. Chem. Inf. Model.* 2007, 47, 488–508.
- [179] Fawcett, T., An introduction to ROC analysis. *Pattern Recognit. Lett.* 2006, 27, 861–874.
- [180] Berthold, M.R., Cebron, N., Dill, F., Gabriel, T.R., et al., KNIME - The Konstanz Information Miner. *SIGKDD Explor.* 2009, 11, 26–31.
- [181] Cereto-Massagu??, A., Ojeda, M.J., Valls, C., Mulero, M., et al., Molecular fingerprint similarity search in virtual screening. *Methods* 2015, 71, 58–63.
- [182] Sood, M., Bansal, S., K-Medoids Clustering Technique using Bat Algorithm. *Int. J. Appl. Inf. Syst.* 2013, 5, 20–22.
- [183] Lauria, A., Bonsignore, R., Bartolotta, R., Perricone, U., et al., Drugs Polypharmacology by In Silico Methods: New Opportunities in Drug Discovery. *Curr. Pharm. Des.* 2016, 22, 3073–3081.
- [184] Martorana, A., Perricone, U., Lauria, A., The Repurposing of Old Drugs or Unsuccessful Lead Compounds by in Silico Approaches: New Advances and Perspectives. *Curr. Top. Med. Chem.* 2016, 16, 2088–2106.



Assessment of the quality of the PhD Thesis “Development and optimisation of computational tools for drug discovery” presented by the candidate Ugo Perricone

The main focus of the current thesis was the development and application of novel methods for computer assisted ligand design. A couple of novel methods were developed and were tested using data sets from literature. Good screening results could be obtained by the developed tools. Especially the testing of pharmacophore models derived from trajectories of molecular dynamics simulations (Chapter 4) represents a highly interesting part of the thesis. The dynamic pharmacophore models outperformed traditional pharmacophore models for the selected training sets. Further evaluations tests will show the general applicability and performance of dynamic pharmacophore models.

The thesis starts with a short introduction to the field of computer assisted design approaches and pharmacophore modeling. The introduction is clearly written and shows the deep scientific knowledge of Ugo Perricone. The references included contain all relevant publications in the field.

The thesis is of very good presentation and style, and shows evidence of the student’s ability to investigate critically a specific field of study, demonstrating an adequate knowledge and discussion of the literature in that field.

The thesis of Ugo Perricone generated significant new knowledge in the development and application of dynamic pharmacophore models. Several approaches have been applied to different targets and novel hypothesis were obtained. A minor criticism is the not always clearly defined own contribution of the candidate. As it is clear that all application work has been done by him, this is not always clear in the other disciplines, such as programming. A clear statement at the beginning of the individual chapters would have been helpful.

It must also be stated that some parts of Ugo Perricone's work is published in very good journals and I am confident that other parts of this PhD work will result in high impact papers.

The scientific value of her work is further demonstrated by the number of already published manuscripts and the novel molecules optimized by chemical synthesis and structure-based design. Based on the overall scientific value of the material, I can confirm the high scientific quality of the work. In summary, I can confirm that the candidate's contribution to the research and publications is sufficiently large to award him with PhD (Doctor Europaeus).

A handwritten signature in black ink, appearing to read 'Sippl', enclosed within a thin black rectangular border.

Prof. Dr. Wolfgang Sippl

To Whom It May Concern

Report on the Thesis document by Mr Ugo Perricone

Mr Ugo Perricone has submitted a thesis dissertation entitled "***Development and optimisation of computational tools for drug discovery***" to Università di Palermo, as an application to obtain the academic degree "*PhD*" from their appropriate PhD program. The study was performed under the supervision of Prof. Anna Maria Almerico.

The thesis manuscript is divided into three major chapters, together with an excellent preface provided as a sort of introduction to the field. All the results presented in these chapters are published in top tier, high quality international scientific journals with strict peer review and high impact.

In the preface, Mr Perricone provides the aim of the thesis and gives an overview on the state of the art of computational methods in modern drug design. Tools for virtual screening (VS), such as docking, and pharmacophore based approaches are one focus, as well as refinement of structures using molecular dynamics (MD) simulations in order to be able to rank compounds based on correctly predicted binding affinities. In this preface, Mr Perricone gives the rationale for the approaches used later during his studies, and also for the targets chosen as proof of principle, in order to identify new bio-active compounds interacting with proteins of therapeutic potential.

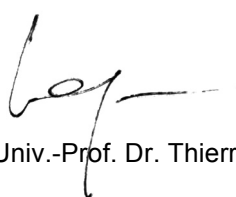
In particular, the first chapter of the thesis, *CHEMOMETRIC PROTOCOLS IN DRUG DISCOVERY*, Mr Perricone reports on the development and enhancement of a VS protocol calculating 3D molecular descriptors on the docked conformation of ligands. The so-called VLKA method was used and then further enhanced to the Conf-VLKA approach to predict the possible biological target for new molecules starting from the structural information contained in molecular descriptors calculated on a set of known inhibitors. Results of this part are published in the article *Conf-VLKA: A structure-based revisitation of the Virtual Lock-and-key Approach* in J. Mol Graph. Model. 2017, 71, 50-57.

In the second chapter, *THE APPLICATION OF MOLECULAR DYNAMICS TO VIRTUAL SCREENING* Mr Perricone describes the efforts towards integration of pharmacophore approaches for the analysis of molecular dynamics trajectories. The first paper in this context, *Evaluating the stability of pharmacophore features using*

molecular dynamics simulations, published in *Biochem. Biophys. Res. Comm.* 2016, 470, 685-689, indicates that the frequency information obtained from the MD simulations can be used to refine the pharmacophore model by adding or removing features and weighting their importance. In the paper *Comparing pharmacophore models derived from crystal structures and from molecular dynamics simulations* published in *Monatsh. Chem.* 2016, 147, 553-563, Perricone and co-authors suggest that even very simple structure refinement approaches, like the ones reported in their study, can lead to pharmacophore models that perform significantly better in virtual screening. In this chapter, two more application case studies are described in the target areas of PPAR alpha and the IGF-1R kinase domain. Also here, Mr. Perricone can demonstrate the advantage of combining molecular dynamics with pharmacophore methods for optimizing the performance of the experiments.

Finally, in the third chapter, *COMPUTATIONAL CHEMISTRY IN POLYPHARMACOLOGY AND DRUG REPURPOSING*, Mr. Perricone investigates one of the most interesting method for finding new drug candidates. Here, he studies virtual screening protocols for identifying drug polypharmacology. Finally, he summarizes the results published in two reviews dealing with the above mentioned topics (*Curr. Pharm. Des.* 2016, 22, 3073–3081 and *Curr. Top. Med. Chem.* 2016, 16, 2088–2106).

Overall, Mr Perricone has produced an impressive amount of data using a selection of the most advanced methods used in virtual screening. The results of his studies further contribute to the knowledge of compounds interacting with different targets since he has identified potential compounds that were shown to be active. He has undoubtedly shown his ability to derive scientifically correct conclusions. In view of these facts, this reviewer suggests the grade 'Excellent' for the present thesis. Clearly, he is in addition eligible for the international Doctor Europaeus title.



Univ.-Prof. Dr. Thierry Langer

Vienna, 2017-01-13