Alma Mater Studiorum – Università di Bologna

DOTTORATO DI RICERCA IN
CHIMICA

Ciclo XXVI

Settore Concorsuale di afferenza: 03/D1

Settore Scientifico disciplinare: CHIM08

Computational strategies to include protein flexibility in Ligand Docking and Virtual Screening

Presentata da: Dott.ssa Rosa Buonfiglio

Coordinatore Dottorato
Prof. Aldo Roda

Relatore
Prof. Maurizio Recanatini

Correlatore
Dott. Matteo Masetti
Prof. Andrea Cavalli

Esame finale anno 2014

To my lab mates

“In conclusion, I would like to emphasize my belief that the era of computing chemists, when hundreds if not thousands of chemists will go to the computing machine instead of the laboratory, for increasingly many facets of chemical information, is already at hand. There is only one obstacle, namely, that someone must pay for the computing time.”

Robert S. Mulliken, 1966.

TABLE OF CONTENTS

1. INTRODUCTION

1.1 BIOMOLECULAR RECOGNITION MECHANISMS

Proteins are dynamic structures subject to fluctuations occurring on a wide range of time scales from femtoseconds for ultrafast bond vibrations, to milliseconds and even seconds for large motions, resulting in a broad spectrum of conformations. These motions are at the base of biomolecular recognition, where ligands and receptors move towards complementary conformations to improve the binding affinity and regulate vital biological processes, such as signal transduction, metabolism and catalysis. The mechanisms underlying biomolecular recognition have been investigated extensively, in order to understand these processes and define novel therapies in a drug discovery context. The first explanation for ligand recognition was introduced in 1894 when Emil Fischer proposed the lock-and-key model.[1] According to this theory, ligands involved in biological reaction fit perfectly into proteins like a key into a lock, with almost no change in conformation. At that time, when molecular structures were poorly explored, this hypothesis gave an exhaustive and useful visual image of protein functions and was widely used as textbook explanation for biomolecular recognition events. However, the lock-and-key model based on a rigid body collision between ligands and proteins, neglected the emerging role of conformational plasticity supported by X-ray crystallography, NMR spectroscopy and single molecule fluorescence detection.[2] On the other side, although underestimating the conformational flexibility, this theory introduced the concept of shape complementarity of the bound components in a macromolecular complex and lasted more than 50 years. Indeed, the lock-and-key model is not completely incompatible with the existence of a protein conformational ensemble. In particular, multiple conformations that interconvert on a time scale faster than binding and dissociation events are indistinguishable from a single structure.[3] From a kinetic standpoint, rapidly interconverting species result in the same properties, without changing the observed relaxation kinetics of the system. In this case, for multiple conformations, constants for ligand binding and dissociation (k_{on} and k_{off}, respectively) are defined as ensemble averages of rate constants over the entire population.[4] Therefore, the lock-and-key model may describe properties of single structures or structural ensemble, unless the conformational transitions occur over time scales which are relevant for binding/dissociation processes. The introduction of this model allowed to investigate mechanisms at the base of biomolecular recognition, so as to provide a number of insights to processes spanning from binging events to catalysis. It is clear that this theory did not cover all cases and then, further models were introduced to explain inconsistencies such as noncompetitive inhibition and other relevant discrepancies.[5] In the last decades, the concepts of induced fit[6] and conformational selection[7] emerged to explain the intricate biomolecular recognition mechanisms. In details, the first theory, proposed by Koshland in 1958, was based on the following postulates: “a) a precise orientation of catalytic groups is required for enzyme action; b) the substrate may cause an appreciable change in the three-dimensional relationship of the amino acids at the active site; and c) the changes in protein structure caused by a substrate will bring the catalytic groups into the proper orientation for reaction, whereas a non-substrate will not”.[6] This theory, generalized to diverse biological mechanisms, pointed out that a fit between macromolecules and ligands occurs *only after* the changes induced by the substrate itself, likened to the fit of a hand in a glove. It incorporated Fisher’s concepts of structural complementarity but, at the same time, took into account the enzyme flexibility. However, kinetic studies showed that the induced fit theory failed in describing all possible binding mechanisms.[8]

In 1965, Monod, Wyman and Changeux proposed the allosteric model assuming that, when an allosteric binding event occurs, a shift of the equilibrium of two(or more) *pre-existing* conformational states is observed.[9] Conformational selection (also referred to as population selection, selected fit and fluctuation fit) is conceptually identical to the original Monod, Wyman and Changeaux model but has been extended to describe a large variety of monomeric non regulated metabolic[10-13] or signaling enzymes,[14] their intrinsic dynamics, binding mechanisms[15] and folding of disordered structures.[16] Although this concept was used in the 1980s,[17-18] only during last decade it has emerged by the insightful contribution of Nussinov and coworkers as one of the prevalent mechanisms related to biomolecular recognition.[16, 19-21] In particular, the description of the energy landscape of proteins by Fraugenfelder, Sligar and Wolynes[22] led, in 1999, to the generalized concept of “conformational selection and population shift”. The energy landscape is a map of all possible conformations of a molecular entity and their corresponding energy levels on a multidimensional Cartesian coordinate system. Conformations in dynamic equilibrium in the energy landscape are called microstates or conformational substates. It is also referred to as “folding funnel” with many highly unfavorable states that collapse via multiple routes into possibly several favorable folded states.[23-25] The populations of these substates follow statistical thermodynamic distributions and the timescale of conformational transitions depends on the height of the energy barriers between substates. In case of low free energy barriers in terms of Boltzmann energy, more than one conformational state can exist in solution. Ligands can bind the most energetically favorable conformations or one of the high energy conformational substates existing in solution. In all cases, this ligand binding causes a population shift, that is a redistribution of the relative populations of conformational substates *pre-existing* in solution.

In summary, in both the induced fit and the conformational selection models, macromolecules assume multiple conformations and protein flexibility is considered in binding mechanisms. However, two different scenarios are depicted: one where the conformational ensemble is generated after ligand binding (induced fit) and the other where the conformational transitions pre-exist (conformational selection). In other words, in the former case, the free (unbound) macromolecule is trapped as a single conformation and undergoes conformational transitions when bound, while the opposite situation occurs in the case of conformational selection.

Figure 1 shows the free energy landscape of macromolecular structures and binding according to the lock-and-key, induced fit and conformational selection theories. In the first process (Figure 1a), no conformational changes appear after ligand binding and protein exists as the same structure in both the unbound and bound state. In this context, the lock-and-key model can be considered a limiting case of conformational selection where ligand binds to the lowest energy and unique conformation. In Figure 1b, induced fit mechanism is reported: a single protein first binds the ligand and then undergoes conformational changes to optimize the complex, corresponding to multiple substates. In the last case, the free macromolecule pre-exists as conformational ensemble and the ligand binding stabilizes one of them, resulting in a single final substate.

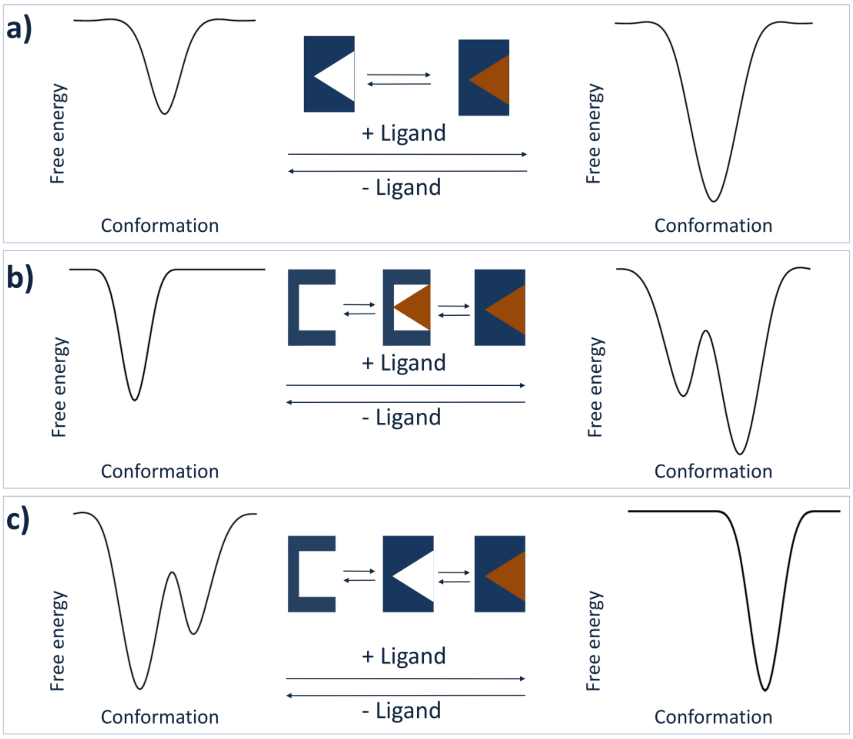


Figure 1. *The energy landscape view of protein structures and binding. a) Lock-and-key model. b) Induced fit model. c) Conformation selection model. Receptors and ligands are represented in blue and orange, respectively.*

Induced fit and conformational selection can be also described through a thermodynamic cycle (see Figure 2 for a simplified scheme). According to this cycle, when the concentration of the higher energy unbound conformation (P_2) is larger than the concentration of P_1L, that is the intermediate complex of the induced fit process, the prevalent biomolecular recognition pathway will be the conformational selection.[26] In this latter process, the rate of formation of the final complex P_2L depends linearly on the concentration of P_2 and nonlinearly on the total concentration of the protein ($[P_1 + P_2]$). On the other side, when the concentration of the higher energy conformation (P_2) is lower than 5%, it is difficult to distinguish the main biomolecular recognition mechanism between the induced fit and conformational selection.

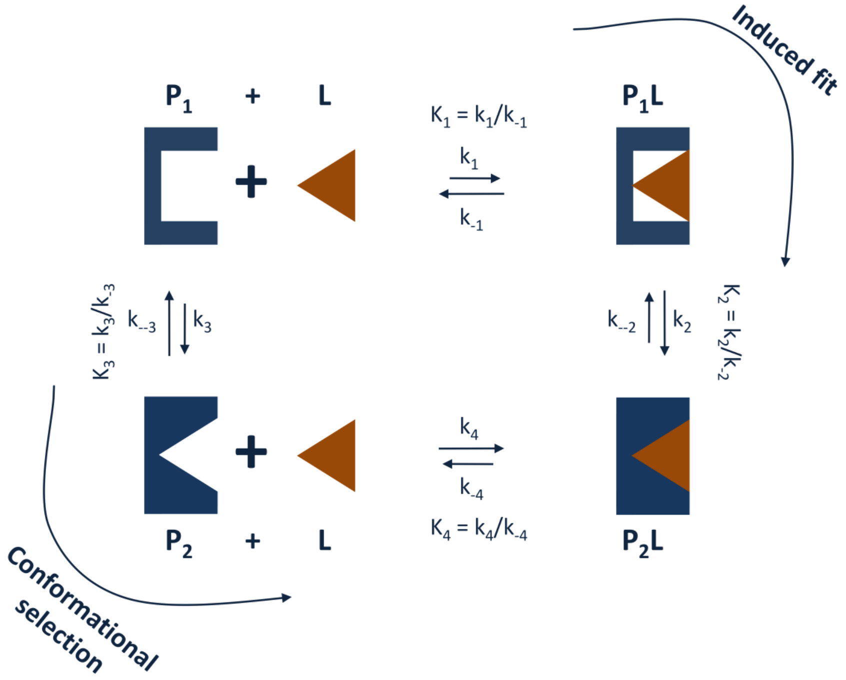


Figure 2. *Induced fit and conformational selection in a thermodynamic cycle. P_1 and P_2 are pre-existing unbound conformations, in agreement with the conformational selection theory. P_1L is the intermediate protein-ligand structure which undergoes conformational changes resulting in the final complex P_2L. The capital and lower-case letters for each equilibrium describe the thermodynamic and kinetic rate constants, respectively. Receptors and ligands are represented in blue and orange.*

The mentioned models seem to be antagonistic, as “all or nothing” phenomena. In reality, in many instances, both mechanisms may cooperate in the same processes. For instance, various cases in which conformational selection event is followed by optimization of side chain and backbone interactions based on induced fit mechanisms have been reported.[27-28]

In this scenario, the original conformational selection model has been extended embracing a repertoire of selection and refinement mechanisms. Induced fit can be thus perceived as a subset of this repertoire contributing to stabilize interactions between bound partners.[20] All mechanisms described so far consider a macromolecule as multiple conformations and the second binding partner as a single conformational state corresponding to small and/or rigid ligands, such as small molecules or DNA. On the contrary, in the extended conformational selection model, the “native states” of both binding partners are accounted for as multiple conformations, located at the low-energy regions of the energy landscape. According to this process, binding partners undergo conformational selection and preceding or subsequent adjustment steps, forming the binding trajectory. The mutual encounter modifies the populations of the partner conformations and, thus, the shape of their energy landscape. This mutual condition consisting of step-wise selection and encounter mechanism has been termed as "interdependent protein dance",[29] where the conformation of a partner represents the environment (preconditions) for the second molecule and vice versa. The lock-and-key, the induced fit, the original conformational selection and the conformational selection followed by the induced fit are special cases of this extended conformational selection model. It can describe biomolecular recognition processes involving proteins and also the binding mechanism of RNA molecules, as recently reported.[30-31]

The biomolecular recognition mechanisms can be potentially distinguished by kinetic measurements, based on the dependence of the rate of relaxation to equilibrium, k_{obs}, on the ligand concentration [L].[3] For decades, a value of k_{obs} increasing with [L] has been seen as diagnostic of induced fit, while a value of k_{obs} decreasing hyperbolically with [L] indicated conformational selection. However, this simple conclusion is only valid under the unrealistic cases in which conformational transitions are rate-limiting in the kinetic mechanism compared to binding and dissociation events. In general, induced fit occurs when values of k_{obs} increase with [L], but conformational selection is more versatile and is related with values of k_{obs} that can increase, decrease or are independent of [L]. According to this more adaptable repertoire of kinetic properties, conformational selection is considered always sufficient and often necessary to explain all experimental systems reported in literature so far and then, is fundamental for binding mechanisms. On the other hand, induced fit is never necessary and only sufficient in some biological events. While these two mechanisms can be distinguished by kinetic measurements, direct characterization of sparsely-populated states of proteins in the free and bound forms are experimentally challenging. In fact, unless trapped, such short-lived states are invisible to conventional biophysical techniques like X-ray crystallography and NMR spectroscopy.[32-33] Recently, developments in NMR field involving paramagnetic relaxation enhancement (PRE) have led to the detection and exhaustive visualization of these sparsely-populated states, applicable to a wide-range of proteins characterized by transient, short-lived conformational transitions. Moreover, single molecule measurements are emerging tools able to characterize the conformational properties related to protein functions.[34] These approaches used to study biomolecular recognition at the single molecule level include the single molecule fluorescence resonance energy transfer (smFRET) and single molecule functional studies. The former reveals the extent and lifetime of conformational motions in catalytic steps and ligand-mediated interconversions between conformations, typically averaged in non-synchronized ensemble measurements.[35-36] On the other hand, single molecule functional studies are useful for the direct observation of activity heterogeneities for single and multiple enzymes.[37-38] In general, these methods quantify the activity, abundance and lifetime of multiple conformations and transient intermediates in the energy landscape.

The development of the mentioned theories revealed the critical role of protein flexibility for the recognition mechanisms in all biological systems. That being so, protein plasticity has important consequences on drug design and must be considered as fundamental requirement to study biological processes and identify and design small molecule inhibitors.

1.2 PROTEIN FLEXIBILITY IN DRUG DISCOVERY

Protein flexibility is required for biological effects, metabolism, transport and function. For example, residues involved in the catalytic mechanism of enzymes are often flexible, some receptors need to transmit the message of ligand binding from outside the membrane to the inside, channels can exist in an open or close form, etc.[39] Protein functions are poorly understood when a single protein structure is under investigation, since typical intrinsic dynamics and wide range of motions are neglected. In other words, using a single protein conformation in a drug design approach to accommodate all ligands corresponds to accept the outdated lock-and-key model of binding.

Conformational changes of side chains within binding sites are essential to identify novel inhibitory compounds. It is known that 8 out of the 20 amino acids can undergo structural rearrangements with a chance of 10-40%, making rigid structures unsuitable for structure-based drug discovery applications.[40]

In a more realistic scenario where multiple conformations pre-exist, design of molecules stabilizing specific structures within the ensemble may improve the probability to succeed in discovering novel active compounds and better understand biomolecular recognition processes. Therefore, proteins can adopt several conformations forming what is known as conformational ensemble. Protein motions are also important in terms of selectivity. Several therapeutic targets belong to families of highly related proteins and then, selectivity is required for pharmacological activity. In particular, binding sites consist of conserved residues located in almost the same positions, as a consequence of binding of the same natural ligands or because these macromolecules carry out reactions with the same chemistry. On the other side, the remaining residues are less homogeneous and promote different types of protein motions. This flexibility bestows a kind of diversity among proteins otherwise similar, to be exploited in the search for selective ligands.[39] Also, the understanding of binding mode and affinity based on interaction among compounds, proteins and water molecules, is really important for rational drug design. Free energy that quantifies the preference of a state (free or bound) compared to the others, conveys the affinity of binding events. In the light of these considerations and of the influence of flexibility on drug design, several tools are now available to explore the conformational space and estimate of free energy of binding.

The first attempt to take advantage of protein plasticity has been the use of multiple protein structures derived from NMR studies and X-ray crystallography, considered as the key methods to characterize molecular structures. In many cases, using conformational ensemble improves predictivity of Virtual Screening, but only a limited set of conformations for all existing proteins has been experimentally solved so far. Crystal structures are sometimes heterogeneous due to crystal packing effects, even errors or uncertainties introducing inaccuracies in structures, especially when they are solved at a resolution higher than 1.6Å.[41] Another way to generate multiple conformations to be employed in structure-based drug design approaches is the use of computational tools including ligand docking,[40] low-frequency normal modes,[42] and sampling approaches, such as Molecular Dynamics (MD),[43] enhanced methods[44] and Monte Carlo simulations.[45] They represent a valid alternative to experimental techniques, since they provide an atomistic-level detail to several processes such as protein folding[46] or biomolecular recognition,[47] and help to define biologically and pharmacologically significant conformations sometimes difficult to characterize with experimental methods. In addition, these techniques allow the identification of additional putative binding sites that may result hidden in the unbound protein structures, shedding light on potential allosteric sites. This scenario points out the importance of protein flexibility in discovering new drugs involved also in allosteric pathways. Ultimately, computational and experimental approaches together seem to be the best combination to study protein fluctuations and interactions.[48]

Overall, these methods probe both local receptor flexibility in close proximity to the binding site and global flexibility of protein domains, corresponding to side-chain and backbone fluctuations, respectively. In particular, backbone flexibility concerns large-scale domain fluctuations and disordered regions such as flexible loops.[49] In the next paragraphs the available computational methods used to probe local and global flexibility are overviewed.

1.3 PROTEIN FLEXIBILITY IN COMPUTER-AIDED DRUG DISCOVERY

In consideration of the wide diversity of computational methods employed to study protein flexibility, some kind of classification can be useful. In particular, we can distinguish: *i*) approaches investigating local and/or global flexibility and *ii*) protocols that make use of unique or multiple structures. However, an absolute classification cannot be defined and, also, the selection of the most appropriate approach depends on the features of the biological system of interest.

Procedures taking into account single protein structures subject to slight side chain movements are described. The first and easiest way to consider small conformational fluctuations within a protein binding site is the use of implicit methods.[50] Many force-field docking algorithms compute the van der Waals forces through the Lennard-Jones potential which increases rapidly to infinity when interatomic distance reduces to zero. It results in large energy penalties when minor steric clashes arise. In other words, potential good binders that do not fit exactly into the rigid binding site, may result in a low rank and bad score. With this method, known as "soft docking", the Lennard-Jones potential is softened through the introduction of a more forgiving potential; thus, minor steric clashes are tolerated and pose orientations including slight overlaps between ligands and proteins are retained, allowing an approximate consideration of protein plasticity. Soft docking is efficient in terms of computational costs, because no additional calculation time is required to evaluate the scoring function and is easily implemented in docking software. However, its main limitation is that it increases the number of false positives and the flexibility is only indirectly taken into account.

Another simple way to include side chain torsions in ligand docking algorithms arises from the reorientation of hydrogen atoms. In particular, rotational sampling of hydrogens and lone pairs of hydrogen-bond donor and acceptor atoms have been included in genetic algorithms, allowing for the promotion of stronger hydrogen bonds.[51] Some computational methods accounting for full side chain flexibility are based on an extension of this simple approach, where torsional sampling around single and double bonds are evaluated while bond lengths and angles are kept fixed. These algorithms incorporating side chain mobility use rotamer libraries, introduced for the first time by Leach.[52] According to his method, no rotameric states have been assigned to alanine, glycine, and proline. Methionine, lysine, and arginine have been given 21, 51, and 55 rotameric states, while the remaining amino acids range from 3 to 10 rotameric states. Thus, the most energy favorable combination of side chain conformers and ligand orientations is defined. A further approach combines the use of rotamer libraries to generate multiple side chain conformations and an energetic optimization of the docked system, in particular of the side chains in close contact with ligands, so as to strengthen molecular interactions.[53] Optimization techniques include simulated annealing, steepest-descent minimization, Monte Carlo (MC) sampling, or other methods.[54]

A different procedure, that still considers side chain rearrangements, is introduced with the "InducedFit" docking developed by Sherman and coworkers and implemented as a tool of GLIDE docking software, in the Schrödinger suite.[55-56] In this case, it accounts for both ligand and protein flexibility and iteratively combine rigid receptor docking with protein structure prediction techniques (GLIDE and PRIME tools).[55] By paraphrasing the name, induced fit rearrangements and local movements within the active site upon ligand binding are included in docking protocol. The first step is the ligand docking with GLIDE. During this step, highly flexible side chains can be temporarily removed and converted with alanine residues, so as to reduce steric clashes. For each pose, PRIME tool is used for structure prediction, in order to reorient side chains and accommodate the ligands. Then, ligands and binding site are energy minimized. Finally, a re-docking of each ligand in the low energy protein structure is performed, and small molecules are ranked according to the GLIDE score. A similar algorithm known as IFREDA (ICM-flexible receptor docking algorithm) has been developed by Cavasotto and Abagyan, to define multiple conformations when a single protein structure is available.[57] IFREDA generates a set of protein conformations through a flexible ligand docking of known active compounds to a flexible receptor. The resulting conformational ensemble is used for flexible ligand–rigid receptor docking and scoring. Finally, a merging and shrinking procedure allows to condense results of the multiple Virtual Screenings, so as to improve the enrichment factor. With the exception of some few cases, the mentioned techniques probe mainly local receptor flexibility.

An alternative method is RosettaLigand, implemented in the popular software Rosetta, which incorporates full protein backbone and side chain flexibility and considers both ligand ensemble generation and receptor movements during docking steps.[58-59] The latest release includes new features, such as docking of multiple ligands simultaneously and redesign of the binding interface during docking.[60] However, treating backbone flexibility in docking protocols remains certainly challenging, because a large number of degrees of freedom needs to be considered. Several other methods have been published so far, which consider global flexibility, and also slight side chain movements. They are mainly based on the use of multiple protein structures and are known as ensemble-based methods. Conformations can derive from NMR or crystal structures or computational methods exhaustive in sampling large conformational space and derived from the protein flexibility analysis, as already introduced in the previous section.

Among docking-based methods exploiting pre-existing protein ensembles to consider protein flexibility, the algorithm FlexE, an extension of the FlexX software, defines rigid and flexible regions based on superimposition of experimental structures.[61] Backbone and side chains in agreement are averaged, while disordered regions including diverse orientations of flexible side chains are collected as rotamer library. During docking, alternative flexible regions are explicitly taken into account, that can be combined to generate novel conformations not experimentally observed. The advantage of this method is the structural novelty introduced with the recombination approach; however, this split and join procedure to combine protein fragments is fast and accurate mainly when side chains are involved, instead of wider conformational changes. The four-dimensional docking is another method applied to multiple protein structures.[62] According to this approach, receptor flexibility is represented as the fourth discrete dimension of the small molecule conformational space, with multiple recomputed 3D grids from optimally superimposed conformers merged into a single 4D object. The four-dimensional docking seems to be advantageous in terms of computational costs and accuracy compared to other developed methods.

A further way to dock ligands in multiple structures has been developed by Knegtel and coworkers, consisting of docking ligands into an ensemble-average energy grid that is defined as the average of grids calculated for all the protein conformations. However, this approach gives a single docking score for each ligand that may result in inaccurate outcomes, due to weakness of some scoring functions compared to a consensus of multiple scores.[63] Most recent studies, carried out by Craig et al.[64] and Rueda et al.[65] have demonstrated that the use of multiple conformations generated through known active compounds leads to a better enrichment compared to the initial protein structures. A similar conclusion has been observed by Novoa and coworkers by using homology models as starting multiple structures for docking and Virtual Screening.[66]

Multistep approaches can also be used to take advantage of single methods, some of which already described. For instance, InducedFit docking and IFREDA are two algorithms in which ligand poses generated with a fast procedure are subject to refinements with longer and more accurate computational methods.

Docking methods are useful in all the available versions to screen large number of ligands in a reasonable computational time and identify novel hit compounds. As reported above, several new improvements have been introduced to include receptor flexibility, referred as backbone and side chain fluctuations. However, when protein undergoes large-scale movements, docking approach is unsuitable because a large number of degrees of freedom is added to the space sampling. In this context, other computational tools are exploited in drug discovery.

The Relaxed Complex Scheme (RCS) is a protocol, introduced by Lin et al. and inspired by the experimental “SAR by NMR” and “tether” methods to discover molecules with high binding affinity.[67] This approach arises from the idea that ligands may interact with conformations that only rarely are explored and experimentally solved. In the first step, a long MD simulation of unbound proteins is performed, in order to sample extensively the conformational space. Subsequently, MD snapshots are selected to carry out docking of mini-libraries of potential active compounds. In this way, ligands are associated with a spectrum of binding scores and ranked according to various properties of this score distribution. In the extended version of this scheme, the MM/PBSA (Molecular Mechanics/Poisson Boltzmann Surface Area) approach is used to re-score the docking results, allowing to find the best ligand–receptor complexes concerning the correct binding modes.[68] It is important to keep in mind that, when MD simulations are used to collect conformations for a Virtual Screening, reliability of resulting structures and enrichment introduced with sampling are unknown a priori.[69] In general, ensemble consisting of crystal structures may lead to better estimate binding affinity, but, conversely, structures from MD simulations may improve predictivity.

In a different perspective, multiple protein structures extracted by MD simulations, as reported in the RCS approach, can be used to define a dynamic pharmacophore model and predict ligand binding. Carlson and coworkers have developed this approach, where pharmacophoric description is extracted from each snapshot through ligand probes and then, collected pharmacophores are clustered and analyzed, in order to define the key elements preserved during the MD simulations and exploited to discover novel ligands with complementary chemical features.[40, 70]

Molecular Dynamics is widely used as tool for several purposes, such us studying receptor flexibility before docking or including solvent effects. Also, MD simulations are used to stabilize ligand-receptor complexes resulted from docking studies, in order to enhance the strength of binding and rearrangements toward more favorable energy structures.[54, 71] Therefore, this computational approach is applied not only to understand mechanisms playing a key role in the biomolecular recognition process, but also to improve predictivity and enrichment in Virtual Screening context, of which the Relaxed Complex Scheme represents a successful example. The timescales simulated with MD span between nanoseconds to microseconds, and recently even milliseconds, allowing generation of multiple low-energy protein conformations and also refinements of active site residues. With improvements of computer hardware and software in terms of efficiency and speed (e.g. GPU and the construction of the special machine Anton),[72] longer and longer MD simulations are performed which are extending knowledge on biomolecular recognition at molecular level. Moreover, binding events to protein targets,[47] full atomic resolution of protein folding and description of folding mechanisms are investigated by long MD simulations.[46] However, some limitations prevent a more extended applicability of this method. For example, to recover a Boltzmann ensemble of structures and obtain converged statistics, the energy barriers between relevant states need to be crossed repeatedly and then, fast and accurate sampling is required. Very long MD simulations may reach this goal. However, in order to optimize computational costs and, at the same time, investigate interesting processes taking place on seconds or longer timescales, new methods have been developed. They explore exhaustively the conformational space, sometimes poorly sampled with conventional MD simulations, and also accelerate the crossing of high energy barriers. The enhanced sampling is possible through the introduction of artificial biases in the simulations. Umbrella sampling,[73] and metadynamics[74] are some possible computational tools to speed up conformational exploration. In these cases, collective variables or a transition pathway between known initial and final state are defined a priori. Briefly, umbrella sampling uses a bias potential to ensure exhaustive sampling along the coordinate reaction. Separate simulations or windows are carried out that overlap to connect the initial and final states. Metadynamics employs collective variables to describe the system of interest. A history-dependent potential bias is introduced in the Hamiltonian of the system, by addition of Gaussians aimed at preventing the system to revisit configurations that have already been explored.

Replica exchange[75] and accelerated Molecular Dynamics[76] represent other computational methods, widely used to accelerate conformational space sampling. In details, replica exchange exploits high temperatures and multiple parallel MD simulations to overcome energy barriers on the potential energy surface and enhance sampling of new conformational space. Accelerated Molecular Dynamics does not require definition of a reaction coordinate or collective variables. It relies on modifying the potential energy surface based on the application of a boost potential at each point of the MD trajectory which depends on the difference between a user-defined reference energy and the real potential energy of the MD force field. Caflisch and co-workers have proposed an alternative method to guide the exploration of conformational space, known as free-energy guided sampling (abbreviated as FEGS).[77] Also, this method does not require the choice of collective variables or reaction coordinates, but uses the cut-based free energy profile (cFEP)[78] and Markov state models (MSMs)[79] to speed up sampling of relevant conformations.

In alternative to these MD-based methods, normal mode analysis (NMA) has been widely exploited to define flexible protein domains with a lower computational cost.[80] Normal modes are harmonic oscillations defining the intrinsic dynamics of the protein within an energy minimum.[81] Elastic Network (EN) analysis, that still relies on a NMA framework, is an alternative method based on a more simplified protein representation, consisting of a network of elastic connections between atoms.[82] Applications of these computational approaches in drug discovery contexts have been extensively reported. For example, Zacharias and co-workers modeled global backbone flexibility in a docking protocol, by relaxation in a few pre-calculated soft collective degrees of freedom of the receptor corresponding to eigenvectors of the protein.[83] They were computed through a Normal Mode analysis related to Gaussian network models as described by Hinsen.[84] Also, determination of the relevant normal modes representing binding pocket flexibility followed by perturbation of the protein along these modes was proposed by Cavasotto et al. to define novel conformational ensembles.[85] More recently, Abagyan group derived multiple protein conformations through Monte Carlo sampling performed on the collective variable space defined by all heavy-atom EN-NMA.[86] All these methods take advantages of Normal mode-based analysis to guide the sampling of conformational space resulting in the final generation of multiple conformations.

In the light of the mentioned approaches, a parallelism between the biomolecular recognition theories and the strategies used to account for protein flexibility in computer-aided drug discovery can be advanced. Figure 3 shows the computational techniques aimed *i*) to generate multiple conformations (conformational selection box) and *ii*) properly accommodate molecules within the binding site (induced fit box). Normal Mode Analysis can be employed to define soft modes as additional variables for rapid relaxation of the receptor structure during docking (induced fit), as described by Zacharias et al., or as tool to guide the generation of multiple conformations. In general, computational methods linked to induced fit theory treat side chain flexibility, whereas sampling methods and low-frequencies normal modes result in the definition of multiple geometries of the protein backbone, along with side chain reorientations. This separation is useful to understand the applicability of each method, but has not to be considered as a strict classification. In fact, ligand binding can also be seen as a combination of a conformational selection stage followed by slight structural rearrangements within the complex, according to the induced fit theory. In other words, upon ligand binding of one of the sampled conformations, the complex can be subjected to side chain rearrangements within binding pockets, through the diverse docking methods dealing with local flexibility. Finally, FlexE and four-dimensional docking can be simplistically mentioned among the methods related to conformational selection theory, since they start with multiple pre-existing structures. On the other side, the method developed by Knegtel and co-workers based on the ligand docking into an ensemble-average energy grid, is difficult to classify on the basis of the ligand recognition theories, since multiple conformations are condensed as a single entity in docking run.

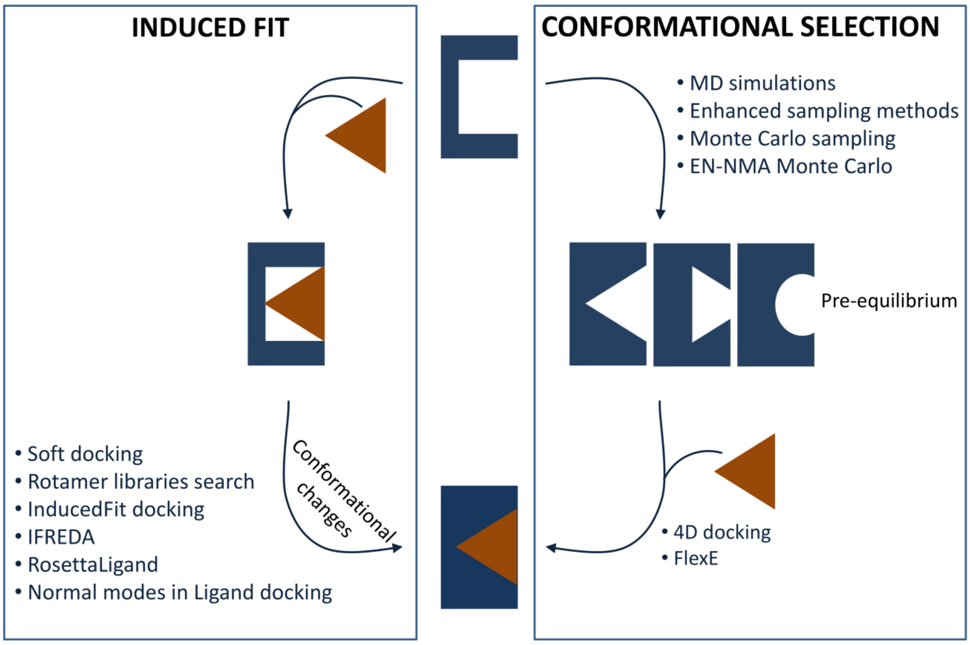


Figure 3. *Induced fit and conformational selection, with the corresponding computational methods dealing with local and global flexibility. Receptors and ligands are represented in blue and orange, respectively.*

All the mentioned computational approaches represent excellent tools to explore the conformational space and identify relevant conformations, sometimes not experimentally solved. More and more methods and applications have been developed, aimed at accounting for receptor flexibility and improving predictivity of drug discovery. They are described in details in several reviews.[87-90]

1.4 ENSEMBLE-BASED VIRTUAL SCREENING

Virtual Screening is a computational strategy based on ligand docking, aimed to reduce the large virtual space of chemical compounds to a more reasonable number for further synthesis and pharmacological tests against biological targets of interest, which could lead to potential drug candidates. The Ensemble-based Virtual Screening is based on using multiple protein conformations in a docking study. Receptor structures from X-ray, NMR studies or homology models can represent the starting geometries to be used as they are, or as input for computational methods to generate multiple conformations. An extensive overview of such *in silico* tools has been depicted in the previous section. Figure 4a below describes the main steps of this protocol.

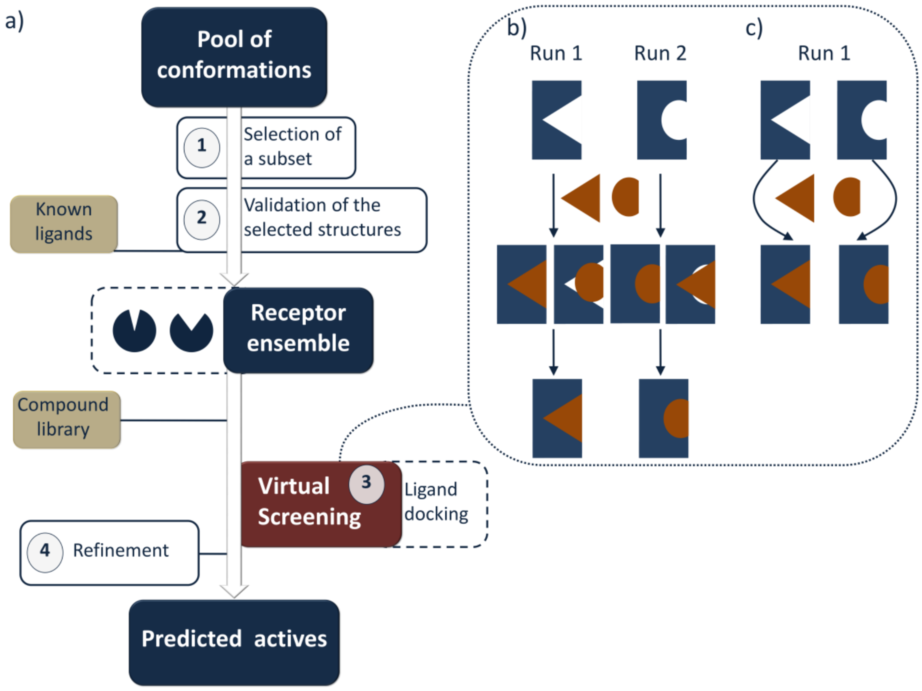


Figure 4. *a) Schematic representation of the Ensemble-based Virtual Screening. b) and c) Combinations of multiple conformations in an Ensemble-based Virtual Screening.*

Step 1. The pool of conformations is analyzed, in order to select the most biologically relevant structures. It is possible to extract conformations from MD or MD-based simulations at regular intervals, but this approach may introduce redundancy and low structural diversity. In alternative, clustering methods are widely used to select representative conformations for Virtual Screening.[65, 91] Several criteria can be used to cluster multiple structures, such as structural similarity in terms of RMSD values. To focus on the most significant geometries, this parameter may be calculated on regions of interest like the binding site, residues involved in binding mode or flexible protein domains linked with the active site. The selection of pairwise metric, clustering metrics and atoms used for the comparison can strongly affect the cluster analysis performance. Although cluster algorithms give generally satisfying results, they can exhibit limitations, such as the tendency to generate small, singleton or homogeneously sized clusters or the low stability to noise and changes in clustering parameters.[92] Alternatively, reduced dimensionality of data can be exploited to manage reasonable numbers of conformations.[93] For example, after long MD simulations, the most relevant degrees of freedom can be obtained by the Principle Component Analysis.[94] Network analysis is another emerging tool to analyze protein dynamics allowing to extract information regarding not only the structural diversity but also connectivity of the collected conformations.[95]

Step 2. Subsequently, the selected conformations are validated through ligand docking. In particular, a set of known active and inactive chemicals are docked in all the single structures of the ensemble. The goal of this retrospective study is to assess the predictive power of the conformations belonging to the final ensemble in perspective of the Virtual Screening. Several metrics can be used, including the Receiver Operating Characteristic and corresponding Area Under the Curve (ROC and AUC, respectively).[96] These parameters describe the probability that actives will be ranked earlier than inactives in a rank-ordered list, as obtained after a docking run. This analysis can be useful to know if docking results bring added value compared to the random picking of compounds. In fact, this scenario should be translated in a real Virtual Screening, where a small proportion of the best ranked compounds is selected for biological tests. A good early recognition in a retrospective docking results in a high probability to find active compounds among the small proportion of ligands selected for experiments.

Sometimes, the validation study is used directly to select conformations, skipping step 1. In other words, the collected conformations are subject to a retrospective Virtual Screening and the best performing structures are selected based on their early recognition capability. However, the applicability of this procedure is restricted to a few ligands and a low number of protein conformations. In fact, a retrospective docking study of thousand ligands on large number of conformations collected by long Molecular Dynamics simulations as such, maybe computationally demanding. An example of this application is the ALiBERO protocol, recently developed by Rueda et al.[97] It is based on EN-NMA Monte Carlo simulations to collect multiple conformations, and selects the best pockets maximizing the recognition of ligand actives from decoys.

Step 3. The validated conformations are then used for docking a ligand database. Several docking software and scoring functions are available and are used to rank large numbers of ligands and identify hit compounds. Multiple conformations can be combined in an Ensemble-based screening in two different ways: *i*) independent docking runs and selection of the best ranked pose for each ligand in one of the collected conformations, or *ii*) a single docking run with all the protein structures.[87] Figures 4b and 4c schematize these procedures.

Compounds used for docking are selected from databases of commercially-available compounds[98] consisting of large number of molecules (≈ 10^6, 10^7), filtered out according to drug-likeness or lead-likeness criteria, which are based on the Lipinski's rule of five.[99] These filters ensure that molecules will be chemically stable and reduce risk of toxicity. Substructure filters to remove Pan Assay Interference Compounds (PAINS), or promiscuous compounds are also used.[100-101]

Step 4. A lead optimization phase can follow the hit identification, with the goal to improve binding affinity with more sophisticated and accurate methods. Molecular mechanics/generalized Born surface area (MM/GBSA)[102] and the molecular mechanics/Poisson-Boltzmann surface area (MM/PBSA)[103] are efficient techniques in calculating free energy of binding by means of molecular mechanics force fields and continuum solvent models. They can be used in conjunction with scoring functions or in a postprocessing step to enhance ranking and binding energy prediction of ligands.[104] In the lead optimization step, the most popular computational technique is the Free Energy Perturbation (FEP), a very useful tool for guiding molecular design. This method consists of alchemic transformations between two states, in conjunction with MD or Monte Carlo simulations in implicit or explicit solvent. Alchemic free energy methods are efficient and accurate but computational demanding. Moreover, an optimal accuracy is obtained only within a congeneric series of compounds.

1.5 AIM OF THE WORK

The dynamic character of proteins strongly influences biomolecular recognition mechanisms. With the development of the main models of ligand recognition (lock-and-key, induced fit, conformational selection theories), the role of protein plasticity has become increasingly relevant. In particular, major structural changes concerning large deviations of protein backbones, and slight movements such as side chain rotations are now carefully considered in drug discovery and development. In the light of what described so far, it is of great interest to identify multiple protein conformations as preliminary step in a screening campaign. Through the projects described in details in the Chapters 3 and 4, protein flexibility has been widely investigated, in terms of both local and global motions, in two diverse biological systems. On one side, Replica Exchange Molecular Dynamics has been exploited as enhanced sampling method to collect multiple conformations of Lactate Dehydrogenase A (LDHA), an emerging anticancer target. The aim of this project was the development of an Ensemble-based Virtual Screening protocol, in order to find novel potent inhibitors.[105] On the other side, a preliminary study concerning the local flexibility of Opioid Receptors has been carried out through ALiBERO approach, an iterative method based on Elastic Network-Normal Mode Analysis and Monte Carlo sampling. Comparison of the Virtual Screening performances by using single or multiple conformations confirmed that the inclusion of protein flexibility in screening protocols has a positive effect on the probability to early recognize novel or known active compounds.

2. METHODS

2.1 MOLECULAR MODELING

Molecular modeling is the ensemble of all theoretical and computational methods required to describe and evaluate the properties of biological systems. These methods represent an important tool to investigate and better understand experimental data, gain knowledge on molecular structures and biological mechanisms, systematically explore structural/dynamical/thermodynamic patterns, etc. Experimental structures from X-ray or nuclear magnetic resonance provide the basic elements for molecular modeling. In fact, they are often used as starting point to study biological mechanisms, protein flexibility, energetic contributions to ligand-protein interactions, etc. In general, all methods of which molecular modeling takes advantages allow atomistic level description of the molecular systems, at different degrees of detail. Mainly, molecular modeling uses two different approaches: quantum mechanics and molecular mechanics. The former approach aims at solving the Schrödinger equation for each atom of the system, in order to explore the electronic structure of molecules. In other words, nuclei and electrons are explicitly treated. On the one hand, it results in a very accurate and robust method and produces a high level of detail of the phenomenon under investigation. However, it requires long calculation time with a considerable computational effort, especially for systems consisting of a large number of atoms, then limiting its applicability to specific cases, such as reaction mechanisms in which few atoms are involved. When biological systems consisting of hundreds to even thousands atoms are taken into account, the Molecular Mechanics (MM) turns out to be a more appropriate approach. It allows to evaluate the energy of a system as a function of the nuclear positions only, while electrons are implicitly considered through an adequate parameterization of the potential energy function.

2.2 MOLECULAR MECHANICS AND FORCE FIELD (FF)

Molecular Mechanics is regularly used when the biological system of interest contains a significant number of atoms, although some properties depending on electronic distribution are better treated with quantum mechanical approaches. The Born-Oppenheimer approximation is a fundamental principle that allows to consider the energy as function of nuclear coordinates only.[106] Nuclei that are heavier in terms of mass compared to electrons, are characterized by negligible velocity. It means that they are considered stationary with electrons moving around them. In the light of this phenomenon, the Born-Oppenheimer approximation states that the Schrödinger equation can be split into a part describing the electronic motions and the other regarding the motions of the nuclei, resulting in two independent wave functions. Molecular mechanics treats a molecule as a collection of masses interacting each other through harmonic forces. In other words, atoms can be simplified as hard spheres with radii obtained from experimental or theoretical data and net charges, while bonds can be represented as springs. In this way, many of the laws of classical mechanics can be applied for studying biological systems of any size in reasonable computational time. The different types of forces holding together all the atoms within the molecule are described by different terms of potential functions which, summed together, define the molecular potential energy, referred to as Force Field (FF). These terms are connected to changes in internal coordinates, that is bond lengths, angles and rotational or dihedral angles.

A relevant property of Force Fields is the transferability, that means the ability to use the same functional form and parameters to describe several systems, instead of defining specific functions for each single molecule. In this way, it is possible to use the Force Field to reproduce and predict structural properties of a wide range of molecules, although some specific systems require particular models. To this end, it is important to highlight that Force Fields are empirical, in other words a "perfect" form does not exist. Most of the available Force Fields are based on the same functional form, suggesting that a generic model is possible. Actually, this similarity is only due to evaluation of comparable interactions in a system. Especially when novel chemical classes are explored, the appropriate Force Field needs to be defined or, in alternative, chosen carefully. The choice of the most appropriate function depends also on the compromise between accuracy and computational efficiency. Force Fields express parameters in terms of atom types to define atoms of the system of interest. In particular, atom types describe hybridization states and neighboring environment, for example to distinguish atoms of histidine amino acid based on its protonation state or sp^3- from sp^2-hybridized carbon atoms.

Force Fields can be simplistically defined as the sum of four different terms including all the intra- and inter- molecular interactions within a system, as reported below:

$$V_{\mathrm{tot}}(r) = V_{\mathrm{stretching}} + V_{\mathrm{bending}} + V_{\mathrm{torsion}} + V_{\mathrm{non-bonded}} \tag{1}$$

$V_{\mathrm{tot}}(r)$ is the potential energy of a system as function of the atom positions (r). The first three contributes are relative to deviations of the internal coordinates from reference values and, then, refer to atoms directly bound (for this reason defined bonded terms). In details, the first and second terms (stretching and bending) are modeled as harmonic potentials and describe the energy when bond lengths and angles deviate from the reference values, whereas the third term is referred to changes in energy when rotations around bonds occur. The fourth contribute is referred to as non-bonded term and treats the interactions of non-bonded parts of the system. It is calculated for atoms separated by at least 3 bonds and usually concerns electrostatic interactions modeled using Coulomb potential and van der Waals interactions by a Lennard-Jones potential. Additional terms are sometimes included in sophisticated Force Fields.

In particular, the bond stretching and angle bending are calculated through Hooke’s law as follows:

$$V_{\mathrm{stretching}} = \frac{k}{2}(r - r_0)^2 \tag{2}$$

$$V_{\mathrm{bending}} = \frac{k}{2}(\theta - \theta_0\ )^2 \tag{3}$$

where k is the force constant and r_0 is the reference bond length for the first contribution, while k and θ_0 in Eq. (3) are the force constant and the reference angle. In consideration of the strong forces interacting between atoms, energy necessary to deviate a bond from its equilibrium value is high, resulting in large force constants for this contribute. On the contrary, force constants for the angle bending are lower because less energy is required to distort angles from their equilibrium position. In general, bond lengths and angles are defined “hard” degrees of freedom because high energy is required to deform them, compared to the remaining contributions. Structural modifications mainly regard torsional and non-bonded contributions.

The torsional potential expresses the rotational energy about a bond as reported below:

$$V_{\text{torsion}} = \sum_{n'=0}^{N} \frac{V_{n'}}{2} [1 + \cos(n\omega - \gamma)] \tag{4}$$

with ω that is the torsional angle, $V_{n'}$ is the barrier of rotation (the barrier height), n is the multiplicity, that is the number of minima in the function when bond rotates through 360° and finally, γ is the phase factor that indicates where the angle passes through its minimum value. Improper dihedral angles are also considered to select the correct geometry of particular atoms and an out-of-plane bending term is included in the Force Field equation.

As regards non-bonded terms, they are usually treated as function of the inverse power of the distance. The electrostatic contribution is defined through the Coulomb law:

$$V_{\text{C}} = \sum_{i=1}^{N_A} \sum_{j=1}^{N_B} \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \tag{5}$$

where N_A and N_B are the point charges in the molecules A and B, $1/4\pi\varepsilon_0$ is the Coulomb’s constant and r_{ij} is the distance between the two charges. These interactions involve pair atoms, resulting in what is known as pairs potentials that require long calculation time. Another limitation is related to the definition of partial charges to calculate the Coulomb energy that are not a directly observable property of an atom and not a property that can be measured directly by experiment. Then, it cannot be unambiguously defined and several methods are available to determine partial charges that reproduce desirable molecular properties. Finally, the van der Waals potential describes the dispersions and repulsions between atoms. One of the most used functions to describe van der Waals interactions is the Lennard-Jones 12-6 equation:

$$V_{\text{vdW}} = 4\varepsilon_{i,j} \left[\left(\frac{\sigma_{i,j}}{r_{i,j}} \right)^{12} - \left(\frac{\sigma_{i,j}}{r_{i,j}} \right)^{6} \right] \tag{6}$$

with ε and σ corresponding to the collision diameter and the well depth, respectively. The term powered to 12 describes the repulsive forces between the electronic distributions of two atoms that are close each other. The second contribute is the attractive long-range term that derives from London dispersion forces between induced dipoles for atoms with electron dispersion. Similarly to the electrostatic term, pair potentials are used to study the van der Waals interactions, leading to the evaluation of a huge number of interacting elements.

2.3 GENERATING CONFORMATIONAL ENSEMBLE

2.3.1 MOLECULAR DYNAMICS

Molecular Dynamics (MD) evaluates the "real" dynamics of a system through the integration of the Newton equation of motion for an ensemble of particles and results in a trajectory over a given time. MD is a deterministic method, that is predictions of subsequent states can be defined from the current state. At defined steps, forces acting on the atoms are calculated and new positions and velocities are defined. Through MD simulations, microscopic properties are investigated, although macroscopic features are of major interest. The relation between microscopic and macroscopic features is at the basis of the statistical mechanics, composed of basic principles necessary to understand the theory of Molecular Dynamics.

A microscopic state is identified through the positions and momenta of the N elements forming the system. These properties represent the coordinates in a 6N-dimension space (each element is associated with 3 position variables and 3 momentum variables), referred to as phase space. Therefore, the system is defined as a point in this multidimensional space evolving over time, resulting in a trajectory in the phase space. According to the principles of statistical mechanics, the instantaneous value of an observable A depends on the momenta ($\boldsymbol{p}^N$) and positions ($\boldsymbol{r}^N$) and can be reported as follows:

$$A(t) = A[\boldsymbol{p}^N(t), \boldsymbol{r}^N(t)] \tag{7}$$

Since this value fluctuates over time, due to the interactions between the particles, a time average (A_{ave}) of $A(t)$ is defined as the time increases to infinity:

$$A_{\mathrm{ave}} = \lim_{\tau \to \infty} \frac{1}{\tau} \int_{t=0}^{\tau} A(\boldsymbol{p}^N(t), \boldsymbol{r}^N(t)) \qquad (8)$$

From this equation, it is clear that to define average values of the observable *A*, it is necessary to describe dynamically the system in an infinite time. Obviously this is not feasible, especially if it is considered that macroscopic systems include a high number of atoms or molecules. To this end, the conventional approach of statistical mechanics considers a collection of systems (replicas) with the same macroscopic properties, instead of trajectories of a single system in the phase space. Each of these replicas is a point in the phase space and all together, they constitute what is known as statistical ensemble. Replicas progress in the phase space by keeping constant the thermodynamic properties which define the statistical ensemble. According to these constant properties, statistical ensembles can be classified in: canonical (NVT), microcanonical (NVE), isothermal-isobaric (NPT) and grand canonical (μVT). The thermodynamic properties of interest refer to number of particles (N), volume (V), energy (E), temperature (T), pressure (P) and chemical potential (μ).

Replicas are distributed based on a probability density, that is the probability to find a configuration with momenta $\boldsymbol{p}^N$ and position $\boldsymbol{r}^N$. It is computed in different ways depending on the statistical ensemble under investigation. For example, in canonical ensembles in which energy fluctuations are observed, the density distribution is defined by the Boltzmann distribution:

$$\rho(\boldsymbol{p}^N, \boldsymbol{r}^N) = \exp\left(-E(\boldsymbol{p}^N, \boldsymbol{r}^N)/k_B T\right)/Q \qquad (9)$$

where $E(\boldsymbol{p}^N,\boldsymbol{r}^N)$ is the energy, $1/k_BT$ is Boltzmann's constant with *T* standing for temperature and *Q* is the partition function. The latter, as reported in Eq. (10), is the sum of the Boltzmann factors from all microstates and is used in the previous formula as normalization factor.

$$Q = \sum_{\boldsymbol{p},\boldsymbol{r}} \exp\left(-E(\boldsymbol{p}^N, \boldsymbol{r}^N)/k_B T\right) \qquad (10)$$

It can be treated with a classic or quantum mechanical approach. In other words, it can be expressed as sum of all the Boltzmann factors (as shown in the previous formula) or as integral in the 6N dimensional space.

In the microcanonical ensemble, energy is constant and all microstates are equiprobable. All replicas explore each point in the phase space due to their constant energy and then, the time average coincides with an ensemble average. This equivalence, as reported in Eq. (11), is at the basis of the ergodic hypothesis which is a fundamental axiom of statistical mechanics.

$$\langle A \rangle = A_{ave} \tag{11}$$

where

$$\langle A \rangle = \frac{1}{M} \sum_{i=1}^{M} A(\boldsymbol{p}^N, \boldsymbol{r}^N) \tag{12}$$

with *M* corresponding to the number of time steps and angle brackets indicating an ensemble average. Molecular Dynamics takes advantage of this hypothesis through the integration of the ensemble average in order to obtain the thermodynamic averages.

The Hamiltonian (H) controls the evolution of a system over time. It can be calculated for a system of N particles as the sum of kinetic energy *K* which only depends on ***p***, and potential energy *V* only dependent on ***r***. This second part is computed through the Force Field which has been discussed in the previous section. On the other hand, the kinetic energy has a simple quadratic form as reported below:

$$K = \sum_{i=1}^{N} \frac{1}{2m_i} \left(p_{ix}^2 + p_{iy}^2 + p_{iz}^2 \right) \tag{13}$$

with m_i corresponding to the mass of particle *i* and p_{ix}, p_{iy}, and p_{iz} the *x*, *y* and *z* components of the momentum ***p***, respectively. In an MD simulation, the evolution of the system over the time, controlled by the Hamiltonian H, is defined through the integration of the equation of motion. The forces acting on the atoms combined with positions and velocities at a given time step (*t*) allow to define new positions and velocities at time interval $t + \Delta t$. Subsequently, atoms moves, forces are computed at the new positions and so on.

The Molecular Dynamics protocol consists of three main steps: minimization, equilibration and production. The starting structure of the system of interest and then, the resulting minimized geometry are the underlying configurations about which fluctuations occur during the simulation. Therefore, it is important to define a stable energy structure, that is a minimum on the potential energy surface to start the dynamics. For complex molecules, multiple minima are possible, as well as a global minimum may be identified, but in this case a conformational search would be needed. In order to reach the convergence, macromolecules are subject to large numbers of minimization iterations. Two examples of widely used minimization algorithms are the Steepest Descent and Conjugate Gradient. In the subsequent equilibration step, initial velocities are assigned to the atoms. Typically, they are randomly generated through a Maxwell-Boltzmann distribution. Random velocities can lead to a nonzero net momentum, resulting in possible translations or rotations of the system. Thus, in order to limit this effect, velocities are zero mean. Subsequently, the system is initially simulated to reach the equilibrium of the initial conditions. If the simulation is run at constant temperature, the system should be brought to the temperature of interest. During this step, in the light of the relation between temperature and velocity, this latter is rescaled until the desired temperature is reached. Once the properties of interest (V, T, P, E) show stable behavior in the equilibration step, the production can finally begin.

In an ideal system, we should consider lattice which is infinite in all dimensions. However, for computability and complexity reasons, some boundary conditions are necessary in order to simulate a part of an infinite system. As consequence, some molecules of the system are close to the edge of the sample and may be affected by surface effects which can be avoided by using Periodic Boundary Conditions. The box which includes the system of interest is replicated in different directions resulting in a periodic lattice of identical subunits. In this way, molecules located in proximity of the edge interact with the atoms of the neighboring box. In order to avoid that particles see their own periodic images, the minimum image convention is applied. According to this condition, each atom is surrounded by a box which is identical in size and shape to the periodic box and includes the remaining other atoms in the simulation. However, the calculation of the pairwise interactions of a particle with all of the other particles in the system may be too expensive in terms of computational cost. In order to limit this problem, the potential truncation is used, that is the introduction of a cutoff within which the non-bounded interactions are calculated. For consistency with the minimum image convention, the cutoff distance should be lower than half length of the box.

An important parameter in Molecular Dynamics is the time step, which defines how often the integration of the equations of motion is performed. Ideally, to reduce computational costs a large time step should be used. In practice, it is typically restricted to the femtosecond time scale. This value must be shorter than the period of the fastest motions in a system, like the oscillations of the hydrogen atoms. In order to reach a compromise between computational costs and accuracy, use of constraint methods, such as the SHAKE algorithm,[107] allows to fix bond lengths and then, increase the time step. The Verlet algorithm[108] is one of the most used finite difference methods, which approximate the positions and dynamics properties through Taylor series expansions. In particular, the Verlet algorithm determinates each position from the current position at time *t* and position at time *t* – Δt as follows:

$$r_i(t+\Delta t) = 2r_i(t) - r_i(t-\Delta t) + a_i(t)\Delta t^2 \tag{14}$$

In the context described so far, where the total energy of a system is conserved, the MD simulations explore the microcanonical ensemble (NVE). To simulate also the other statistical ensembles, like the isothermal-isobaric (NPT) or the canonical (NVT) ones, the equation of motion should be modified. In these further ensembles, velocities can be rescaled at certain steps so as to reach the desired target temperature, as already introduced earlier. In the isothermal-isobaric ensemble, constant pressure involves volume fluctuations. In this case, a target pressure is maintained by rescaling the simulation volume, through the use of a barostat. The Berendsen thermostat/barostat is widely used to weakly couple the system to an external heat bath with coupling constant for temperature and pressure.[109] The Langevin thermostat is another example that adds a friction term and a fluctuating force to Newton’s second law which is then integrated numerically.[110] Details of these algorithms can be found in the related reference papers.

Although Molecular Dynamics simulations are widely used and have been resulted in successful results to investigate biological systems, its utility is still limited because of two main reasons: *i*) the computational demand to adequately sample motions occurring in long timescales, and *ii*) refinements of Force Fields used to define the potential energy. To overcome the second limitation, methods combining molecular mechanics with quantum mechanics are used and successful results have been reported. Several solutions have also already been introduced to overcome the limits of the short time scales typically simulated. As of today, enhanced sampling methods, including umbrella sampling, methadynamics, accelerated dynamics, replica exchange molecular dynamics, etc. have allowed to observe protein shifts between conformations that would not be accessible given the time scales of conventional molecular dynamics.

2.3.2 METROPOLIS-MONTE CARLO METHOD

In contrast with Molecular Dynamics, Monte Carlo method is a stochastic sampling approach, where configurations are not connected in time and depend only on the previous sampled state.[111] New configurations are randomly sampled and accepted based on set of criteria. In particular, the probability to obtain a configuration is based on Boltzmann factor, as previously described. However, contrary to Molecular Dynamics where total energy includes the kinetic energy contribution, the total energy in Monte Carlo simulations is defined only as the potential energy. Configurations with low energy show a higher probability of acceptance than higher energy states. The average of the properties calculated for all the configurations M is reported below:

$$\langle A \rangle = \frac{1}{M}\sum_{i=1}^{M} A(\boldsymbol{r}^N) \qquad (15)$$

and clearly shows its dependence on $\boldsymbol{r}^N$ and not upon momentum contribution. Nicolas Metropolis made important contributions for the development of this method, which is known as Metropolis-Monte Carlo approach. After random moves of atoms or molecules of a system, the energy of the resulting configuration is higher of the previous state, a random number between 0 and 1 is generated and compared with the Boltzmann factor. If this number is higher than the Boltzmann factor, the novel state is rejected and a new move is attempted, otherwise it is accepted. A Metropolis approach defines a Markov chain of states which is characterized by “memoryless” properties, that is the outcome of each trial depends only on the preceding trial and not on the sequence of trials preceding it. This feature is known as Markovian property. Several Monte Carlo modifications have been introduced to focus on the exploration of the most important parts of phase space. For example, the preferential sampling favors random moves of molecules close to the solute than those far away. Cut off may be used to define regions subject to more frequent move attempts. As alternative, the probability to choose a molecule for random moves may be related to its distance from the solute. A further procedure, known as force bias Monte Carlo considers that each move is selected with higher probability in the direction of the instantaneous force on the particle than in other directions. Thus, these directed moves bring to a lowering of the overall potential energy and, consequently, to a higher acceptance probability than in the classic Metropolis protocol.[112] The traditional statistical ensemble of Monte Carlo method is the NVT ensemble, although sampling from the other ensembles is also possible.

2.3.3 ENHANCED SAMPLING METHODS

2.3.3.1 REPLICA EXCHANGE MOLECULAR DYNAMICS (REMD)

Replica exchange (or parallel tempering) is a sampling method based on *M* parallel and non-interacting replicas of the system of interest which are subject to simulations at different temperatures, typically in the canonical ensemble. At regular intervals, exchanges or swaps of configurations at different temperatures are attempted. This method takes advantage of simulations at high temperatures which allow to sample large volumes of phase space. On the other side, low temperature systems may be trapped in non-representative local energy minima. Therefore, configuration exchanges between lower and higher temperature replicas favor to escape from regions of the phase space in which configurations may be stuck.[113] The first replica exchange simulations trace back to Swedensen and Wang who introduced a replica Monte Carlo approach with a partial exchange of configurations at different temperatures.[114] Thereafter, a complete swap of replicas applied by Hansmann and co-workers to study biomolecules gave a large contribution to the development of the replica exchange method in Monte Carlo simulations.[115] Moreover, Sugita et al. introduced the Molecular Dynamics version of parallel tempering, also referred to as Replica Exchange Molecular Dynamics (REMD).[75]

If we consider a replica exchange simulation with M replicas, each of them at temperature T_i with $T_1 < T_2 < T_3 <\ldots$ and T_1 as temperature of the system of interest, the partition function of this ensemble is defined as:

$$Q = \prod_{i=1}^{M} \frac{K_i}{N!} \int d\boldsymbol{r}_i^N \exp[-\beta_i U(\boldsymbol{r}_i^N)] \tag{16}$$

where K_i is the kinetic energy of the system analytically calculated, $\beta_i = 1/(k_B T_i)$ and U is the potential energy. Usually swaps occur between configurations simulated at adjacent temperatures. The probability to accept the exchanges between replica i and j is given by:

$$\rho = \min\left\{1, \exp\left[\left(\beta_i - \beta_j\right)\left(U(\boldsymbol{r}_i^N) - U\left(\boldsymbol{r}_j^N\right)\right)\right]\right\} \tag{17}$$

In order to satisfy the detailed balance conditions in statistical mechanics, the exchanges must be attempted with a defined probability. The main difference between the Monte Carlo parallel tempering and REMD approach is that in the first case only positions of the atoms are taken into account, while the momenta are also considered in REMD simulations. In this last case, to conserve the kinetic energy between the replicas at different temperatures, new momenta needs to be determined as:

$$\boldsymbol{p}' = \sqrt{\frac{T_{\text{new}}}{T_{\text{old}}}}\boldsymbol{p} \tag{18}$$

where $\boldsymbol{p}$ is the old momentum for replica i and T_{new} and T_{old} are the temperatures before and after the exchange. This rescaling approach guarantees that the average kinetic energy is equal to $3/2Nk_bT$. It is worth pinpointing that replica exchange methods produce unphysical moves and not a real dynamics of a system. They are used with the aim of accelerate the sampling of the conformational space. Potential energy distribution is represented as Gaussian curve; therefore, the exchange probability can be represented as the overlap between two adjacent curves. The choice of number of replicas and temperatures at which systems are simulated requires considerable attention to ensure good overlap of energy histograms. Temperatures should be defined by guaranteeing exhaustive sampling of the phase space, while a right number of replicas is important to ensure a high exchange probability for all adjacent replicas. Several methods have been proposed to define temperatures and number of replicas to optimize successful configuration swaps. For instance, Kofke et al. have introduced a geometrical distribution of replicas in a defined range of temperature, where the ratio $T_{\text{old}}/T_{\text{new}}$ is kept constant.[116-117] Other studies have shown that a good performance of replica exchange simulations should correspond to an acceptance probability of around 20%.[118-119] As regards the number of replicas, it increases as $\sqrt{N}$, where N is the system size. To overcome this growth as function of the size of the system of interest, some methods attempt the exchanges by considering only part of the system. Many other variants of parallel tempering have been developed, such as methods that attempt swaps of alternative parameters. The Hamiltonian replica exchange is an example of this last category, where hydrophobic or van der Waals interactions between replicas are scaled and the acceptance probability takes into account the Hamiltonian of the swapped configurations, instead of potential energy as observed in the original version described before.

The main limitations of the replica exchange approach concern the high computational cost, due to the necessity to run multiple simulations in parallel. In addition, the set up of basic parameters to perform these simulations requires preliminary simulations, resulting in additional computational demand. Moreover, explicit solvent treatment increases the degrees of freedom to consider and decreases the width of Gaussian curves, resulting in a large number of replicas to reach an acceptable probability of exchange. Alternatives to overcome this limitation consist of implicit treatment of solvation effects that, however, only produce an approximation of the “real” behavior of a biological system. Another possibility is the explicit representation of water molecules in the first solvation shells, but, in this case, restraints on the solvent are necessary introducing artifacts because of the solvation surface.

2.3.3.2 hybrid REPLICA EXCHANGE MOLECULAR DYNAMICS (hREMD)

In the last years, parallel tempering has been combined with other simulation methods and, more in general, variants of the original approach have been introduced to deal with the limitations of REMD.[120-121] Hybrid Replica Exchange Molecular Dynamics (hREMD) combines explicit solvent Molecular Dynamics using standard methods such as periodic boundary conditions and inclusion of long-range electrostatic interactions, with implicit solvent models during the exchange attempts.[122] In particular, the swap probability is calculated by taking into account solvation shells defined on the fly during the fully solvated simulation and temporary replacing the remaining water with a continuum representation. After the exchange attempts, the original solvent coordinates are restored, and the simulation can proceed with explicit solvent treatment. In this way, the drastic decrease of degrees of freedom leads to a reasonable number of replicas to model large systems and accuracy of explicit solvent simulations. In fact, an advantage of this method is the full solvation during the entire Molecular Dynamics simulation; also, distribution functions and solvent properties are unaffected by the hybrid model during the exchange attempt. Moreover, in this approach, restraints of the water molecules are not necessary since the solvation shells are defined on the fly at every exchange calculation.

For the continuum solvent treatment, the generalized-Born (GB) model which represents an approximation of the Poisson-Boltzmann (PB) continuum electrostatic model, is widely used in the context of hybrid REMD simulations. Comparative studies by using hybrid and explicit REMD simulations have shown that hREMD is a sophisticated method with a good compromise between accuracy and calculation costs.

2.3.4 ELASTIC NETWORK-NORMAL MODE ANALYSIS

Normal Mode Analysis is a harmonic analysis starting with a minimization of the conformational potential energy as a function of the atomic Cartesian coordinates, followed by the calculation of the "Hessian" matrix, that is the second derivatives matrix of the potential energy with respect to the atomic coordinates. Finally the diagonalization of the Hessian matrix is computed and the eigenvectors (the "normal modes") and eigenvalues are defined.[93] The first most relevant eigenvectors are thought to describe the main protein motions in aqueous solution and are then used to collect significant protein conformations in perspective of a Virtual Screening.

In details, given a system of N atoms, the potential energy of a biomolecules is described as a function of its 3N coordinates through bonded and non-bonded energy terms. At a minimum, the potential energy can be expanded in a Taylor series in terms of mass-weighted coordinates, truncated at the quadratic level, with the first term set to zero:

$$V = \frac{1}{2}\sum_{i,j=1}^{3N} \left.\frac{\partial^2 V}{\partial q_i \partial q_j}\right|_0 q_i q_j \tag{19}$$

The second derivatives of this equation is written in the Hessian Matrix $\boldsymbol{F}$, of which the diagonalization determinates eigenvectors and eigenvalues:

$$\boldsymbol{F}\boldsymbol{w}_j = \lambda_j \boldsymbol{w}_j \tag{20}$$

with $\boldsymbol{w}_j$ and λ_j representing the jth eigenvector and eigenvalue, respectively. The 3N - 6 resulting eigenvectors, ranked according to their corresponding eigenvalues, specify the normal modes coordinates through:

$$Q_j = \sum_{i=1}^{3N} \boldsymbol{w}_{ij}\, q_i \tag{21}$$

while the eigenvalues describe the energy cost to deform the system along the pattern of atomic displacement defined by the eigenvectors. By considering that normal mode coordinates are subject to harmonic oscillations with angular frequency ω_j, the previous equation can be written as:

$$Q_j = A_j \cos(\omega_j t + \varepsilon_j) \tag{22}$$

where A_j is the amplitude and ε_j is the phase. Displacement of Cartesian coordinates along the eigenvector $\boldsymbol{w}_\mathrm{j}$ is then defined:

$$\Delta x_{ij} = \frac{\boldsymbol{w}_{ij}}{\sqrt{m_i}} A_j \cos(\omega_j t + \varepsilon_j) \tag{23}$$

Sometimes, Normal Mode Analysis may result to be computational demanding. In particular, the diagonalization of the 3N × 3N matrix may require long CPU time depending on the number of atoms N in the molecule.

In 1996, Tirion et al. proposed a new model based on Normal Mode Analysis, called Elastic Network Model (EN-NMA).[82] In this approach, interactions between two atoms are described through Hookean pairwise potentials and distances are considered at their minimum, so as to avoid the first step of Normal Mode Analysis, that is the energy minimization. In the case of Elastic Network, the potential energy function is computed as follows:

$$V = \frac{\gamma}{2} \sum_{|r_{ij}^0| < R_C} \left(r_{ij} - r_{ij}^0\right)^2 \tag{24}$$

In this equation, r_{ij} corresponds to the distance between atoms i and j, r_{ij}^0 is the distance of the reference structure, R_C is the distance cut-off and γ is the spring constant. Compared to Normal Mode Analysis, energy minimization is not required and the matrix diagonalization is faster, since it is commonly based on Cα atoms of the system. In fact, since Force Field is not used, Elastic Network model can be performed on a subset of atoms and not necessary on all elements of the system. This reduction results in a considerable gain in terms of computational cost spent to define the Hessian matrix. It has been widely demonstrated, through experimental and theoretical data, that low-frequency normal modes govern the large-scale conformational fluctuations.[123]

The EN-NMA, which analyzes fluctuations of protein backbone and side chains, can guide the Metropolis-Monte Carlo algorithm for the generation of Cartesian displacements, then resulting in novel conformations. In each iteration, the defined modes are displaced randomly, the energy of the resulting system is computed as reported in Eq. (24), and the Metropolis criteria is used. This leads to accepted displacements (at a temperature of 300 K) on each of the modes, which can be projected onto the Cartesian space. It is possible to obtain bigger displacements by increasing the temperature of the Monte Carlo procedure.[86]

2.4 ANALYSIS OF MULTIPLE CONFORMATIONS

2.4.1 CLUSTER ANALYSIS

With the development of sophisticated sampling methods, increasingly long trajectories consisting of large amount of data are collected. In the context of drug discovery, the use of all these stored structures for a Virtual Screening approach would be costly in terms of calculation time and irrelevant in terms of novelty, since these structures (in good part) may be conformationally redundant. Data-mining techniques represent powerful tools to group similar elements and gain useful information from the long trajectories. Among them, clustering is a general method, widely applied to any data collection, that allows to define set of points called clusters, joined according to a function measuring pairwise distance.[124] Thus, the elements of a cluster are ideally more similar to each other than to external elements belonging to other clusters. Several clustering algorithms have been developed and applied in different studies, like the analysis of Molecular Dynamics trajectories or, more in general, to group similar conformations. When used as tool to study configurations collected from sampling methods, cluster analysis allows to extract information about the visited substates of the conformational space. Since sampling follows the Boltzmann distribution, low energy substates are more populated than high energy substates, resulting in differently populated clusters.

The parameters that can critically affect the performance of the clustering algorithms are the selection of the atoms used to compare the initial data, the pairwise metrics for the comparison, and the cluster algorithm to group together similar elements. It is important to highlight that there are no perfect "one size fits all" algorithms and that the performance strongly depends on the clustering criteria.[92] In fact, some algorithms tend towards singleton or small clusters and others to homogeneously sized clusters. Therefore, particular care is necessary when cluster analysis is carried out and a visual inspection of resulting clusters is always recommended. Clustering methods are used in a variety of areas and several algorithms have been developed. In this section, a brief description of the most common algorithms applied in medicinal chemistry and, in particular, to study structural diversity of protein conformations collected from sampling approaches, are reported.

First of all, cluster algorithms can be classified as hierarchical and non-hierarchical. In the first case, clusters and sub-clusters can be identified resulting in a tree structure or dendrogram, as shown in Figure 5. Each cluster in this scheme (except for the leaves of the tree) is obtained by merging its children (sub-clusters), whereas the root represents the large ensemble containing all the initial elements. On the other side, a non-hierarchical method defines a classification by partitioning a dataset, resulting in a set of non-overlapping groups without hierarchical relationships between them. In a simplistic perspective, non-hierarchical clusters can be obtained by cutting any level of the hierarchical dendrogram (red box in Figure 5).

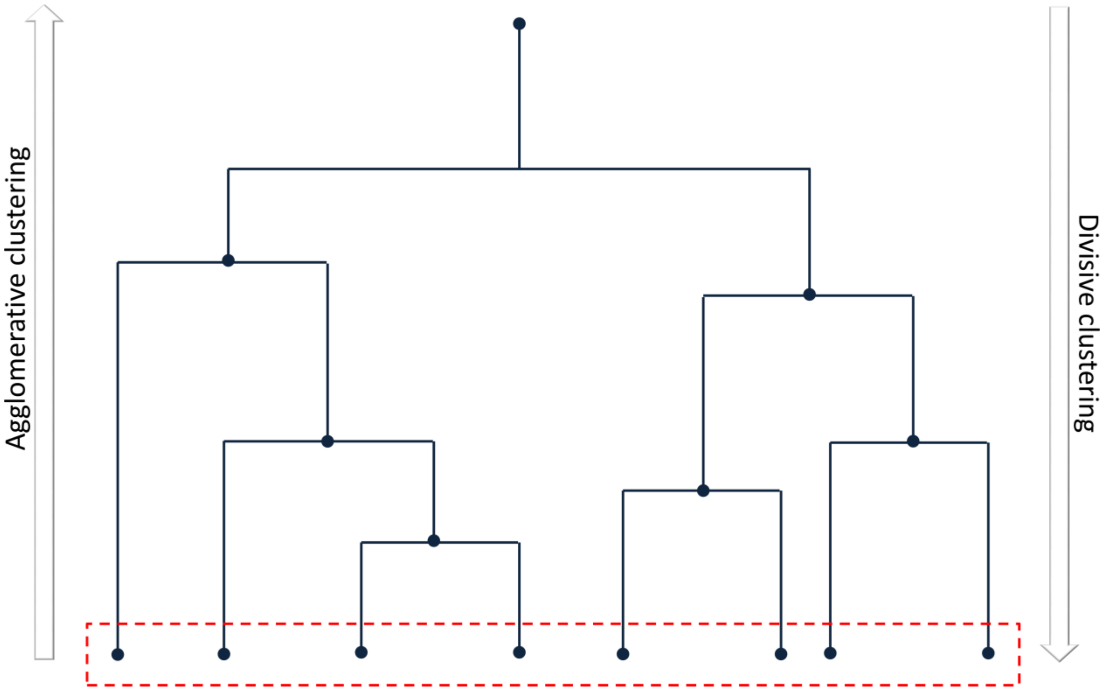


Figure 5. *Hierarchical dendrogram. Clusters are represented as blue dots. Red box indicates unrelated clusters, as result of a non-hierarchical algorithm. The two rows on the left and right of the dendrogram schematize the bottom-up (agglomerative) and top-down (divisive) clustering approaches.*

Also, hierarchical methods can be separated in divisive and agglomerative clustering approaches. In the first case, an initial large cluster including all the points is iteratively split in sub-clusters.[125] In principle, at each step, the largest distance between any two points is computed and the two extrema become the initial centroids for two new clusters. All points closest to one of the two centroids are assigned to this child cluster. This iterative procedure, also referred to as top-down method, is repeated until the desired number of clusters is reached. The opposite workflow which comes up when all starting elements are treated as singletons and combined together to give larger clusters, is known as agglomerative or bottom-up approach. Below, the most used hierarchical algorithms belonging to the latter category are described:

- single-linkage, the distance between clusters is defined as the shortest intercluster point-to-point distance;
- average-linkage, cluster-to-cluster distance is calculated as the average of all distances between each point of the two clusters;
- centroid-linkage, similar to single linkage but the intercluster distance is calculated between the cluster centroids;
- complete-linkage, cluster-to-cluster distance is identified as the largest point-to-point intercluster distance between two clusters.

Centripetal clustering and centripetal complete clustering are two further examples, based on a more sophisticated procedure, to make the algorithm less sensitive to outliers. K-means clustering is also widely used to group conformations from MD simulations and is an example of non-hierarchical algorithm.[126] In particular, it is a relocation method starting with an initial guess as to where the centers of clusters are located. Then, compounds are shifted between clusters to iteratively refine the centers, until stability is achieved.[127] In this case, *k* seed compounds are selected to act as initial cluster centers and each compound is linked to the closest seed. Then, centroids are recomputed and the procedure is repeated. Usually seeds are placed as much as far away from each other. Bayesian clustering and other algorithms are also relevant in computational chemistry and are described in details in several reviews.[92] Although it is not possible to select a priori the best performing algorithm to use for the study of interest, general comments about the features of resulting clusters can be outlined. In general, divisive methods tend to generate uniformly sized clusters, whereas agglomerative algorithms mostly result in a single large cluster. To assess clustering quality, some metrics have been introduced, aimed at evaluating two main measurement criteria: compactness (cluster members should be as close to each other as possible) and separation (clusters should be adequately separated).[128] The assessment clustering methods are divided in external, internal and relative criteria. The first two approaches are computer demanding and are based on user evaluations and specific metrics focusing on the data set, respectively. On the other side, relative criteria are useful parameters to establish the best performance between two clustering algorithms. An example is the Davies-Bouldin index (DBI) that is based on R_{ij} which is the similarity measure between clusters *i* and j.[129] Briefly, it is calculated as follows:

$$DBI = \frac{1}{n_c}\sum_{i=1}^{n_c} R_i \tag{25}$$

where

$$R_i = \max_{j=1\dots n_c, j\neq i}(R_{ij}) \tag{26}$$

This index measures the average of similarity between each cluster and its most similar one. Low DBI values correspond to a better clustering.

A method used to select the optimal number of clusters for a good result is the elbow criterion which take into account the variance as a function of the number of clusters.[130] According to this approach, one more cluster is added to the chosen number of clusters, and percentage of variance is evaluated. In general, a plot of percentage of variance versus the number of clusters shows that the first cluster adds the majority of variance, and after a certain number of clusters the marginal gain decreases, giving an angle in the plot. The number of clusters are usually chosen at this point.[131] Although cluster analysis has been reported as an accurate method to group and analyze conformations from sampling methods, it presents some weaknesses that limit their universal use. For example, application of several algorithms on the same data set has shown widely different results, highlighting the limitations of all single methods. In particular, hierarchical algorithms are highly sensitive to outliers, mean algorithms lead to homogeneous sized clusters and average and centroid linkage algorithms tend to generate singletons.[92]

2.4.2 DIMENSIONALITY REDUCTION

In the field of Molecular Dynamics or, in general, sampling methods, it is difficult to study large amount of data collected from long trajectories, through cluster analysis. As briefly commented in the previous section, clustering algorithms show low stability issues. In addition, this method does not provide a clear interpretation of the phenomenon under investigation linked, for example, to global protein flexibility. A valid alternative is represented by methods aimed at reduce the dimensionality of a data set, that is the number of variables required to describe the system of interest. Principal Component Analysis (PCA)[94] and Multidimensional Scaling (MDS)[132-133] are two techniques, widely used for this purpose. PCA is a way to express a high dimensional data set through a set of dimensions (principal components) that are orthogonal to each other. When data are plotted in the *x*-dimensional space, where *x* corresponds to the principal components (PCs), the first component maximizes the variance within the data set; in other words, the data show the greatest spread of values along this component. The subsequent principal components take into account the maximum variance not considered by the first variable. In general, a few principal components are sufficient to describe salient features of a system.

In details, PCs are computed through two main steps: *i*) the calculation of the covariance matrix, $\boldsymbol{C}$, of the positional deviations, and *ii*) the diagonalization of this matrix. For a covariance matrix computed on an ensemble of protein structures, the elements of $\boldsymbol{C}$ are defined as:

$$c_{ij} = \langle (\mathrm{x_i} - \langle \mathrm{x_i} \rangle)(\mathrm{x_j} - \langle \mathrm{x_j} \rangle) \rangle \tag{27}$$

where x_i and x_j represent the atomic coordinates and the brackets correspond to the ensemble average. The diagonalization of the symmetric matrix $\boldsymbol{C}$ results in:

$$\mathbf{A}^{\mathrm{T}}\boldsymbol{C}\mathbf{A} = \lambda \tag{28}$$

where $\mathbf{A}$ and λ represent the eigenvectors and the associated eigenvalues, respectively. This procedure thus transforms the original matrix in a new orthornormal matrix composed of the eigenvectors. The first principal component shows the highest eigenvalue. MDS is another possible method to reduce dimensionality, based on dissimilarity/similarity data between pairs of elements. The main goal of MDS is to represent these dissimilarities/similarities as distances between points in a low dimensional space such that the distances represent as closely as possible the dissimilarities. With MDS, analysis of any kind of similarity or dissimilarity matrix is possible, in addition to correlation matrices. In technical terms, MDS is based on a function minimization algorithm that evaluates different configurations with the aim to maximize the goodness-of-fit (or minimize "lack of fit").

2.4.3 NETWORK ANALYSIS

Further methods used to analyze and study the diversity within structural ensemble include network analysis. The advantage of this approach lies in the possibility to visualize topological correlations between the elements otherwise difficult to be seen by classical cluster analysis. The elements of a network are defined as nodes, which are bound together by links or edges representing interactions between element pairs.[134] The network architecture based on the disposition of nodes and links defines a layout. According to the nature of interactions, networks are classified as directed or undirected. In the first case, nodes are linked with a well defined direction, whereas no directions are observed in undirected graphs. The degree or connectivity of a node (k) is a basic parameter identifying the number of links of each node with the others, while the degree distribution, *P(k)*, defines the probability that a specific node has exactly k edges. This distribution is calculated by dividing the number of nodes with k links by the total number of nodes, as reported in the following formula:

$$P(k) = \frac{n_k}{n} \tag{29}$$

Other important metrics in network analysis include the shortest path and mean path lengths, measuring the path with the smallest number of edges between the selected nodes and the average over the shortest paths between all pairs of nodes, respectively. Finally, the clustering coefficient is computed as:

$$C_I = \frac{2n_I}{k_I(k_I - 1)} \tag{30}$$

where n_I defines the number of edges between the neighbours of node *I*, and $k_I(k_I - 1)/2$ is the total possible number of these links. In other words, C_I refers to the "triangles" that go through node I and can give insights of how densely clustered the edges in a network are. *C(k)* is the average clustering coefficient of nodes of degree *k* in the network. Different forms of the parameters described above characterize three types of networks: *i*) random, *ii*) scale-free and *iii*) hierarchical networks (Figure 6).

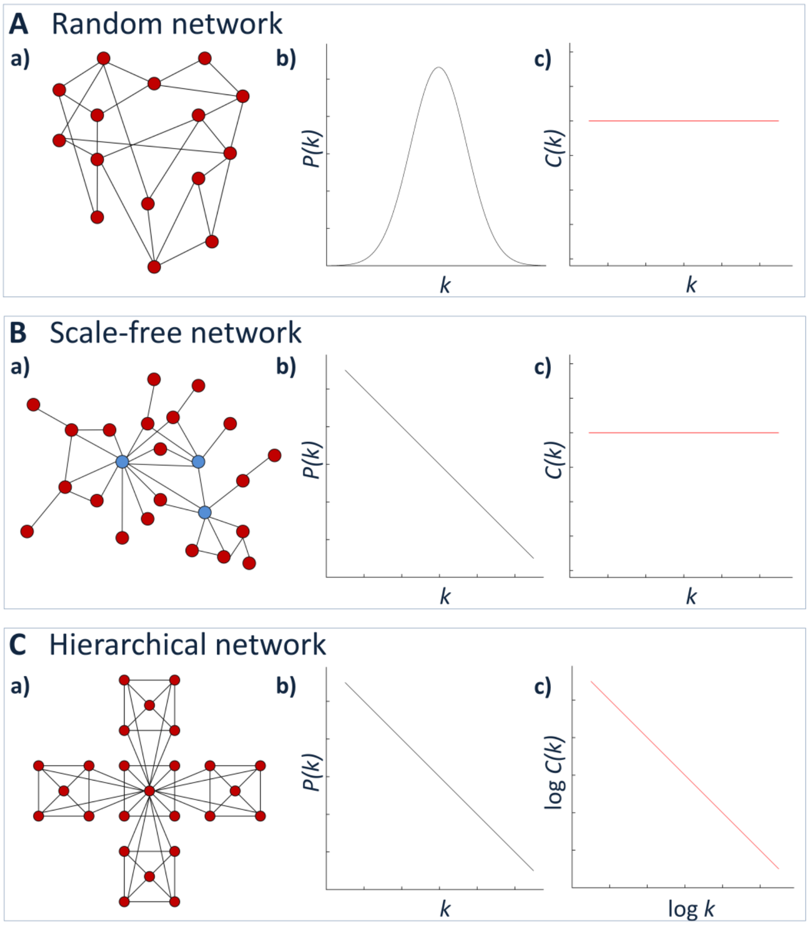


Figure 6. *Layout of random (A), scale-free (B) and hierarchical network (C). Degree distribution and average clustering coefficient are reported in the b) and c) panels, respectively.*

Random networks (Figure 6A) are characterized by N nodes connected with probability p, resulting in graphs with approximately $pN(N-1)/2$ links distributed in a random way.[135] Due to this random positioning of edges, most nodes show approximately the same degree, close to the average degree $\langle k \rangle$ of the network. Indeed, the node degree follows a Poisson distribution (Figure 6Ab) with a peak at $\langle k \rangle$. From this plot, it is clear that nodes which diverge from this average value are rare. The average clustering coefficient as function of k is a horizontal line, that is, the clustering coefficient is independent of nodes connectivity. Moreover, random networks are characterized by small-word property: the average path lengths are small compared to network size *n* and, more specifically, are of the same order of magnitude as $\log(n)$. This means that only a small number of steps are necessary to reach most nodes from every other. Networks characterized by a power-law degree distribution, as reported in Eq. (31), are called scale-free (Figure 6B):

$$P(k) \sim k^{-\gamma} \tag{31}$$

where γ is the degree exponent. According to the power-law distribution, this network includes a large number of nodes with only a few links and a few nodes, known as hubs, with numerous edges. Also in this case, average clustering coefficient is independent of *k*. For scale-free networks, the average path length is significantly smaller (ultra-small property) than in random networks, and scales as $l \sim \log(\log N)$, due to the presence of hubs. The last category of networks, hierarchical networks, is characterized by modules that are replicated to form clusters connected to a central nodes (Figure 6 C). Therefore, a few central hubs are created, resulting in a network with a power-law degree distribution. In this case, the clustering coefficient of a node with *k* links follows the scaling law:

$$C(k) \sim k^{-1} \tag{32}$$

that quantifies the coexistence of a hierarchy of nodes with different degrees of clustering. In order to test the robustness of networks, several tests are available, such as the accidental node failure. Removal of nodes from random networks results in a disintegrated layout, whereas scale-free networks are more robust to these changes, unless their hubs are attached.

In the context of drug discovery, network analysis can be used to analyze conformations resulted from sampling methods, such as Molecular Dynamics. In this case, nodes are the explored conformations and links describe the similarity of nodes, for example in terms of RMSD. Nodes can be weighted or not, based on the importance of conformations and the type of information to extract from the simulations. For instance, the free energy surface can be mapped into a weighted network where nodes and links are configurations and direct transitions among them, respectively.[136] However, when network analysis is used with the only purpose to analyze large amount of data collected from long simulations as alternative to cluster analysis, all nodes are similarly considered and unweighted networks are used. Several layout algorithms have been developed aimed at positioning the nodes and edges for the network. They include grid layout, circular layout, hierarchical layout, data-driven simple layouts that reflect some data property of the nodes themselves and force-directed layout algorithms. The latter is based on some kind of physical simulations and consider the nodes as physical objects and the links as springs connecting those objects together. It will be discussed in details in next chapter.

2.5 LIGAND DOCKING AND SCORING FUNCTIONS

Ligand docking is a key computational method in drug discovery aimed at predicting conformation and orientation of small molecules that interact with protein structures. It consists of two main steps:

- Posing: configurational space sampling to accommodate ligands in the active site and find putative binding modes.
- Scoring: evaluation of each pose based on the fit of ligands to the receptor.

Small molecules can contain a high number of degrees of freedom, so as to increase the size of the configurational space to be sampled. Therefore, definition of binding poses is a challenging step requiring accuracy to define the best configurations and, at the same time, reasonable computational time to screen large databases in a docking run. As regards the second step, some docking software use multiple scoring stages: first, they rank binding modes according to approximate evaluations of shape and electrostatic complementarity and then, they re-score resulting binding modes and estimate the binding affinity by means of more sophisticated treatment of electrostatic, van der Waals, solvation or entropic effects.[137]

Search algorithms for ligand flexibility and pose prediction fall into two major categories: systematic and stochastic methods. Systematic conformational search is deterministic and samples all the degrees of freedom in a molecule at regular intervals, resulting in the problem of combinatorial explosion.[138] Because of this limitation, systematic methods are mainly employed for rigid docking, or replaced, for example, with the incremental construction approach consisting of a core fragment selection, core fragment placement and incremental ligand construction. In details, ligands are divided into a rigid core and flexible side chains which are, in turn, divided based on their rotatable bonds. Then, core fragments are docked followed by the side chains which have been sampled with the systematic search. On the other side, stochastic algorithms make random changes that are accepted or not based on a probability function. This category includes Monte Carlo[111] and genetic algorithms.[139] According to the former, random conformations, translations and rotations are sampled for each ligand, configurations within the binding site are generated and scored. If the new state shows a lower energy, this move is accepted, otherwise is retained based on Metropolis criterion. This procedure is repeated until the defined number of configurations is obtained. Genetic algorithms are based on the principles of biological competition and population dynamics. Individuals represent binding modes and are encoded in a chromosome composed of genes, that is the degrees of freedom. Individuals are subject to genetic operations, such as crossover between two parent chromosomes to generate two new offspring. The next generation can also be produced through mutations by which a gene is selected and randomly changed. Resulting chromosomes are evaluated by a fitness function. Stochastic search ends when a sufficient number of generations or convergence of the docking results on each binding mode are achieved.

As regards the scoring functions, they are classified in four different categories: *i*) force-field-based, *ii*) empirical, *iii*) knowledge-based and *iv*) consensus scoring. The force-field scoring function is based on physical atomic interactions, such as van der Waals and electrostatic interactions.[140] The van der Waals term is calculated by a Lennard-Jones potential function and can be more or less permissive to close contacts, depending on the selected parameters. Electrostatic interactions are computed by the Coulombic term which includes the distance-dependent dielectric function; this factor understates the desolvation effect and consequently scoring functions may be biased on Columbic electrostatic interactions favoring charged ligands. A generic force-field scoring function typically shows a low ability to properly assess solvation and also entropic effects. Another limitation is the difficulty in combining individual energy terms that are calculated from diverse methods with diverse scales. Then, weighted coefficients are necessary and are obtained by fitting experimental data. However, a complete set of these factors for all protein-ligand complexes is not available.

Empirical scoring functions estimate experimental data, like binding affinity based on a set of weighted energy terms:

$$\Delta G = \sum_i W_i \cdot \Delta G_i \tag{33}$$

where ΔG_i corresponds to the energy terms and W_i represents the coefficient obtained from regression analysis with experimental data. Compared to the force-field based scoring functions, they seem to be faster in computing scores thanks to their simple terms. However these methods depend on the molecular data sets for the regression analysis. As regards knowledge-based scoring functions, they aim at reproducing experimental structures instead of binding energies and employ atomic interaction-pair potentials. They are accurate, robust and fast scoring functions permitting efficient screening of large datasets. Finally, consensus scoring combines advantages of different scores to limit weakness of single scoring functions and improve performance in ligand docking.

The main applications of molecular docking concern: *i*) the determination of the binding mode of ligands, *ii*) the prediction of binding affinity which is particularly relevant in lead optimization and *iii*) the identification of novel potential hit compounds, as successful results of a Virtual Screening approach. Different criteria are used to evaluate the performance of a scoring function. First of all, the ability to discriminate native binding modes from decoys is measured through distance metrics like RMSD between the best ranked conformations and native structures. A successful prediction results in an RMSD value less than or equal than 2.0 Å. However, this simplistic criteria to compare docking results may be misleading in some cases. To evaluate the performance of a docking study in terms of binding affinity of a complex, several metrics can be used, while figures of merit assess a scoring function for its ability to early select potential hits from a Virtual Screening. These parameters are extensively discussed in Section 2.6.

A large number of docking software and scoring functions have been published so far. Among them, GLIDE (Grid based Ligand Docking with Energy)[141-142] and

ICM (Internal Coordinate Mechanics)[143] will be described in details in next paragraphs.

2.5.1 GLIDE: DOCKING AND SCORING FUNCTION

GLIDE is a widely used docking software based on an exhaustive systematic search of the positional, orientational and conformational space of docked ligands. Taking into account the main goals of a docking study, this method shows a satisfying performance in terms of computational cost, robustness of binding mode prediction, accuracy of binding affinity prediction and Virtual Screening results.[141-142] An ensemble of hierarchical filters are used to investigate the putative locations of ligands within the receptor active-site: *i*) ligand conformations are generated; *ii*) an initial screening of ligand poses is performed; *iii*) resulting poses are minimized using a molecular mechanics scoring function; *iv*) the best ones are subjected to a Monte Carlo procedure to explore torsional minima and then *v*) they are rescored using Glide SP (Standard-Precision) and/or Glide XP (Extra-Precision) empirical scoring functions. A set of fields on a grid describes shape and properties of the receptor and allows to define an accurate scoring of ligand poses. As first step of this docking protocol, ligand flexibility is treated through a systematic conformational search along with a heuristic screen allowing the elimination of unsuitable high-energy conformations. Each ligand includes a core region and rotamer groups. Multiple conformations are generated for each core, depending on the number of rotatable bonds, conformationally labile five- and six-membered rings, and asymmetric pyramidal trigonal nitrogen centers, while rotamer states are numerated for each rotamer group. Then, they are docked in a preliminary docking run. For each core conformation, all possible positions and orientations within the active site are explored. “Site points” are defined and selected on an equally space 2 Å grid over the active site. At the same time, ligand centers and diameters are identified. Subsequently, distances between the site points and receptor are compared with those between ligand centers and ligand surfaces, and in case of good match, ligands are positioned at those site points. These placements are examined and, if too many clashes are observed with the protein, they are skipped. Next step consists of rotation of ligands around its diameter and scoring of atoms which favorably interact with protein. If this score is good enough, all interactions are scored. A discretized version of ChemScore[144] is used to evaluate these favorable interactions, in particular hydrophobic, hydrogen-bonds, metal-ligation interactions and steric clashes. This step is known as “greedy scoring”, since the score for each atom is dependent not only on its position against the protein but also on the best possible score after movements of 1 Å in the Cartesian directions. The best poses are first minimized on OPLS-AA “smoothed” van der Waals and electrostatic receptor grids and then, on the full-scale OPLS-AA nonbonded energy surface (annealing). In this way, other possible favorable core conformations and rotamer-group torsion angles are sampled, in order to improve the score. A final score of the docked poses is defined by using GlideScore, existing in two different forms: Glide SP and Glide XP. In principle, these two functions are based on similar terms, but Glide SP is a more forgiving function, allowing a reasonable score for imperfect ligand poses, aimed at reduce false negatives. In contrast, Glide XP is a harder function with strong penalties minimizing false positives.

Glide SP scoring function is based on ChemScore function and described as follows:

$$\begin{gathered}\Delta G_{bind} = \Delta E_{Phob} + \Delta E_{Hb_{neut-neut}} + \Delta E_{Hb_{neut-charged}} + \\ \Delta E_{Hb_{charged-charged}} + \Delta E_{met} + \Delta E_{rotb} + \Delta E_{polar-lip} + \\ \Delta E_{C} + \Delta E_{vdW} + \text{solvation terms}\end{gathered} \tag{34}$$

The first term corresponds to the hydrophobic interactions and is the same as in ChemScore. The hydrogen bonding term is divided into three categories: neutral-neutral, neutral-charged and charged-charged. The first contribute has the highest influence on the final score than the charged-charged term. Then, the metal-ligand interaction term is analogue to the ChemScore one, with some variations. The subsequent two terms concerns the rotatable bonds and energy contributes occurring when polar atoms different form hydrogen-bonding atoms are located in hydrophobic regions. The Coulomb and van der Waals interaction energies are included in Glide SP scoring function. Moreover, the last term takes into account solvation contribution to binding energy. To define this effect, Glide performs a docking of ligands with explicit water molecules and employs empirical scoring terms to measure the exposure of functional groups to the solvent. This water-scoring approach has been made efficient by the use of grid-based algorithms.[141] In order to dock a large database in a reasonable computational time with Glide SP, small coefficients with soft penalties are included. However, the docking performance is still satisfying in terms of enrichment factors for several virtual screening protocols.

As regards Glide XP, the sampling method starts with the identification of some fragments of initial ligands (usually derived from Glide SP docking) as “anchors”, typically rings. Clustering of multiple anchors orientations and selection of the most representative members are followed by the growing step, that is the study of flexibility relating to side chains directly bound to the initial fragments. Bond by bond, molecules are grown, minimized and the best ligand poses are score with Glide XP scoring function, including some penalties. If side chains are wrongly oriented resulting in high penalties, growth strategy is repeated and further poses are defined.

The novel features of Glide XP compared to Glide SP concern the use of large desolvation penalties to ligand and protein polar and charge groups and the identification of specific motifs important to recognize particular interactions enhancing binding affinity. The explicitly-water technology is analogue to the SP procedure, but higher penalties to violations of physical principles are used. For example, a desolvation penalty is assigned if polar or charged functional groups are improperly solvated or water molecules interacts with hydrophobic contacts higher than a defined cutoff. Contact penalty is also included in the Glide XP scoring function, considering high strain energy of docking poses. Moreover, this function includes some terms contributing to improve the binding affinity prediction, such as the hydrophobic enclosure model that favors the recognition of hydrophobic ligand atoms surrounded on both faces by lipophilic receptor atoms. Improvements include also the treatment of hydrogen bonds and identification of π-cation or π-π stacking interactions.

2.5.2 ICM: DOCKING AND SCORING FUNCTION

ICM flexible ligand docking is based on Monte Carlo simulations that relies on global optimization of the entire flexible ligand in the receptor field. In particular, proteins are considered rigid and represented as grid potential maps.

The ICM docking algorithm consists of 4 different steps:

1. a random change of one or multiple variables of the ligand in the active site is introduced. Two different random moves are possible: a) a positional Pseudo-Brownian or b) an internal torsional modification. In the first case, the whole ligand is moved or also rotated around its center of gravity, while internal torsional angles are randomly changed one a time;
2. a local minimization of differential energy terms is performed. The conformations resulted from step 1 are optimized through conjugate gradient method, without including the surface-based solvation energy;
3. desolvation energy is computed;
4. Metropolis selection criterion is applied to accept or reject the new conformation.

This procedure is repeated for a defined number of steps. Each molecule is first subject to a stochastic conformational search and then, the stack of low energy conformations are used for docking. Receptor grid maps are pre-calculated and represent hydrogen bonding, van der Waals, hydrophobic and electrostatic potentials.

Ligand poses are scored with the ICM scoring function, defined as follows:

$$\begin{aligned}\Delta \mathrm{G} = \Delta E_{IntFF} + T\Delta S_{Tor} + \alpha_1 \Delta E_{HBond} + \alpha_2 \Delta E_{HbDesol} \\ + \alpha_3 \Delta E_{SolEl} + \alpha_4 \Delta E_{HPhob} + \alpha_5 \Delta Q_{Size}\end{aligned} \tag{35}$$

It is an empirical scoring function, based on physico-chemical properties calculated for receptor-ligand complex. The first term considers describes the internal force field energy of the ligand, including internal van der Waals interactions and torsion energy for the ligand. The basic energy function used in this program is ECEPP/3.[145] A smoothed van der Waals potential is used to reduce noise in the energy function. The second term of the scoring function describes free energy variations related to conformational energy loss upon ligand binding. The remaining part of the function includes weighted terms related to the interaction energy with the five grid potential maps of the proteins subtracted with the energy for the free ligand in solution. Coefficients have been defined by fitting the resulting scoring functions to a training set of protein-ligand complexes. In details these terms describe hydrogen bonds, hydrogen bond donor-acceptor desolvation energy, solvation electrostatic energy upon ligand binding, hydrophobic free energy and a size correction term depending on the number of ligand atoms.[146] The Potential of Mean Force (PMF) is a knowledge based scoring function also available in ICM software. It is based on experimental distance distributions of atom types belonging to protein-ligand complexes extracted from the Protein Data Bank. The score is defined through the atom pair distances of the docked ligands within receptors.

2.6 VALIDATION OF DOCKING RESULTS

Structure-based Virtual Screening consists of docking and scoring of ligands into single or multiple proteins and estimation of ligand binding affinity. The main goal of Virtual Screening is to extract a high number of potential actives from the first percentage of the rank-ordered set of compounds as obtained from ligand docking. In other words, a successful Virtual Screening should rank actives early in a database including also inactives/decoys, since only a minimal set of ligands will be selected for experimental assays. This ability is referred as “early recognition” and will be often recalled in this section. In general, both good separation between actives and decoys and early recognition of actives in a rank-ordered list are important; in fact, a Virtual Screening algorithm turns out to be useless when it selects a large number of actives but ranks them at random positions.

2.6.1 FIGURES OF MERIT

Several metrics are used to evaluate the effectiveness of ranking methods in Virtual Screening and new descriptors aim at treat the early recognition problem. More in general, they can be useful in two contexts: *i*) to analyze the ability to select known active chemicals and discard inactive compounds, in an assessment study or retrospective Virtual Screening, and *ii*) to define a threshold between molecules which should be selected for biological tests and others that should be discarded as likely inactives, in a drug discovery campaign.[147] Most of these metrics are based on two parameters: Sensitivity and Specificity, also referred to as True Positive rate and False Positive rate, respectively. When, in a Virtual Screening, *k* molecules are selected from an initial database including N molecules, True Positive (TP) and False Positive (FP) compounds are the actives and the decoys belonging to the hit list of the selected entries, respectively. The remaining part of the database includes

False Negatives (FN) that are the actives not selected by Virtual Screening, and the unselected inactives representing the True Negatives (TN).[147] Figure 7 below schematizes this classification.

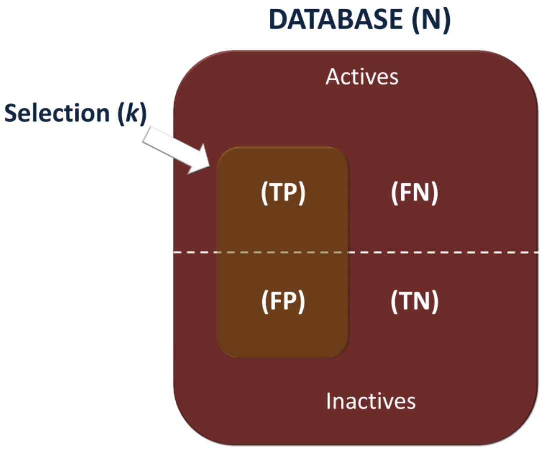


Figure 7. *Schematic representation of TP, TN, FP, FN definition, according to a Virtual Screening selection of k molecules from a database consisting of N compounds.*

Sensitivity is then defined as the ratio of the selected actives and the total number of actives in the database, as follows:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{36}$$

while Specificity describes the ratio of the discarded inactives to all inactives in the database:

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{37}$$

The Receiver Operating Characteristic (ROC) curve is a plot describing Sensitivity for any change of *k* as function of 1 – Specificity reported below:

$$1 - \text{Specificity} = \frac{\text{FP}}{\text{TN} + \text{FP}} \tag{38}$$

In order to use the ROC curve method to assess the performance of a retrospective Virtual Screening, three initial steps are required: *i*) definition of a pharmacological activity cutoff to select known active and inactive compounds, *ii*) definition of a small data set of these molecules and *iii*) Virtual Screening of the sample.

Subsequently, ROC curve is built by defining a first selection threshold (*k*) corresponding to the rank of the first selected molecules in the rank-ordered list. At this threshold, the number of actives and inactives is calculated and, accordingly, Sensitivity and Specificity are computed as well. This procedure is repeated until the selection threshold covers all active chemicals. The ideal condition in which all active compounds are early ranked in respect to inactives leads to an increase along the y-axis, from the origin to the upper-left corner of the ROC curve. After that, all the inactives are retrieved moving the ROC curve as straight line towards the upper-right corner identified by Sensitivity = 1 and 1 – Specificity = 1 (red plot in Figure 8a). On the other side, when actives are randomly distributed in the rank-ordered list, the ROC curve is represented as bisecting line with Sensitivity = 1 – Specificity. Finally, in the most common scenario, ROC curve has an intermediate trend between the ideal and the random case, as shown with the blue plot in Figure 8a. If actives and inactives in a rank-ordered list are represented as two different distributions, the ideal ROC curve is associated to a clear separation between these two distributions, whereas in the more realistic case, an overlap occurs which sheds light on False Positives and False Negatives (see Figure 8b).

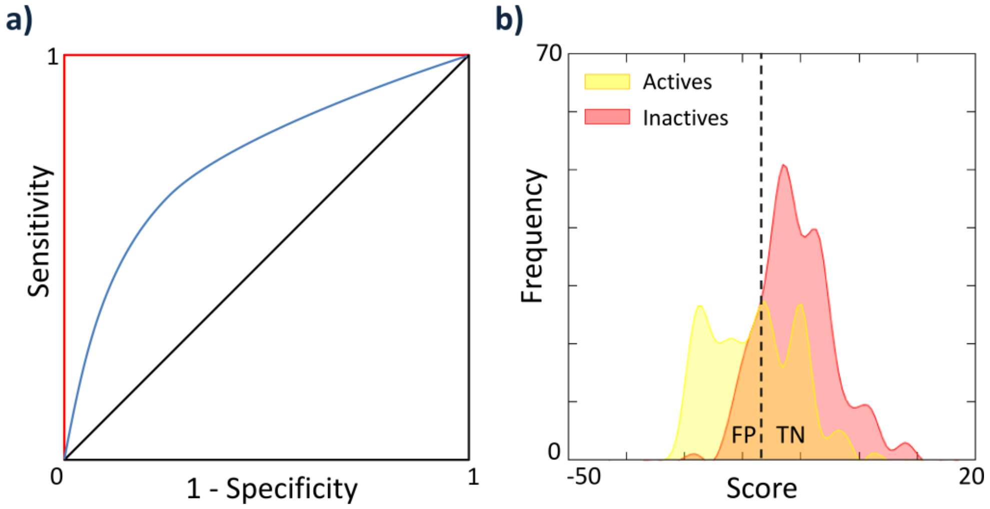
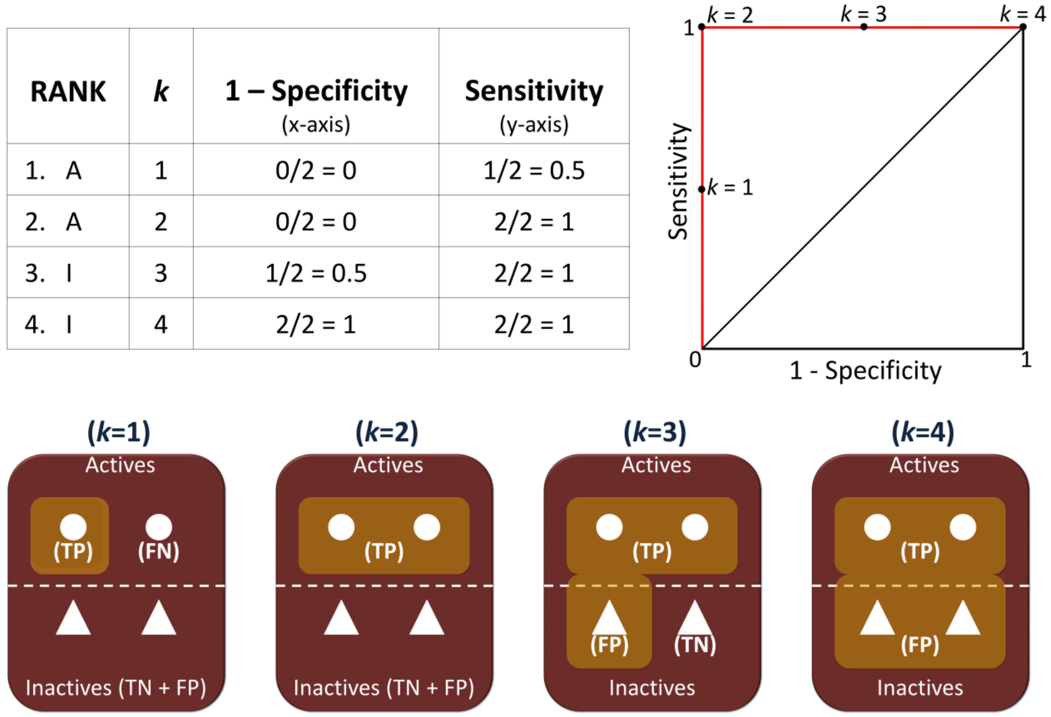


Figure 8. *a) ROC curve for the ideal scenario (red), random distribution (bisecting line) and real situation (blue). b) Theoretical distributions for actives (yellow) and inactives (red) and identification of False Positives and False Negatives based on the selection threshold (dashed line).*

To better understand the ROC curve method, we describe a simple example in which 2 actives among 4 molecules (*N*) are found at rank 1 and 2. Figure 9 shows how the ROC curve is built for any change of the selection threshold *k*, according to the number of True Positives, False Positives, True Negatives and False Negatives found.



RANK	*k*	1 – Specificity (x-axis)	Sensitivity (y-axis)
1. A	1	0/2 = 0	1/2 = 0.5
2. A	2	0/2 = 0	2/2 = 1
3. I	3	1/2 = 0.5	2/2 = 1
4. I	4	2/2 = 1	2/2 = 1

Figure 9. *A and I in the table represent actives and inactives, respectively. After a Virtual Screening, the two actives are found at rank 1 and 2. Based on the selection threshold k, Sensitivity and 1 – Specificity can be defined, so as to build the ROC curve. When all actives are found among the selected compounds (k = 2), False Negatives are equal to zero. Then, all inactives are found moving the ROC curve towards the upper-left zone of the plot.*

ROC curves represent a very intuitive mean to predict Virtual Screening performance. Closely related, the area under the ROC curve (AUC) is a parameter widely used that describes the probability of actives to be early ranked compared to inactives.[148]

In general, the cumulative distribution function *F(x)* is the probability that a random variable has a value between 0 and *x* and is simply the integral of the probability distribution function *f(x)*. In the Virtual Screening context, *x* is the normalized rank in an ordered list resulting by division of the rank of a compound by the total number of compounds. The cumulative distribution function is referred as accumulation curve and indicates the number of actives found at a rank position.

The discrete formula of AUC is defined as follows:

$$AUC = \frac{1}{nN}\sum_{k=2}^{N} F_a(k)\,[F_i(k) - F_i(k-1)] \tag{39}$$

where $F_a(k)$ and $F_i(k)$ are respectively the accumulation curves for actives and inactives and correspond to Sensitivity and 1 – Specificity. In a simplistic way, AUC can be seen as the sum of all rectangles formed by Sensitivity and 1 – Specificity coordinates for the different thresholds.[149]

The continuous definition of AUC is given through the following formula:

$$AUC = \int_0^1 F_a(x) f_i(x)\,\mathrm{dx} \tag{40}$$

with $f_i(x)$ that is the probability distribution function of the inactives. AUC value of 1 and 0.5 correspond respectively to the ideal and random distribution of actives and inactives, while a scenario in which Virtual Screening performs better than random picking, entails AUC values between 0.5 and 1. Although AUC is widely used in several fields and very useful for preliminary evaluations of Virtual Screening performance, recent works have shown some limitations. In particular, the most relevant drawback is the low Sensitivity to the early recognition problem. Figure10 highlights the limited ability of AUC metric to define the best Virtual Screening protocol, in terms of early recognition.

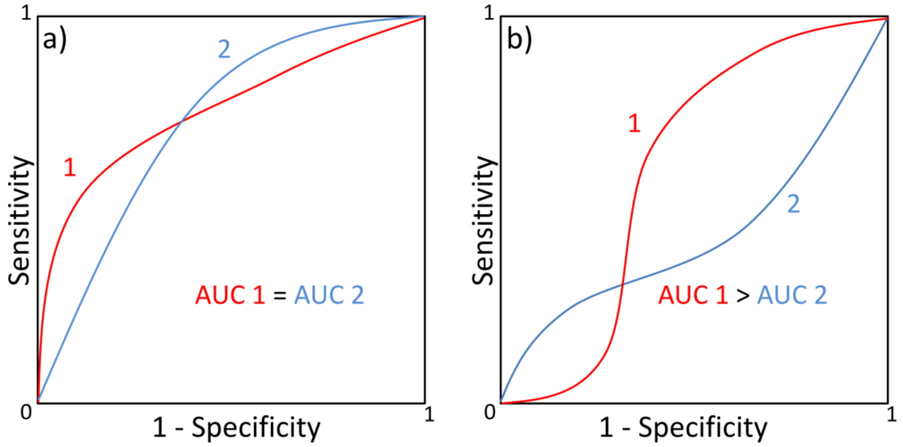


Figure 10. *ROC curves of two different Virtual Screening results, with the same and different AUC value (a and b, respectively).*

In Figure 10a, even if AUC values are identical, Virtual Screening 1earlier retrieves actives in the rank-ordered compared to Virtual Screening 2. In Figure 10b, the correlation of AUC metrics with the “early recognition” is even more misleading, since the docking study with the highest AUC value extracts a lower number of actives at the beginning of the ordered list. An alternative version of the AUC metrics is the Normalized Square Root Area Under Curve, or NSQ_AUC, which emphasize the “early” enrichment in screening results.[150] In details, for each compound rank k, True Positives among k top-scoring compounds in the rank-ordered list are plotted against the square root of False Positive rate, *x = Sqrt(FP).* Also in this case, a perfect separation of actives and inactives returns a value of 1, and 0 for a random selection.

The Accumulation curve (or enrichment curve) is also used to study ranking performance and is defined by the rank of compounds on abscissa vs. the cumulative count of actives in the rank-ordered list along the y-axis. The Area Under the Accumulation Curve is a related metric, computed through the discrete formula:

$$AUAC = \frac{1}{2nN}\sum_{k=0}^{N-1}[F_a(k) + F_a(k+1)] \tag{41}$$

or the continuous formula:

$$AUAC = \int_0^1 F_a(x)\,\mathrm{dx} \tag{42}$$

where $F_a(x)$ is the normalized cumulative distribution function. It describes the probability to early retrieve actives in respect to a randomly selected compounds and spans between the minimum value of *n/(2N)* and the maximum value of *1-n/(2N).* Similarly to the AUC trend, it ranges between 1 (best performance) and 0 (worst performance) and 0.5 for a uniform distribution of actives in the ordered list. The use of this metric to evaluate Virtual Screening performance is limited because of the strong dependence of the number of actives in the dataset.

AUC and AUAC metrics are related as reported below:

$$AUC = \frac{AUAC}{R_i} - \frac{R_a}{2R_i} \tag{43}$$

where R_a and R_i are the ratio of actives and inactives in the dataset. If n « N, then R_i tends to 1 and R_a towards 0 and consequently AUC ≈ AUAC. In contrast to AUAC, AUC metrics has been extensively considered independent of the proportion of actives vs. inactives. However, the formula described above is a strong evidence supporting the contrary.

Another commonly used descriptor is the Enrichment Factor EF, that measures how many more actives are found within a defined ''early recognition'' fraction χ of the ordered list relative to a random distribution.[96] The Equations (44) and (45) describe the discrete and the continuous formulas, respectively.

$$EF = \frac{\sum_{i=1}^{n} \delta_i}{\chi n} \tag{44}$$

and

$$EF = \frac{\int_0^{\chi} f_a(x)}{\chi} \tag{45}$$

δ is a function with a value of 1 if an active i is found before the threshold position defined by χN, otherwise it takes a value of 0. χ spans from 0 to 1, while EF ranges between 0 and the maximum value of 1/χ if χ ≥ nN and N/n if χ < n/N, and an average value of $\lfloor \chi N \rfloor / (\chi N)$in case of a uniform distribution of the actives, where the lower brackets stand for "the largest integer smaller than". The main limitations of this metric concern the dependency on the ratio of actives of the database and the low discrimination ability after the χ threshold. Moreover, all actives have the same weight within the cutoff which means that with this descriptor, it is not possible to discriminate a Virtual Screening result that early recognizes actives from another protocol where all the actives are ranked at the end of the list.

Sheridan et al. have developed a further descriptor, the Robust Initial Enhancement (RIE)[151] that addresses the "early recognition" problem, as opposite to the other metrics described before and is less sensitive to low number of actives within the screened database. It relies on decreasing exponential weight as a function of rank. Below, the continuous and discrete formulas are reported:

$$RIE = \frac{\int_0^1 f_a(x)\mathrm{e}^{-\alpha x}\mathrm{dx}}{1/\alpha(1-\mathrm{e}^{-\alpha})} \tag{46}$$

$$RIE = \frac{\frac{1}{n}\sum_{i=1}^{n} \mathrm{e}^{-\alpha x}}{\frac{1}{N}\left(\frac{1-\mathrm{e}^{-\alpha}}{\mathrm{e}^{\alpha/\mathrm{N}}-1}\right)} \tag{47}$$

where the parameter α is the corresponding 1/χ as reported for the EF metric. Similarly to this latter, the RIE metric describes how many times the average of the distribution of the ranks for actives is better than random distribution. The RIE value is higher than 1 when a large number of actives is better ranked than a random distribution, while a value of 1 corresponds to a random distribution. This metric addresses the "early recognition" problem and then, is more advantageous than previous metrics. However, a significant disadvantage is the dependency on n, N and α.

The Boltzmann-enhanced discrimination of ROC (BEDROC) is closely related to the RIE. It possesses the same ability to take into account the "early recognition" and, as added value, is a standardization of the RIE ranging between 0 and 1. The BEDROC metric is a generalized AUC parameter including a decreasing exponential weight.

Simplistically, it can be described as follows:

$$BEDROC = \frac{RIE}{\alpha} + \frac{1}{1-\mathrm{e}^{\alpha}} \tag{48}$$

if $\alpha R_a \ll 1$ and $\alpha \neq 0$. In this case, the BEDROC is independent from the ratio of actives and corresponds to the probability to retrieve an active before a random compound from a hypothetical exponential probability distribution function with the early recognition parameter α. The BEDROC parameter is very useful to compare the performance of different Virtual Screening and combines the advantages of RIE and AUC. RIE and BEDROC needs the setting of the α parameter. Truchon et al. have shown that a high value allows to count more the early part of the accumulation curve. In particular they recommend an α value of 20 corresponding to say that the first 8% of the list contributes to 80% of the BEDROC value.

3. LDHA - CASE STUDY 1

3.1 WARBURG EFFECT

In order to maintain homeostasis, cells need energy, mostly in the form of adenosine triphosphate (ATP). The main mechanism for energy production is the oxidative phosphorylation (OXPHOS), an oxygen-dependent process which takes place in mitochondria. According to this pathway, oxidation of NADH and $FADH_2$ is coupled with the phosphorylation of ADP resulting in ATP. As second option, cells can produce energy through glycolysis, that is an anaerobic mechanism occurring in the cytosol. In this case, glucose is converted through several steps in piruvate; NAD^+ is consumed and ATP is produced. Contrary to the oxidative phosphorilation, glycolysis is less productive in terms of energy, resulting in 2 ATP molecules versus 38 in the aerobic process, when a molecule of glucose is processed. In summary, glucose represents an important source that cells use to provide energy and, in aerobic conditions, the preferred metabolism for normoxic cells is the oxidative phosphorylation which supplies the 70% of the total energy.

In spite of the low efficiency in providing ATP, it has been found that glycolysis is the selected mechanism during the proliferation process of many fast-growing unicellular organisms, regardless of oxygen availability.[152-153] Similarly, enhanced glycolysis during fast growth has been observed in multicellular organisms.

In 1924, Otto Warburg discovered that cancer cells produce energy predominantly by glycolysis despite oxygen availability, a phenomenon called Warburg effect or aerobic glycolysis. For several decades, this phenomenon has been discredited since Warburg hypothesized that the cause of the aerobic glycolysis was the mitochondrial injury.[154] Unfortunately, experimental evidences showed that cancer cells infrequently present respiratory injury and, for years, the cancer research was mainly focused on the role of genetic mutations, by overlooking the importance of respiratory effect on tumour process. With the introduction of the positron emission tomography (PET), the Warburg effect was verified and emerged as acquired ability which cancer cells use for energy production. In details, glycolysis in cancer cells contributes to ATP production at a rate of 1-64%, depending on cell and tissue types.[155]

Tumours are heterogeneous diseases and the metabolic phenotype can vary even in the same process from one cell to another.[156-157] Recent studies have proven that mitochondrial oxidative phosphorylation is intact in some cancers or, alternatively, can be compromised or insufficient to support the energy requests of cancer cells, resulting in the Warburg effect. Therefore, high levels of glycolysis does not necessarily lead to this process as unique energetic pathway in cancer mechanisms.

In tumour context, aerobic glycolysis is more suitable than OXPHOS for rapidly growing cancer cells, since it provides both ATP and glycolytic intermediates which are used to generate macromolecules essential for proliferation. In particular, glycolysis is a source of carbons and nitrogens for the biosynthesis of nucleotides, phospholipids, fatty acids, cholesterol and proteins, so as to provide biomass production required for cancer growth and proliferation. Precursors for lipids can also be provided by glutamine, representing an alternate pathway for energy production.[158] Another advantage of aerobic glycolysis consists of the faster rate of ATP production, compared to OXPHOS, that meets the high energy demand of rapidly growing and dividing cells, when combined with an increase of glucose uptake.[159] The continuous request of energy for cancer cell proliferation indirectly regulates the consumption of ATP, of which high concentrations can inhibit glycolytic enzymes, such as phosphofructokinase 1 and pyruvate kinase 1.[160] Moreover, when aerobic glycolysis is activated, resulting lactate is transported outside the cells, leading to environmental acidosis. This microenvironment favours cancer cell growth over normal cells, promoting the invasion of neighbouring tissues and metastatic process.[161]

In many cancer lines, the Warburg effect is the result of different factors, including *i*) mutations and deletions of mitochondrial DNA, *ii*) mutations or abnormal gene expression of nuclear DNA, *iii*) oncogenic transformations and *iv*) influence of tumour microenvironment.[162] First of all, mitochondria have an important role in producing ATP, regulating apoptosis, and generation of reactive oxygen species (ROS). Thereby, mutations of mitochondrial DNA lead to increased ROS and mitochondria mass.[163-164] Experimental evidences show that accumulation of glucose metabolites derived from glycolysis activates the hypoxia-inducible factor 1 (HIF-1) which, in turn, increases the transcription of genes encoding for proteins involved in cancer development, glucose metabolism, apoptosis resistance, invasion, metastasis and angiogenesis.[165] Mitochondria dysfunction has been also correlated with aerobic glycolysis through HIF-1 protein expression. It is important to highlight that some cancer cells retain mitochondrial oxidative phosphorylation, and use this energetic pathway as principal route to produce ATP. Therefore, mitochondrial mutations lead to partial defects in oxidative phosphorylation. Other mechanisms favouring mitochondrial dysfunction are the mutations of nuclear genes encoding for mitochondrial protein components, such as succinate dehydrogenase (SDH) and fumarate hydratase (FH), important for the tricarboxylic acid cycle (TCA cycle). Moreover, overexpression of some glycolytic enzymes coded by nuclear DNA, like hexokinase II, glyceraldehyde-3-phosphate dehydrogenase (GAPDH), lactate dehydrogenase (LDH), can contribute to the Warburg effect. Decreased mitochondrial respiration is also linked to a loss of the tumour suppressor p53 which regulates the balance between oxidative phosphorylation and glycolysis, resulting in the promotion of the latter to generate ATP. Oncogenic transformations involving, for example, Ras, Src, PI3K/Akt, and Bcr-Abl, represent another mechanism promoting the Warburg effect. In particular, glucose uptake favouring the initial steps of glycolysis is stimulated by Ras or Src transfection and activation of PI3K/Akt/mTOR pathway, and also Bcr-Abl inhibitor Gleevec has supported the evidence that this oncogene is involved in the Warburg effect.[162] A further condition connected with the switch to glycolysis for ATP production is the hypoxia, which is frequently observed when the tumour mass enlarges exceeding the capacity of blood supply. For example, the high contribution of oxidative phosphorylation to energy production in cervical and breast carcinomas are severely reduced in hypoxic conditions, suggesting that glycolytic phenotype in cancer cells is strongly connected to hypoxia.[166] HIF-1α is involved in this process by mediating the cellular response to hypoxia. For example, it induces expression of glucose transporters and inhibits the conversion of pyruvate to acetyl-CoA by activating pyruvate dehydrogenase kinase 1, resulting in the reduction of oxidative phosphorylation. Finally, the acidification of microenvironment due to high lactate production not only favour tumour progression and metastasis, but also induces multiple glycolytic enzymes that promote the Warburg effect.[167]

3.2 LDHA AS ANTICANCER TARGET

In the light of the essential role of aerobic glycolysis in energy production, cell proliferation, tumour invasion and metastasis, the dependence of cancer cells on this energetic metabolism for ATP generation has been exploited for anticancer therapy. In consideration of the glycolytic mechanism, antitumor strategies are based on inhibition of glycolytic enzymes and reduction of glucose uptake by inhibition of the synthesis of glucose transporters, so as to decrease glucose entry into cells. Figure 11 shows the glycolytic enzymes subject to inhibition in anticancer therapies.

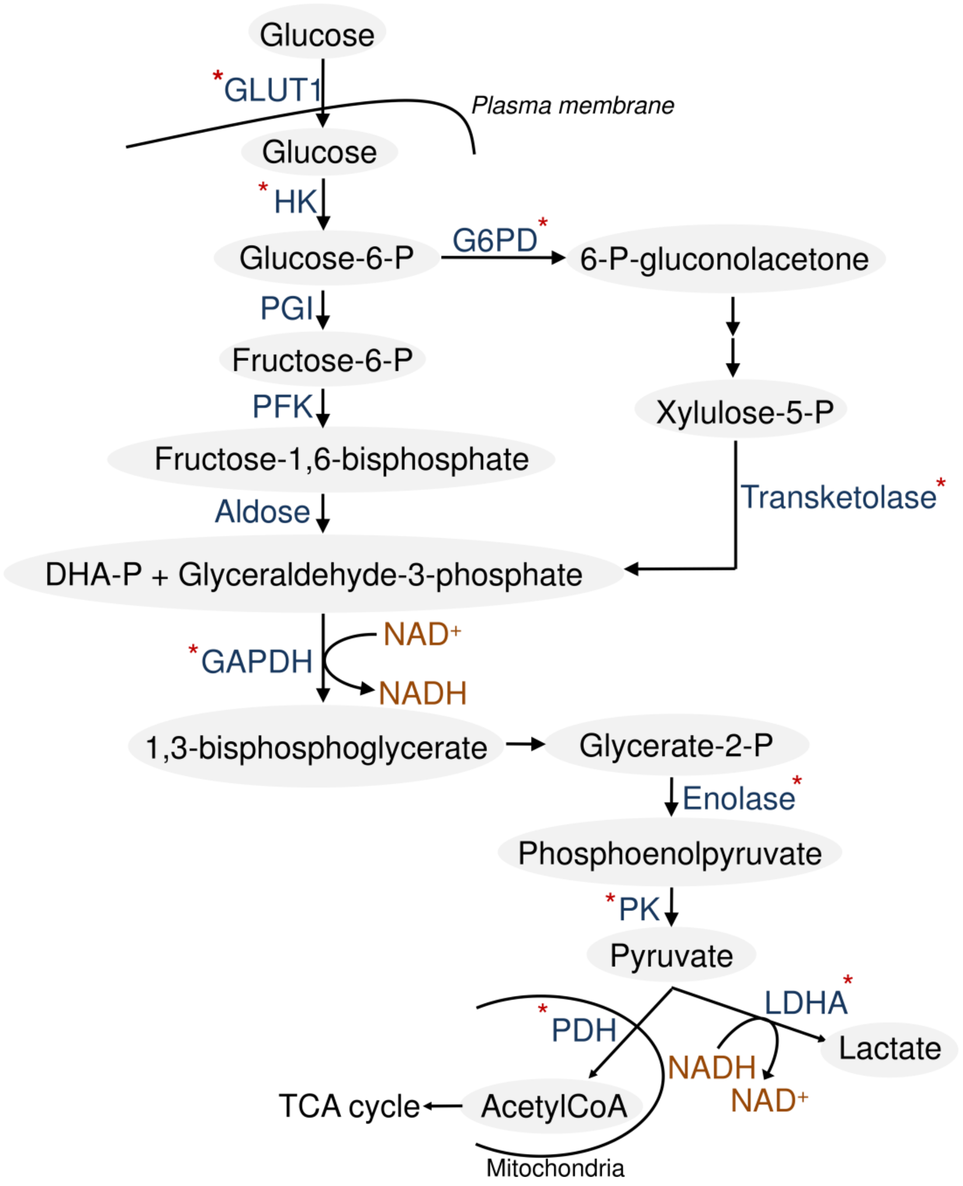


Figure 11. *Scheme of glycolytic pathway. Asterisks remark the possible glycolytic enzymes which can be inhibited for anticancer therapies.* ***GLUT1****: glucose transporter 1,* ***HK****:*

Hexokinase, ***PGI****: phosphoglucoseisomerise,* ***PFK****: phosphofructokinase,* ***GAPDH****: glyceraldehyde-3-phosphate dehydrogenase,* ***G6PD****: glucose-6-phosphate dehydrogenase,* ***PK****: pyruvate kinase,* ***PDH****: pyruvate dehydrogenase,* ***LDHA****: lactate dehydrogenase A,* ***TCA****: tricarboxylic acid cycle.*

Among the glycolytic checkpoints, LDHA represents one of the most attractive anticancer target. In fact, it catalyzes reduction of pyruvate to lactate, with simultaneous oxidation of NADH to NAD^+, which is essential to continue glycolysis, in the step catalyzed by GAPDH (see Figure 11). NAD^+ can also be regenerated through the electron transport chain, but the transport of electrons into mitochondrial matrix via cytoplasmic shuttle systems, such as the malate-aspartate shuttle, requires diverse reactions and is a slower process compared to the pyruvate reduction. Therefore, LDHA represents a fundamental means to regenerate the oxidized cofactor and proceed glycolysis. Moreover, since NAD^+ is required for nucleotide and amino acid biosynthesis, lactate production by LDHA may promote incorporation of glucose metabolites into biomass, so as to sustain rapid cell growth.[168] As consequence, LDHA inhibition leads to reduced glycolytic metabolism of cancer cells and suppression of cell proliferation.

It is important to highlight that LDHA inhibition can be a selective strategy against cancer and is less toxic than blocking the other glycolytic enzymes. In fact, when LDHA is inhibited, pyruvate can be transported into mitochondria and used for energy production through the oxidative phosphorylation. On the other side, the block of other glycolytic enzymes would decrease the total pyruvate generation, resulting in a more severe impact concerning ATP production in both normal and aberrant cells. It has been also observed that deficiency of LDHA is asymptomatic, confirming that its inhibition may be a nontoxic approach to contrast tumor growth.[169] However, since some cancer cells maintain a basal level of oxidative phosphorylation, LDHA inhibition alone could be inadequate to completely abolish ATP production in cancer cells.

To date, only a few LDHA inhibitors have been discovered, most of them with poor inhibitory activity in the low μM in terms of IC_{50} and are reported in Table 1.

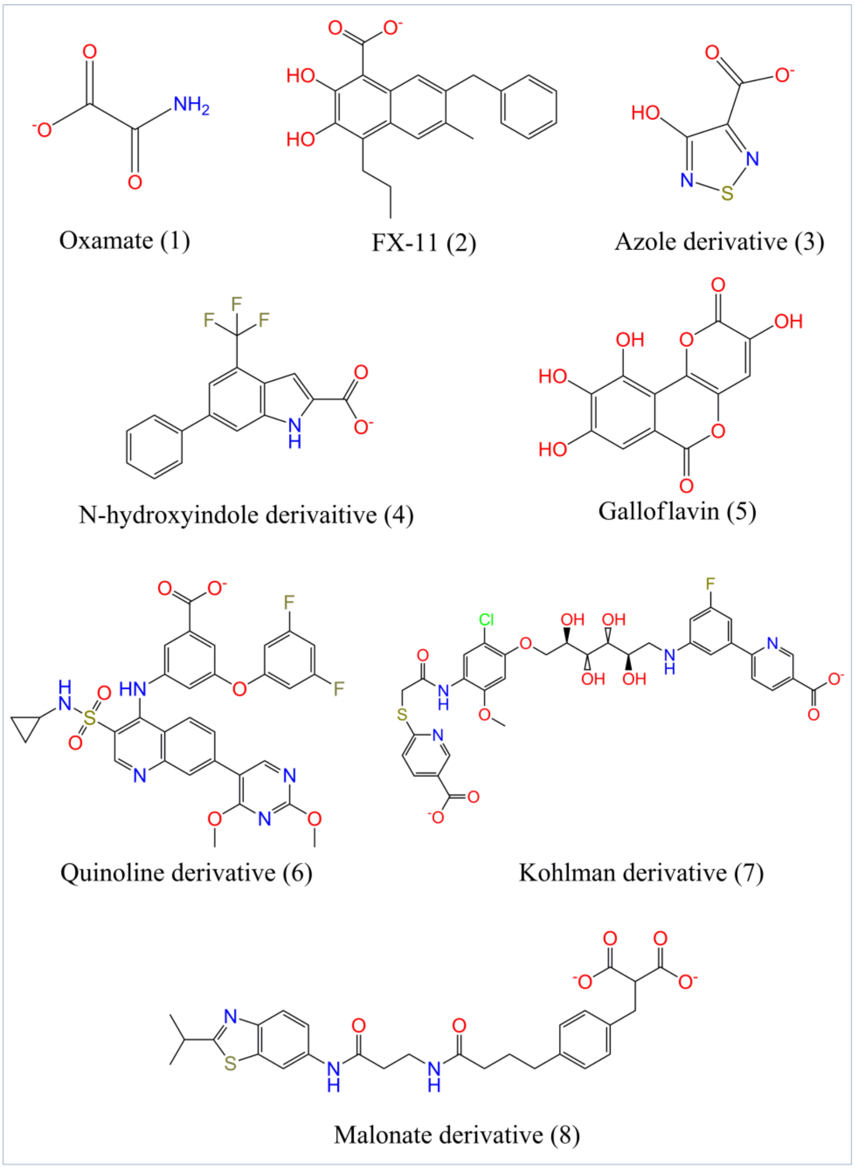


Table 1. *Chemical structures of representative LDHA inhibitors.*

They include oxamate that is a mimic of pyruvate displaying weak toxicity and good selectivity for LDH. However, it shows poor cellular penetration causing a good inhibition activity in vitro only at high concentrations.[170] Gossypol is considered a non-selective competitive inhibitor of LDH since it shows inhibition activity on different isoforms. Use of this molecule as therapeutic agent has been restricted due to the presence of aldehyde functional groups which leads cardiac arrhythmias, hypokalemia, muscle weakness and other severe side effects. Subsequently, gossypol derivatives in which aldehyde groups were converted with other functional groups were developed, retaining biological activity. For example, FX-11 is a gossypol derivative which strongly inhibits LDHA by competing with the cofactor NADH.[171] A series of heterocyclic, azole-based compounds (e.g. compound 3 in Table 1) have been described as inhibitors of Plasmodium falciparum LDH which is involved in malaria. They were also tested on human LDH showing micromolar activity.[172] Moreover, N-hydroxyindole-based LDHA inhibitors (compound 4)[173] and other small molecules, such as Galloflavin,[174] have been discovered. Recent more potent inhibitors have been reported in literature, with LDHA inhibition activity in the nanomolar scale, such as quinoline 3-sulfonamides and malonate derivatives (compounds 6-8 in Table 1).[175-178] Compared to the previous molecules, they are bulky ligands able to bind simultaneously substrate and cofactor pockets.

Interestingly, it has been discovered that short hairpin RNAs (shRNAs) cause knockdown of LDHA in cancer cells, resulting in stimulation of mitochondrial respiration, decrease of cell proliferation under hypoxia and suppression of tumorigenicity.[169]

3.3 LDHA: STRUCTURAL FEATURES

Lactate dehydrogenases are 2-hydroxy acid oxidoreductases which exist as tetrameric isoenzymes, coded by different genes but with a high sequence similarity. The combination of two different types of monomers, that is the M- and A- subunit, coded respectively by ldh-a and ldh-b, give rise to five possible isoforms: LDH1 (H_4 or LDHB), LDH2 (M_1H_3), LDH3 (M_2H_2), LDH4 (M_3H_1) and LDH5 (M_4 or LDHA). Moreover, a sixth isoform, LDHC (C_4 or LDHX), belongs to this enzyme family and is coded by ldh-c gene. Subunits M and H are differently located in the tissues. In particular, isoforms mainly formed by M-monomers (LDHA and LDH4) are found in anaerobic tissues, like skeletal muscles, liver and neoplastic tissues, whereas H-subtypes (LDHB and LDH2) are situated in tissues with predominant aerobic metabolism, such as cardiac muscle, kidney, brain and erythrocytes. LDH3 is placed in lymphatic tissues and platelets, and finally LDHC is found in testes and sperm.[173]

From a structural standpoint, LDH tetramers exhibit a D2 symmetry, with three orthogonal rotational axes (Figure 12a). Each monomer consists of 330 residues forming a bilobal structure of two main domains: *i*) the cofactor binding site and *ii*) the substrate binding domain (Figure 12b). In details, the former includes residues

20-162 and 248-266 adopting a “Rossmann” fold, that is a structural motif consisting of six parallel β-strands linked to two pairs of α-helices.[179] On the other side, the substrate binding domain consists of residues 163-247 and 267-331 which form four β-strands and three α-helices. Among these domains, α-1G2G helix is located in proximity of the active site and includes some amino acids directly involved in substrate binding (Figure 12c). The binding site domain defines the active site which is arranged in a 10 Å depth cavity, reducing solvent accessibility during catalysis. Another characteristic domain is the “active site” loop formed by residues 98-110 which is highly conserved among the LDH isoforms (Figure 12c). This flexible portion can exist in different conformations that can promote the substrate binding and the catalytic reaction. In particular, the mobile loop includes the catalytic residue Arg105 which has a key role for the initial step of the catalysis.

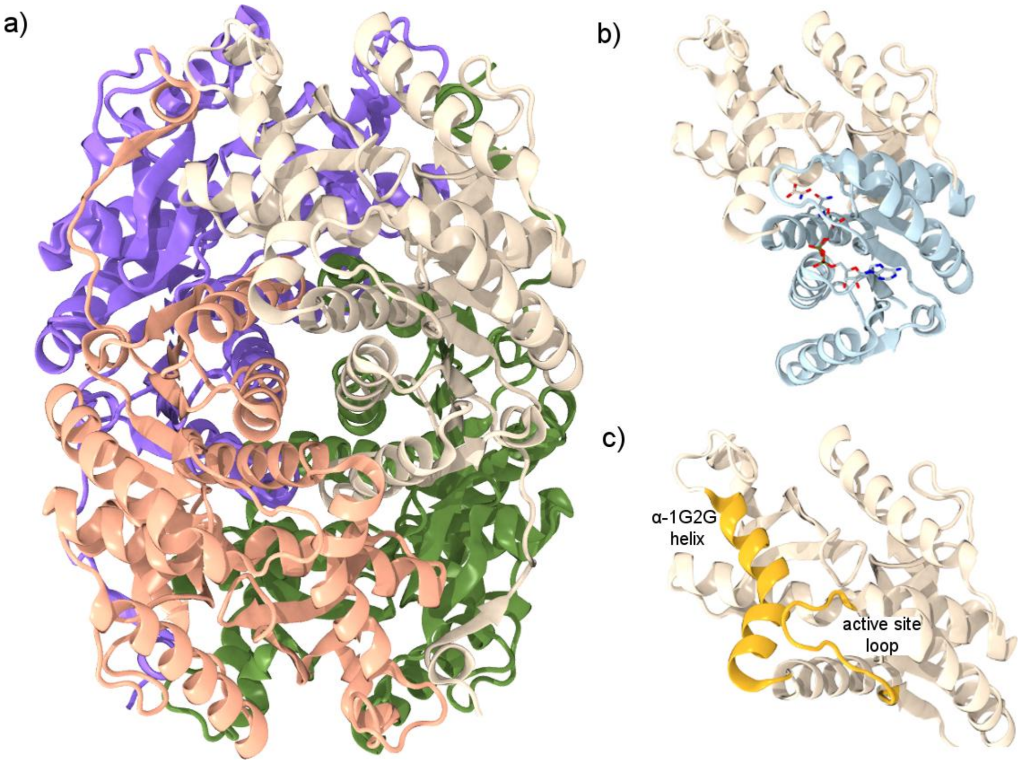


Figure 12. *a) Representation of LDH tetramer. Monomer A-D are shown in white, pink, violet and green. b) The cofactor binding site and the substrate binding domains are shown as light blue and white cartoon representations, respectively. c) The active site loop and the α-1G2G helix are highlighted in the LDH monomer. For the sake of clarity, some helices have been removed.*

An important structural feature of LDH is the interface between the monomers, in proximity of which the active site is located. Computational studies have been carried out, aimed at investigating the role of the second monomer on the stability of the first subunit.[180] Molecular dynamics simulations have confirmed that the tetrameric isoform environment plays a significant role in maintaining the active-site geometry, whereas simulations of the monomeric form resulted in low structural stability of the active site. As a consequence, at least part of the neighboring monomer is necessary to prevent the unfolding of the α-1G2G helix immediately adjacent to the active site. In particular, the α-C helix of the second monomer is in close contact with α-1G2G helix, so as to preserve the structural stability of the first subunit. These computational results were consistent with biochemical studies led by Wang and co-workers, proving that the dimer in the tetrameric enzyme is the minimal functional unit.[181] Figure 13 highlights some domains located at the interface of two monomers. For the sake of clarity, nomenclature of the α-helices characterizing the secondary structure is reported.[182]

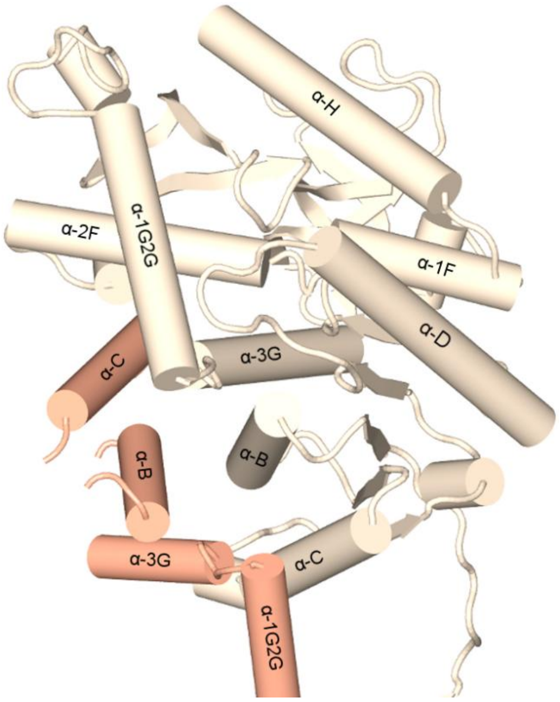


Figure 13. *Secondary structure of the minimal functional unit. Some α-helices of the second monomer (pink) in close contact with the first subunit (white) are highlighted.*

The most expressed LDH isoforms in Homo sapiens are LDHA and LDHB, which are mainly located in anaerobic and aerobic tissues, respectively. They have a high sequence identity (~75%) with 81 different amino acids out of the 330 forming the monomer, mainly located in peripheral regions distant from the cofactor and substrate binding sites. In spite of their high sequence similarity, LDHA and LDHB show a different degree of affinity for the endogenous substrates. In particular, LDHB generally favours the reaction of lactate to pyruvate, whereas the muscle isoform LDHA promotes lactate formation. By calculating electrostatic potential of both isoforms, Read et al. have shown that surface electrostatic effects underlie the different activity response and different kinetic properties between LDHA and LDHB.[183] In fact, the different charged surfaces can influence the pK_a value of the His192 which, in the protonated form, has an important role in both the binding of substrate and the chemical reaction.

As regards the catalytic mechanism, it consists of an ordered sequence of events, initially requiring the cofactor binding. Subsequently, the substrate binds the binary complex and, once the endogenous ligand is close enough to the active site, the mobile loop closes over the ligand to bring the bound substrate and NADH together in a proper geometry for reaction. This loop motion prevents solvent to access to the active site and promotes the hydride transfer. It has been demonstrated that the mobile loop enclosure represents the rate limiting step of the catalytic mechanism occurring in the millisecond timescale.[184] In details, it enhances k_{cat} in LDH by over a factor 1200.[185] Moreover, loop enclosure is associated with slight movements of α-D helix at its base, and a concerted movement of the α-H helix.

The loop enclosure favours the proper rearrangement of Arg105 which is involved in the polarization of the carbonyl group of the substrate, so as to trigger the reaction. Substitution of this residue with a glutamine resulted in a slower hydride transfer than in the wild type enzyme and proved that Arg105 reduces the activation energy barrier for the chemical mechanism by more than 4.2 kcal.[185]

Experimental studies based on the laser-induced temperature-jump relaxation spectroscopic techniques have allowed to examine the kinetics of cofactor and substrate binding to LDH.[186] The formation of the secondary complex LDH-NADH is a multistep process and multiple interconverting conformations have been observed. Some of these structures are defined competent and others noncompetent to form the catalytically productive complex.[187-188] Therefore, the formation of the tertiary complex can be attributed to a conformational selection mechanism. Callender and co-workers have demonstrated that the competent complex is a minor population of conformations characterized by an open mobile loop, whereas noncompetent structures show a closed loop geometry. In particular, different networks of hydrogen bonds and solvent exposure of the binding pocket describe the two species. On one side, the binding competent species can bind ligands at diffusion-limited speeds, indicating that their binding site are highly solvated, whereas in the closed loop structures, the binding domain lies deep within the protein. According to the experimental results, the substrate binds the competent system forming the encounter complex, which is subject, in turn, to structural rearrangements leading to the catalytically competent ternary complex. The mechanism at the basis of this process is in agreement with the induced fit theory. Therefore, the catalytic mechanism can be connected to both the biomolecular recognition models. The collapse of the encounter complex to form the catalytic competent Michaelis complex has been also investigated. Callender et al. have developed some hypotheses in presence of the substrate mimic oxamate and found two possible pathways: *i*), a major populated structure in which all key H-bonds involving His192, Arg105 and Arg168 within the active site are formed and *ii*), a minor structure with unfavorable interactions to promote the chemical reaction.[189] Figure 14 schematizes this kinetic model.

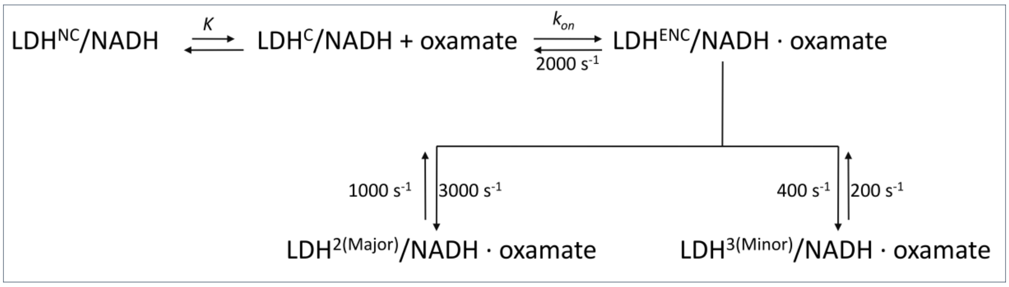


Figure 14. *Kinetic scheme describing the formation of the enzymatically productive ligand-protein LDH complex. NC, C and ENC correspond to noncompetent, competent and encounter complex, respectively.*

Once the Michaelis complex is formed, lactate dehydrogenase catalyzes the reversible reduction of pyruvate to lactate with simultaneous oxidation of NADH to NAD^+ (Figure 15), accelerating the solution chemical reaction by some fourteen orders of magnitude:[190]

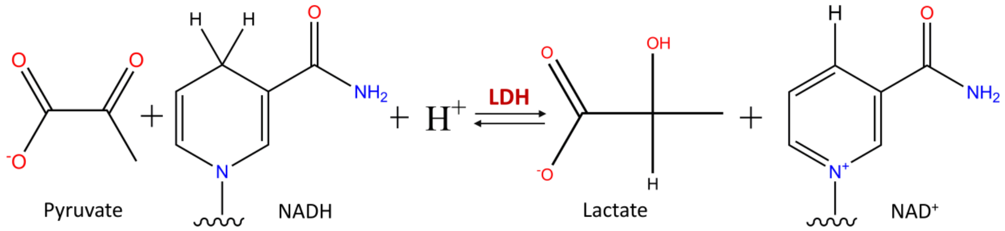


Figure 15. *Schematic representation of the chemical reaction catalyzed by LDH.*

In details, the enzyme catalyzes the transfer of a hydride ion from the pro-R face of the nicotinamide group of NADH to the C2 carbon of pyruvate. In the major form of the Michaelis complex, the pro-R and pro-S hydrogens of the cofactor have been found in the pseudoaxial and pseudoequatorial geometry, respectively.[189] The proton transfer to the carbonyl moiety of the pyruvate to favour the formation of the alcohol lactate is promoted by the catalytic diad His192-Asp165. This acid residue stabilizes His192 through an H-bond with the imidazole group. The positively charged His192 has the function of proton donor through a charge reinforced hydrogen bond with the oxygen atom bound to the C2 carbon of pyruvate. The opposite reaction also occurs with this residue acting as proton acceptor. Arg105 has the key role to polarize the ketone functionality of the pyruvate, so as to promote the hydride transfer from NADH, and interacts also with the carboxylic group of the substrate. Finally, two important residues anchor the substrate within the binding pocket: Arg168 which binds the carboxylate group of the substrate through a two-point electrostatic interaction involving its guanidine group, and Thr247 which forms an hydrogen bond. Figure 16 below schematizes this network of binding interactions.

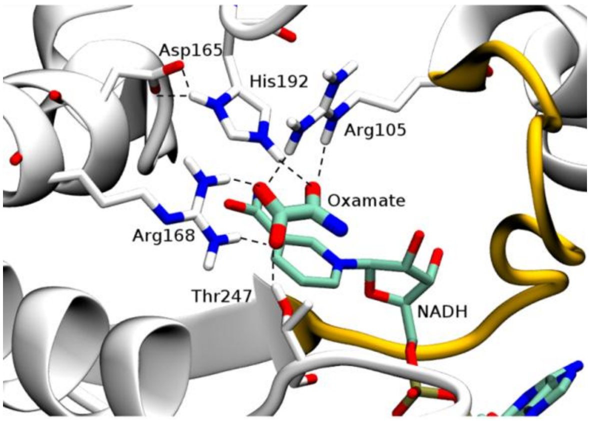


Figure 16. *Schematic representation of the binding mode of oxamate within the binding site.*

3.4 AIM OF THE PROJECT

In the light of the relevant therapeutic role of LDHA in cancer process and the importance of protein flexibility in drug discovery, this project aimed at developing a computational protocol to include the active site loop flexibility in a Virtual Screening campaign.

The crystal structure of human LDHA (PDB ID:1i10) consists of two tetramers (A-D and E-F) cocrystalized with NADH and the LDHA inhibitor oxamate.[183] Chain D presents a different mobile loop conformation compared with the A, B, and C monomers crystalyzed with a closed active site loop state. In theory, different conformations offer a great opportunity to enrich Virtual Screening procedures for novel inhibitors taking into accout protein flexiblity. However, comparison of stability parameters for the chrystallographic A and D chains shows a substantial level of uncertainty about the coordinates of the openmobile loop atoms. In details, the high β-factor valuesindicate a large displacement from their main positionrelated to the closed loop conformation and consequently denotes a highdegree of flexibility. Some portions including residues 100-104 and Arg105 side chain even show zero occupancy. Based on these prerequisites, this study arises from the need to consider large conformational changes LDHA undergoes during ligand recognition that are not sufficiently and accuratly described from the human crystal structure, through enhanced sampling methods.

3.5 RESULTS AND DISCUSSION

Preliminary Molecular Dynamics simulations were performed to identify a minimal model of LDHA structure. The purpose of this step was the definition of a system which preserved the main features of the LDHA functional unit, that is the dimeric form, but with a smaller size, so as to optimize the computational demand and the performance of hREMD simulations. In particular, conventional MD simulations of monomeric and dimeric systems were carried out. Details of the MD setup are reported in the next section. Analysis of the resulting trajectories showed that the monomeric form displayed low structural stability and lost its initial folding structure, especially regarding the α-1G2G helix in close proximity of the flexible mobile loop. Also α-C helix was subject to wide fluctuations; however, the latter was disregarded in this analysis, due to its low influence on the binding site stability. The calculation of the Root Mean Square Deviation (RMSD) and Root Mean Square Fluctuation (RMSF) highlighted the structural stability and the most flexible domains, as reported in Figure 17a and 17b, respectively.

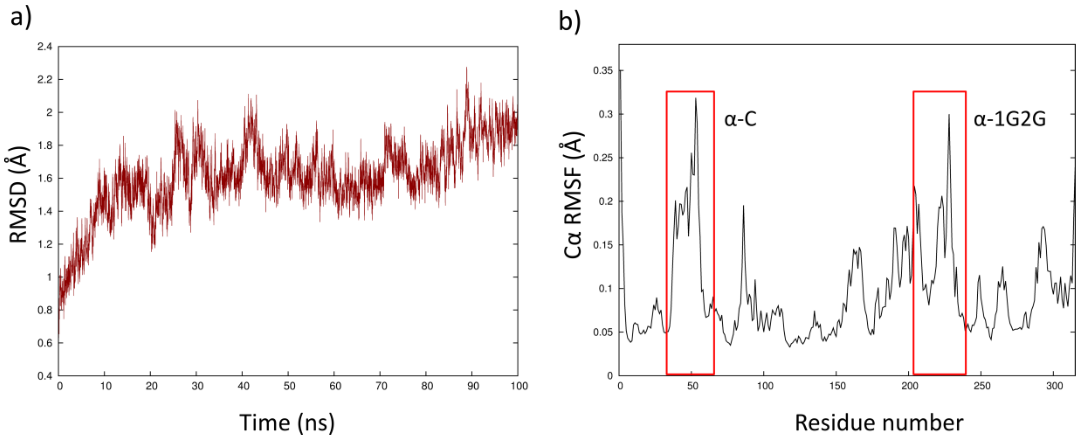


Figure 17. *a) RMSD plot and b) RMSF plot of the chain A. Red squares show the protein domains subject to the highest fluctuations during the 100 ns MD simulation.*

On the other side, the MD studies of the dimeric form displayed a higher stability of the unfolded portions reported in the previous simulations. This result confirmed that the presence of the chain B was necessary to stabilize some domains of monomer A directly involved in the substrate binding mode. Starting from this evidence, the minimal system (Figure 18a) was defined by including a small portion of chain B (α-C helix) aimed to stabilize the flexible α-1G2G helix of chain A and to preserve the folding of the binding site. Most importantly, the presence of the α-C helix was also necessary to shield the active site from an unrealistic ingress/egress of water molecules from the interfacial side of the dimer. A 100 ns MD simulation and the resulting analysis (RMSF calculation) confirmed this hypothesis, as shown in Figure 18b.

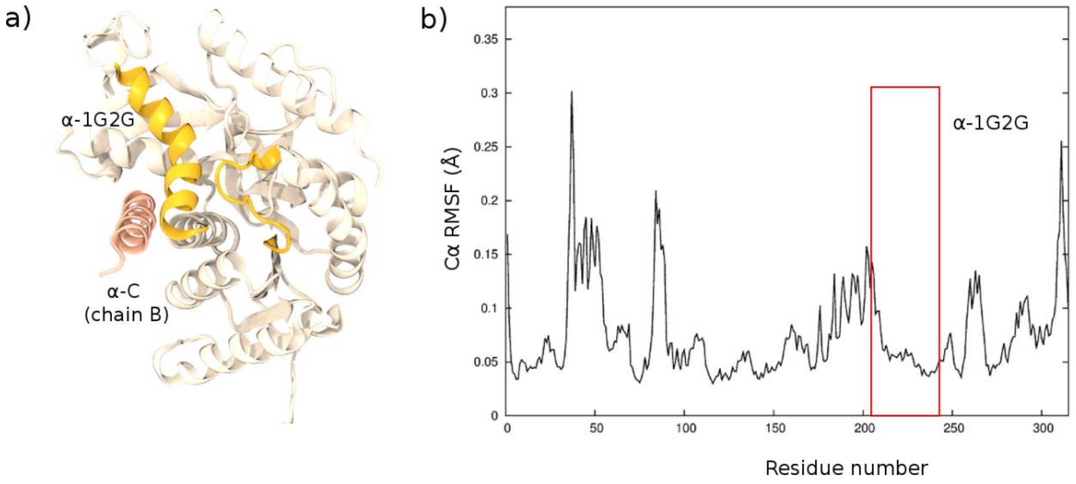


Figure 18. *a) Minimal LDHA system. α-1G2G helix of chain A and α-C helix of chain B are labeled and showed in yellow and pink cartoon representations, respectively. b) RMSF plot of the minimal LDHA system. The red box highlights the structural stability of the α-1G2G helix of chain A, compared to the monomer showed in Figure 17.*

In retrospect, the comparison of RMSD calculated for the α-1G2G helix over time for chain A, chain AB and the minimal system, demonstrated that the monomer presented a low structural stability, whereas the minimal system exhibited a similar stability compared with the chain AB (Figure 19).

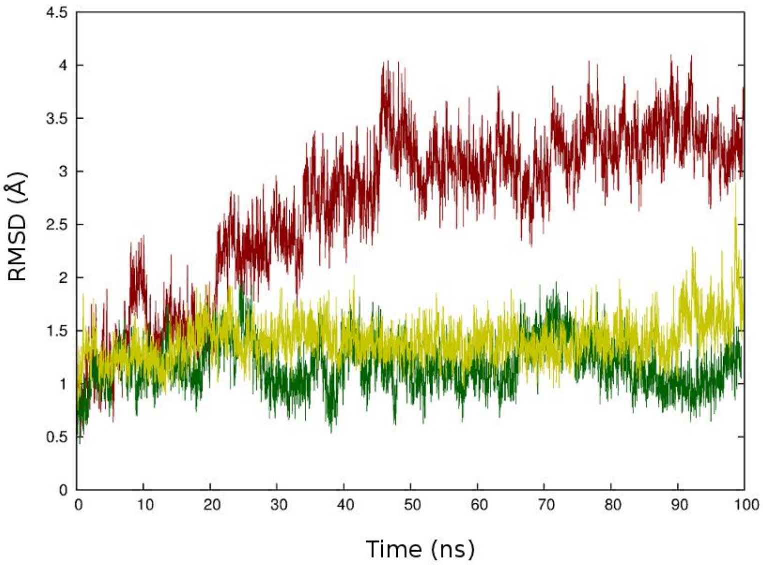


Figure 19. *RMSD calculated for the α-1G2G helix alpha-carbon atoms over time for chain A (red), chain AB (green), and minimal model system (yellow).*

This minimal system was used for the hREMD simulations which led to collect a large amount of configurations, different in terms of the active site loop geometries. The almost 700 ns simulation resulted in 13000 structures. These configurations were analyzed firstly through cluster analysis, which is extensively used to classify configurations sampled by MD methods. In this study, we tested different atom selections as clustering criteria. In details, we computed the RMSD based on all atom backbone and, subsequently, on the mobile loop backbone alone. However, a careful visual inspection showed that the resulting clusters were highly heterogeneous. In other words, conformations with a high diversity in terms of mobile loop rearrangements were assigned to the same clusters. In this scenario, the simplistic selection of the most representative structures from each cluster could negatively affect the subsequent screening campaign. In fact, a careful definition of the initial protein structures is necessary for a successful ensemble docking and the conformations extracted from this cluster analysis were unrelated with the most populated structures sampled with the hREMD simulations. In addition, since the fluctuations of some side chains belonging to the mobile loop were supposed to affect its geometry, the information of these amino acids was overlooked in the similarity measures based on the overall protein backbone. Although a further analysis was performed by including specific side chains as clustering criteria, results were unsatisfying.

Then, network analysis was selected to study the structural diversity of the conformations.In order to optimize the huge memory requirements to allocate all the configurations explored by hREMD simulations, 6500 out of the 13000 sampled structures were selected to build the network layout. In this study, nodes consisted of the collected configurations described as the 12 amino acid loop backbone and Glu101:CD, Glu103:CD, andArg105:CZ, while links described similarity between conformations below a cutoff RMSD value of 1.00 Å. Therefore, nodes having a RMSD below this threshold were connected by links. The resulting network layout is reported in Figure 20.

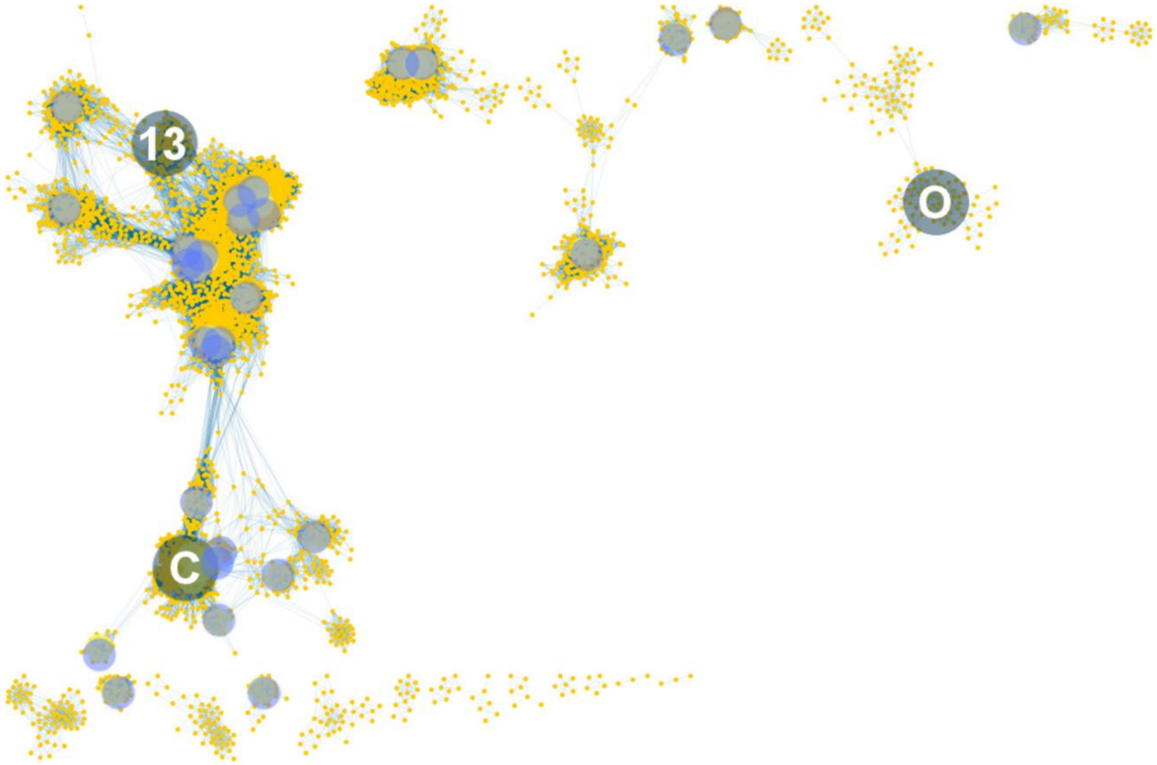


Figure 20. *Network representation of the sampled configurations. Nodes are represented as yellow dots, while links are shown as blue lines. Closed, open loop states and conformation 13 are highlighted. Light yellow dots correspond to the most representative conformations from the first 30 best-ranked clusters.*

To compare these conformations, reference structures were necessary. For this reason, closed and open loop conformations from crystal structures were included in the network analysis. The resulting network included a dense region with sparsely branched islands. Most of the sampled configurations corresponded to intermediate loop structures, between open and closed loop state, although they were structurally more related to the closed than to the opposite state. In fact, the open loop conformation was found in an isolated region and was linked to a few more nodes, by reflecting its low structural stability. The latter region also included some configurations with an even wider loop opening compared to the crystallographic open loop conformation. Moreover, nodes without connections due to unfavorable loop rearrangements were located in further sparse regions. Specific conformations were promoted through interactions involving mobile loop amino acids, such as electrostatic interactions of Arg105 with Asp194 and/or Glu101, interactions of Glu103 with Arg111 and Asn107 or mobile loop backbone atoms.

Subsequently, cluster analysis was performed directly on the network analysis outcome, in order to identify a manageable number of conformations for a subsequent Virtual Screening campaign. Similarity and connectivity of the nodes were used as clustering criteria. Details of the algorithm are discussed in the next section. Visual inspection of these clusters demonstrated that the combination of network and cluster analyses was accurate in discriminating loop conformations and side chain orientations. Figures 21a and 21b show the structural distribution of conformations within the most populated clusters from the initial cluster analysis and the network analysis, respectively, highlighting the structural homogeneity for the latter.

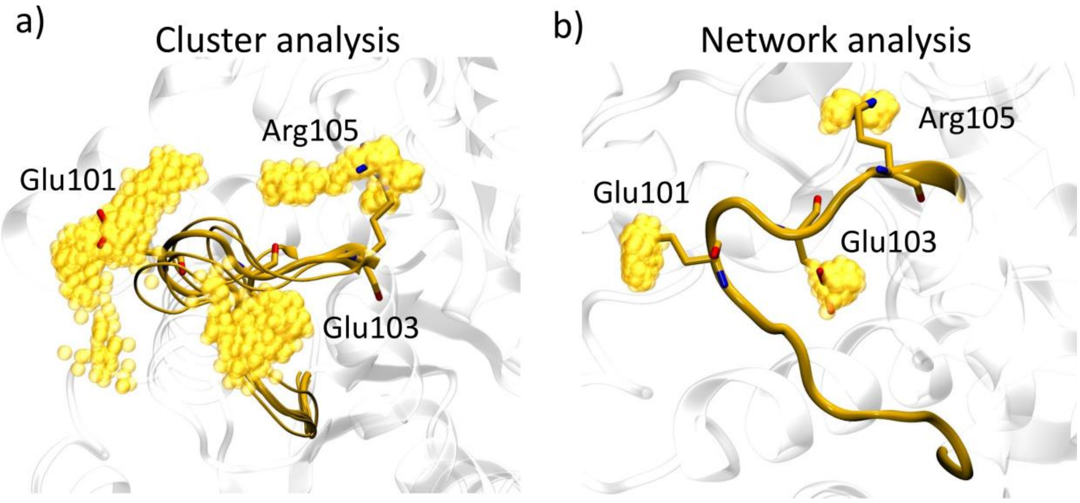


Figure 21. *Distribution of side chains and mobile loop in the largest cluster from cluster analysis (a) and network analysis (b).*

This analysis resulted in 238 clusters which were scored and ranked. Among them, 117 clusters consisted of less than 20 nodes and were not considered as representatives of the structural ensemble. Light yellow circles in Figure 20 correspond to the most representative conformations (referred to as seeds) from the first 30 top-scored clusters. They were mainly located in the central core of the network layout and then were structurally linked to the closed state. The 30 protein structures were employed for the subsequent retrospective Virtual Screening. Besides these seeds, the closed conformation coming from the crystal structure was included in the screening because of its unquestionable experimental significance. Moreover, a representative conformation of the open state was also considered. In particular, the most similar representative to the open loop conformation was selected among the entire ensemble of configurations sampled with the hREMD simulations.

Two ligand datasets were used to validate the conformational ensemble, including: *i*) known active and inactive compounds coming from the BindingDB[191] (active/inactive ratio: 21/17), and *ii*) known active compounds and decoys generated by the DUD-E database (active/decoy ratio: 21/1950).[192] A cutoff of 30 μM was selected to classify active/inactive chemicals and both IC_{50} and K_i activity data were considered. The choice of this threshold and the unbalanced number of actives/inactives were imposed by the low availability of high-affinity actives and, in general, experimental LDHA inhibition data collected in BindingDB. Furthermore, known BindingDB LDHA ligands were properly filtered out in order to limit chemical redundancy within the datasets. Docking calculations were carried out into the LDHA binding pocket using all the 32 conformations. Figures of merit were computed for all docking results, including AUC, AUAC, RIE, BEDROC metrics.[96] Analysis of DUD-E screening results showed that all the LDHA conformations had comparable recognition capability. In fact, AUC values spanned between 0.75 and 0.99. On the other side, BindingDB dockings resulted in good recognition performance with AUC value above 0.8 only for the conformation 13 (0.81), while crystal structures of open and closed conformations showed lower AUC values (0.56 and 0.77, respectively). Figure 22a and 22b displayed the AUC values computed for the 32 ensemble structures and the related mobile loop conformations.

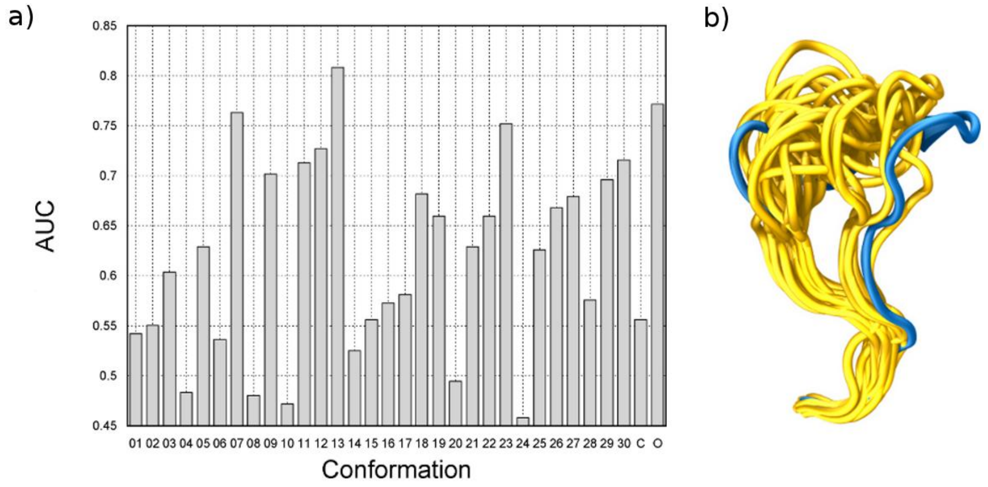


Figure 22. *a) Area Under the ROC curve calculated for the BindingDB data set using 30 seeds plus the C and O conformations. b) Active site loop conformations adopted by the 30*

collected seeds (yellow tubes)together with conformation C and O (blue tubes on the left and on the right, respectively).

This result could be related to the different binding site volume displayed by these three conformations. Conformation 13, indeed, showed an intermediate loop geometry located between the open and the closed state, likely resulting in a more permissive structure, able to accommodate ligands of different size (see Figure 23). On the contrary, too small or widely open pockets could limit a good compound screening. The calculation of protein site volumes were performed for the 32 conformations by means of SiteMap, so as to correlate the docking results with binding site accessibility.[193] Different loop geometries and side chain orientations led to distinct binding site volumes, mainly spanning between the closed and open state (355.00 to 841.07 Å^3). In details, some seeds showed lower or higher binding site volume compared to the closed and open conformation respectively, due either to a different position of side chains mobile loop and NADH binding site, or a different backbone rearrangement of the same regions. Conformations with a small binding site volume, such as seed 20 or 24, displayed an unsatisfying docking performance, because of a narrow cavity reflecting a limited capability to properly accommodate and recognize molecules of diverse size. For instance, small inactive ligands were well ranked within conformation 24 probably due to favorable interactions in a well enclosed pocket. On the other hand, seed 13 exhibited a well defined binding site volume (734.02 Å^3), so as to promote binding of both small and large ligands.

It is worth underlying that the AUC values as obtained using BindingDB and DUD-E decoys were well correlated (r^2 = 0.65), highlighting that some conformations resulted in good docking performances regardless of the ligand dataset. However, while open loop state and seed 13 showed a high AUC value in both data sets, a dissimilar performance wasobtained for the closed loop conformation. Indeed, the latter early recognized largemolecules along with small actives, compared to decoys. Nevertheless, some BindingDB inactives resulted well ranked probably because of overestimated favorable interactions promoted by a well-defined enclosed cavity.

The results of individual runs were combined through ensemble-averaged scores, ensemble-averaged ranks and Boltzmann-weighted average, in order to evaluate the enrichment improvement by using a multiple receptor approach. The average of the ensemble scores showed the best results. In particular, we evaluated the enrichment by combining docking scores for the ensemble consisting of crystal closed and open loop structures, with each of the 30 sampled conformations. Figure 23 shows the ROC curves of docking results from closed and open states, seed 13 and the combination of these three structure that was one of the best combinations to increase enrichment.

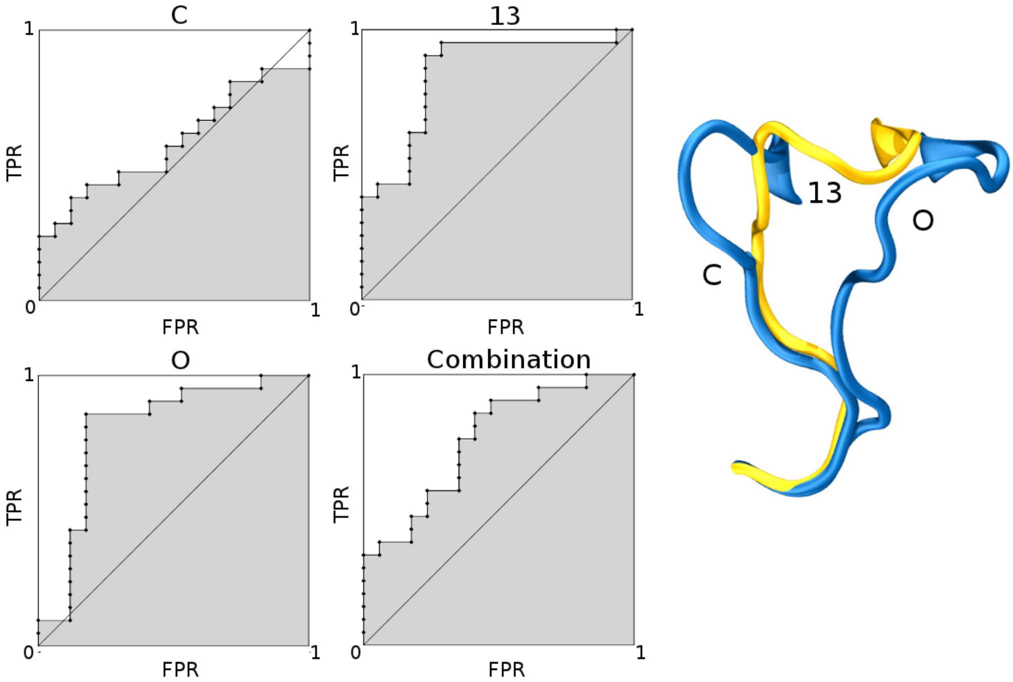


Figure 23. *ROC curves calculated for the BindingDB data set using conformations C, O, seed 13, and their combination. The ROC space is defined as true positive rate (TPR, y axis) versus false positive rate (FPR, x axis). Corresponding active site loop conformations are also shown in blue and yellow tubes.*

To validate the best performing conformations in view of prospect Virtual Screening campaigns, the most recent published datasets were used for further docking studies. They are referred to as Kohlmann, Dragovich, and Fauber datasets, consisting of 6/7, 5/11, and 18/15 actives/inactives, respectively.[176-177, 194] Docking results from the Kohlmann database showed AUC values between 0.64 and 1.00. In details, closed and open loop states had a perfect recognition capability, whereas conformation 13 exhibited an AUC value of 0.67. Analysis of binding poses highlighted that the lower docking performance of conformation 13 was due to the incorrect binding mode of two active compounds that were wrongly docked into the binding site. Even though this result was not in line with BindingDB dockings, binding mode analysis and comparison with crystallized ligand−protein complexes highlighted that seed 13 reproduced well accommodated small molecules, such as the cognate ligand from PDB ID 4I8X. Moreover, bulky molecules (for example, PDB ID:4I9H) were properly located within the binding pocket, for C, O and seed 13, with the exception of the two molecules above-mentioned for this latter conformation. Dragovich dataset included molecules cocrystallized with the LDHA active site loop in a relatively open state and Arg105 displaced from the binding pocket. The recognition performances of conformation 13 and close and open loop states were in line with the experimental results. In other words, the open state conformation resulted in the highest AUC value, followed by seed 13 among the best ranked conformations. Conversely, the poor recognition performance showed by closed loop state was probably due to the orientation of the Arg105 side chain. Therefore, large binding pockets, as found in the conformation with the open mobile loop, promoted a good docking performance. Finally, docking run with the Fauber dataset confirmed seed 13 as one of the best conformations in recognizing active compounds, in agreement with BindingDB results. The docking results also highlighted the good capability to reproduce binding mode of cocrystallized ligands.

According to validation results, conformation 13, along with crystallographic closed and open loop states were selected for consecutive Ensemble-based Virtual Screening. Three parallel screenings were performed, following the workflow illustrated in Figure 24. Glide SP and XP were used to dock Asinex database, opportunely filtered as described in the next section. In details, the first 10% of the docking poses scored according to Glide SP were selected for a more accurate docking study with Glide XP. Then, 1000 best-ranked compounds per conformation were clustered and visually inspected, in order to carefully investigate the resulting binding poses. In particular, the most representative ligands for each cluster, along with other compounds from the most populated clusters were taken into account.

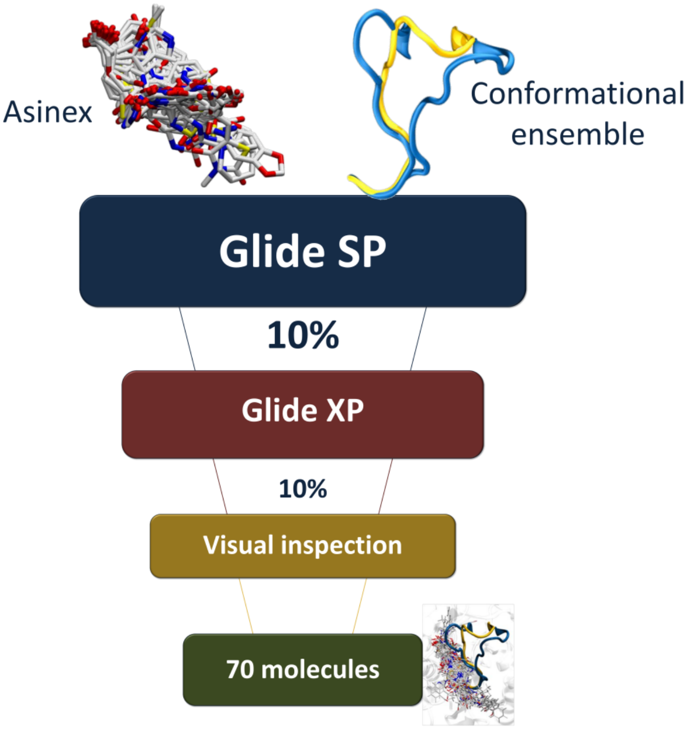


Figure 24. *Ensemble-based Virtual Screening workflow.*

In principle, large molecules tended to bind within both substrate and NADH pockets making contacts with protein residues generally involved in binding with the endogenous molecules, whereas small molecules showed variable poses, depending on whether the mobile loop was in the open, intermediate or closed state. For example, some ligands were located within binding pockets defined through alternative side chain rearrangements or around the open active site loop. Other ligands bound pockets normally filled by the adenine moiety of NADH. Figure 25 shows some binding poses as obtained from Glide XP docking.

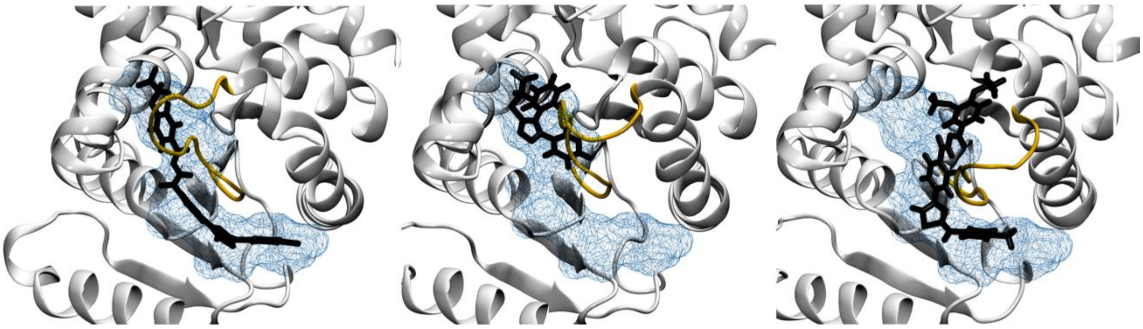


Figure 25. *Binding poses extracted from ligand docking on the closed loop (left), seed 13 (middle) and open loop state (right). The small molecules are located within the binding*

site (blue grid) but in different regions based on the accessible pockets defined by the mobile loop.

The conserved interactions with the catalytic residues along with the binding pockets (substrate and/or cofactor binding sites) in which ligands were located were the main criteria to select the most promising hit compounds.

In total, 70 promising molecules were chosen for biological tests: 39 and 12 compounds from closed and open loop states respectively, and 19 ligands from the seed 13. Some of these molecules were simultaneously extracted from multiple structures. As regards the ligands selected from the closed loop state, most of them bound within the substrate binding site and reached the distant cofactor pocket. In particular, large molecules extended in the direction of the cavities in which sugar moiety and nicotinamide of NADH are usually located. The 12 ligands extracted from the docking with the open state conformation showed non-conventional binding modes, since a larger substrate pocket gave access to additional cavities. However, most of them preserved the electrostatic interactions with the catalytic amino acids. Due to a more accessible binding pocket, a few molecules from this docking run bound the nicotinamide pocket. Finally, the intermediate loop geometry of seed 13 allowed the accommodation of ligands within the binding sites in diverse modes. In particular, some molecules presented alternative binding modes within additional small cavities around the mobile loop, whereas others preserved the simultaneous binding within the substrate and cofactor pockets.

The most frequent moiety of the 70 molecules was a carboxylic acid, mimicking the same functional group of the endogenous substrate. In details, it interacted with the catalytic amino acids of the substrate binding site, such as the Arg168, protonated His192 and Arg105. Moreover, this ligand subset also included some molecules with tetrazoles, in order to investigate the biological activity of bioisosters of the carboxylic acids.

Finally, they were purchased and tested for their LDHA inhibitory activity. The enzymatic activity was measured by monitoring NADH oxidation over time, by recording the decrease in NADH fluorescence after addition of the inhibitors. Moreover, biological evaluations in terms of cellular assays were carried out. In particular, inhibition of lactic acid production and effects on cell proliferation were monitored. Through these experimental measurements, four molecules were found with an inhibition activity (IC_{50}) lower than 100 μM: 3 of them related to the close active site loop and the other from the open loop state. Interestingly, one of them showed a promising activity from both enzymatic and cellular assays. A few other molecules exhibited inhibitory activity in the high micromolar scale, including molecules extracted from the Virtual Screening carried out on seed 13.

In order to investigate the novelty of the discovered inhibitors, we evaluated the chemical diversity between nine μM hit compounds found through the screening and known LDHA inhibitors from BindingDB. In details, we generated MOLPRINT 2D fingerprints with Canvas tool, describing the structure of the molecules based on atom environments.[195-196] In general, the most common type of fingerprints consists of a series of binary digits (bits) that represent the presence or absence of particular substructures in the molecule. Comparison of these descriptors allows to define the similarity between two molecules, to find matches to a query substructure, in perspective of clustering and screening. In this project, a distance matrix was computed based on MOLPRINT 2D fingerprints. In particular, the Tanimoto similarity was calculated between the hit compounds and BindingDB data set. In general, high Tanimoto values correspond to high similarity between two molecules. The BindingDB ligands selected for this analysis consisted of 13 compounds from Deck publication and 25 compounds extracted from Ward project.[171, 178] The resulting distance matrix is reported in Figure 26. According to the color scale, the novel LDHA inhibitors showed a low Tanimoto similarity, that is they were chemically diverse compared to the BindingDB data set. Therefore, we speculated that new regions of the chemical space were explored through the LDHA hit compounds.

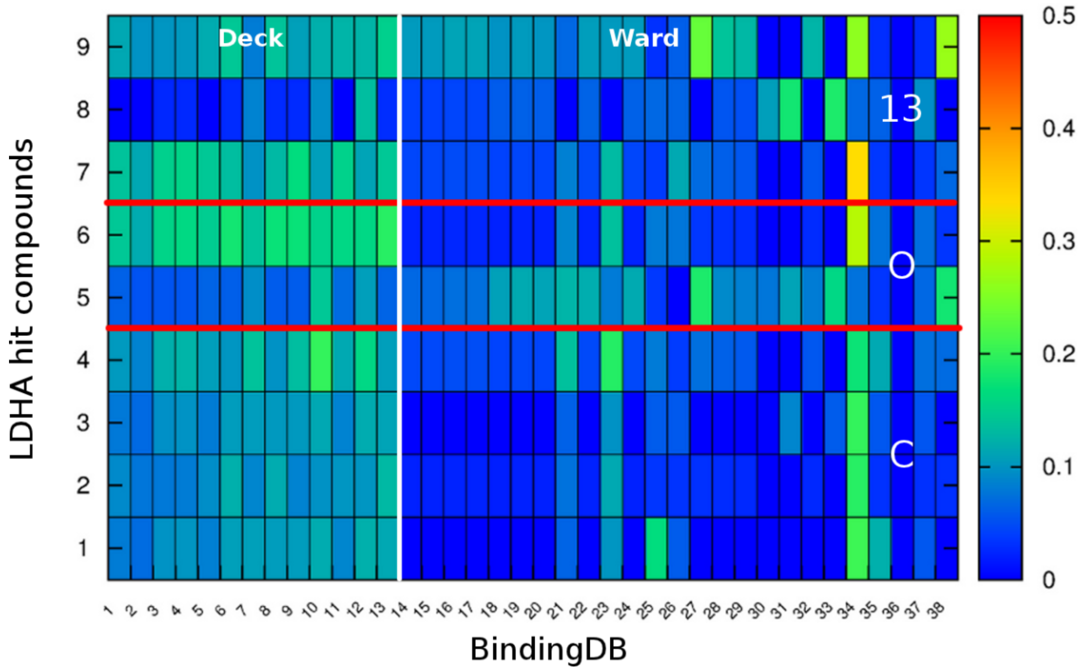


Figure 26. *Distance matrix based on Tanimoto similarity. Blue indicates low chemical similarity (Tanimoto < 0.1). White line separates Deck and Ward subsets within BindingDB database (compounds 1-13 and 14-38, respectively). Red line separates the novel LDHA inhibitors according to the related protein structures of Virtual Screening.*

It is evident that the distance matrix was divided in two blocks, based on the two ligand sources within the BindingDB database, as described before. Also, the hit compounds were ordered according to the LDHA conformations from which they were selected. In particular, ligands 1-4, 5-6 and 7-9 were extracted from Virtual Screening on the closed loop state, open loop state and seed 13, respectively (red line in Figure 26). Analysis of the distance matrix according to these separations highlighted that compound 8 selected from seed 13 was chemically different from both Deck and Ward subsets. This interesting result confirmed the enrichment of multiple conformation approaches compared to Virtual Screening campaigns carried out with single crystal structures. It offers the possibility to design, synthesize and test more potent inhibitors in a poorly explored chemical space. Then, this protocol can enhance the probability to discover novel and also chemically diverse LDHA hit compounds.

3.6 SIMULATIONS SETUP AND ANALYSIS PROCEDURES

3.6.1 THE MINIMAL LDHA MODEL

Explicit solvent MD simulations of chain A, chains AB, and the minimal model were performed with NAMD2.8[197] using the parm99SB-ILDN force field.[198] All the systems were prepared with the LEaP module of AMBER11. The N-terminal residues (1-19) were removed from each monomer. TIP3P water model was used and all hydrogen atoms were added. All ionizable amino acids were charged based on their standard protonation state, whereas the catalytic His192 was protonated according to the proposed catalytic mechanism. Counter-ions were employed to neutralize all the systems and the simulations were performed including the NADH cofactor in the binding site. Four cycles of 1000 steps of conjugate gradient were performed to energy minimize the systems, followed by three steps of 100 ps to thermalize the system to 300 K in the NVT ensemble. The restraints initially added on alpha-carbons were gradually released during the heating procedure. At the temperature of 300 K, an adjustment of the volume of the cell was reached through an additional step of 100 ps in the NPT ensemble. Long-range electrostatics was treated with Particle Mesh Ewald, Periodic Boundary Conditions were applied, and the SHAKE algorithm was used to allow an integration time-step of 2 fs. Simulations of 100 ns in the NPT ensemble were carried out for each system, by using a Langevin thermostat with a collision frequency of 5 ps-1, and a combination of Nosé-Hoover and Langevin barostat implemented in NAMD 2.8 at the pressure of 1 atm. Resulting trajectories were analyzed with the Plumed software.[199]

3.6.2 hREMD SIMULATION SETUP

The closed and open loop of the minimal LDHA model were studied through two different hREMD simulations with the AMBER11 package using the parm99SB-ILDN force field. First of all, 500 steps of steepest descent followed by 500 steps of conjugate gradient were used to minimize the initial structure. Equilibration was performed in explicit solvent in the NVT ensemble, with Periodic Boundary Conditions and Particle Mesh Ewald, followed by additional 500 ps in the NPT ensemble. Finally, a further equilibration in the NVT ensemble was performed for

200 ps. This procedure guaranteed to maintain a correct average pressure in the physical replica when performing simulations in the canonical ensemble, as it is required for most REMD implementations. Twenty four replicas were simulated in parallel in a temperature range between 300 K and 500 K. A geometric progression was used for the distribution of temperatures within this range. In details, the ratio between consecutive elements is constant over the temperatures. On the other side, in case of constant intervals between temperatures, variations of the overlaps between consecutive Gaussian curves resulted in variable probability to accept the exchanges.

First of all, q parameter was computed as follows:

$$q = \sqrt[n-1]{\frac{T_{max}}{T_{min}}} \tag{49}$$

where q is the incremental factor, n is the number of replicas and T_{max} and T_{min} are 500 K and 300 K, respectively. Then, T_i was assigned to each ith replica, based on the following formula:

$$T_i = T_{min} \times q^i \tag{50}$$

To limit the protein unfolding due to high temperatures, restraints were applied to all alpha-carbons, except for the active site loop and a portion of 7 residues in the α-1G2G helix, so as to allow a mutual conformational adaption. A force constant of 0.50 kcal/Å^2 was used. It was defined through classic Molecular Dynamics simulations performed at 500 K. In order to neutralize the system showing a low net charge (+1/331 residues), the charge excess was uniformly spread over the protein atoms as an alternative to use counter-ions. As regards the solvent treatment, 700 water molecules were preserved in the hybrid representation. This solvent shell was defined by considering water molecules within a radius of 3 Å over the protein during 200 ps of conventional MD at 300 K. The total aggregate time from the two simulations was almost 700 ns.

3.6.3 CLUSTER ANALYSIS

The average linkage algorithm, as implemented in Amber tool was selected to perform a cluster analysis. In order to optimize cluster analysis, different parameters were tested. First, the 13000 configurations were aligned with the exclusion of the active site loop, RMSD was computed by considering only the LDHA mobile loop backbone and, finally, configurations were clustered according to an RMSD threshold of 1.00 Å. However, visual inspection of the results highlighted poor homogeneity within each cluster, due to the loss of the alignment of the initial elements during cluster analysis. Subsequently, a second trial was carried out using the same RMSD cutoff computed on the overall backbone instead of the mobile loop. However, an all atom RMSD calculation reduced the differences of the mobile loop orientations, resulting in unsatisfying results.

3.6.4 NETWORK ANALYSIS

Cytoscape software[200] was used to build network graph. In details, the sampled configurations were aligned based on the backbone atoms with the exclusion of the mobile loop, RMSD pairwise matrix was computed, without realignment and nodes with a pairwise RMSD below a threshold of 1.00 Å were used to define the network by means of the Force-directed layout algorithm.[201] Different trials were performed in order to define the best RMSD cutoff able to well connect the most similar configurations. This algorithm considers the graph as a virtual physical system of interacting forces; nodes interact through electrical repulsion that separates each other, whereas edges act as springs and represent attractive forces. During the optimization step, the forces are computed and new equilibrium states are defined, until a local minimum energy layout is achieved. Default parameters were used and 10000 steps were set for the iterative optimization process. The resulting network was clustered according to the connectivity and similarity of nodes. Highly interconnected subgraphs were found through the MCODE algorithm[202] available as a plug-in for the Cytoscape software. It consists of three main steps: *i*) node scoring, *ii*) cluster finding and *iii*) post-processing. In details, higher scores are assigned to nodes whose direct neighbors are more interconnected and subgraphs composed by the top scoring nodes are identified. This second stage recursively moves outward from the seed (defined as the highest scored node), by including nodes whose score is above a given threshold, defined as follows:

$$\text{Threshold} = (1.0\text{-Node Score Cutoff}) \text{ x } (\text{Score of Seed Node}) \qquad (51)$$

A low threshold results in large clusters. After different trials, a low node score cutoff value of 0.1 was set, in order to obtain accurate and appropriate subgraph sizes. The cluster score is defined as follows:

$$\text{Cluster Score} = \text{Number of Nodes} \times \text{Density} \qquad (52)$$
$$= \text{Number of Nodes} \times (\text{Number of Edges} / \text{Number of Possible Edges})$$

Therefore, the best-scored clusters correspond to highly interconnected subgraphs.

3.6.5 RETROSPECTIVE VIRTUAL SCREENING PROTOCOL

3D geometry, tautomers, stereoisomers, and protonation states of all ligands were defined with LigPreptool. 1950 3D decoys were then generated by the DUD-E database. All the docking calculations and analysis were performed with the Schrödinger suite. As regards the protein structures, they were prepared by the Protein Preparation Wizard workflow. Water molecules coming from hREMD trajectories as well as the NADH cofactor were deleted, and the resulting structures were energy-minimized using the OPLS2005 force field.[203] The binding site was defined by a box of 12 Å that included both the substrate and the cofactor pockets, centered in proximity of the bound oxamate inhibitor. Glide XP protocol and GlideScore were used respectively to dock and rank actives, inactives, and decoys.[141-142] The performance of each structure was evaluated by means of the scripts from Schrödinger suite.

3.5.6 ENSEMBLE-BASED VIRTUAL SCREENING PROTOCOL

Proteins were prepared as reported in the previous section. As regards ligand dataset, structure-based Virtual Screening approach was employed to obtain new hits from a commercial library of Asinex database containing around 500000 unique structures.[204] Ligands were filtered according to physical and chemical descriptors based on a slightly modified Lipinski’s rule of five (see Table 2).

Physical-chemical filters	Threshold value
Hydrogen bond acceptors	≤5
Hydrogen bond donors	≤10
Molecular weight	≤600
Number of chiral centers	≤2
Number of rotatable bonds	≤10

Table 2. *Physical-chemical filters applied to Asinex database.*

These filters were selected to optimize pharmacokinetic properties, preventing poor absorption and permeation. Additional filters were applied at this step to discard ligands with specific chemical substructures associated with poor chemical stability or toxicity. In fact, the rule of five takes into account physiochemical properties that impact the earliest stages of drug discovery, but overlooks the chemical reactivity that might lead to in vitro false positives in biochemical screening.[205] Specific functional groups have been identified as reactive false positives or promiscuous inhibitors. In particular, molecules containing Michaelis acceptors and aldehydes were filtered out, since these groups are responsible of electrophilic protein-reactive false positives in biochemical assays. In other words, they are prone to decompositions by solvolysis or hydrolysis and are characteristically reactive towards biological nucleophiles including target protein, serum protein and glutatione.[205] Moreover 1,2-dicarbonyl moieties were discarded from Asinex database, since they are responsible for artifact data in biochemical screens. Finally, only one nitro group per molecule was allowed to limit the toxicity. Glide standard precision mode was used for the current docking study, with default parameters. 52000 poses were retained and re-docked through Glide extra precision, so as to obtain more accurate scores and docking poses. This workflow was applied for each conformation. Finally, the first 10% of docked compounds

(1000 per protein conformation) were group together through a cluster analysis with Canvas and visual inspected.

3.5.7 MOLECULAR SIMILARITY SEARCHING

MOLPRINT 2D fingerprints were computed for the most potent LDHA hit compounds from the Ensemble-based Virtual Screening and 38 known inhibitors from BindingDB database. MOLPRINT 2D fingerprints are descriptors based on atom environments to define chemical structures.[206] They are computed through two steps: *i*) Sybyl atom types are assigned to every heavy atom of the molecule and *ii*) an individual atom fingerprint is calculated for every heavy atom in the molecule. A count vector is constructed with the vector elements being counts of atom types at a given distance from the central atom. Atom environments are stored as binary presence/absence features for each molecule. Subsequently, a distance matrix based on Tanimoto similarity was computed by taking into account the MOLPRINT 2D fingerprints.

3.7 CONCLUSIONS

A combination of an enhanced sampling method with network and cluster analysis led to a collection of multiple LDHA binding pocket conformations which were validated in order to assess their ability to discriminate active from inactive inhibitors. In particular hREMD simulations were carried out to collect multiple conformations of the active site loop.[122] The choice of this technique derived from the necessity to investigate fluctuations occurring in a long time scale, difficult to sample through standard Molecular Dynamics simulations. Then, the resultling structures were analyzed through the network analysis, providing a visual and intuitive classification of conformations, difficult to be observed by classical cluster analysis. The 30 most representative conformations, were validated in a retrospective Virtual Screening simulations for their ability to discriminate active from inactive compounds and the most promising structures were selected for a structure-based drug discovery campaign.

This protocol led to the discovery of novel LDHA inhibitors with a micromolar inhibition activity, confirming that taking into account protein flexibility can improve Virtual Screening predictivity in the search for novel active compounds.

Through this protocol, the main atomistic features of the binding pockets and active site loop were explored and a relevant conformation, seed 13, was discovered showing a satisfying recognition capability. However, the validation study with different datasets confirmed that individual conformations were unable to optimally perform with all available ligands and that a multiple receptor approach was effective in discriminating all compounds regardless their chemical features. Based on this retrospective study, three conformations were selected and employed for an Ensemble-based Virtual Screening aimed at discovering novel LDHA inhibitors. The preliminary experimental results led to the discovery of novel LDHA inhibitors in the micromolar range, confirming the success of the described computational protocol in the context of drug discovery. The heterogeneity of the three protein conformations in terms of local rearrangements within the binding site, allowed to explore different pathways aimed to discover novel inhibitors. Even if the most potent hit compounds were related to the crystallographic closed mobile loop state, the other conformations improved the quality of docking results. In fact, analysis of the structural diversity of the hit compounds compared to the known LDHA inhibitors highlighted a good novelty of the Ensemble-based Virtual Screening. Surprisingly, some of the most diverse hit compounds were discovered through the seed 13. In spite of the low LDHA inhibitory activity, these molecules represent the starting point for Structure-Activity Relationship studies, in order to design chemically diverse inhibitors and optimize their pharmacological profile.

In summary, this study highlighted the importance to include flexibility in drug discovery process and how a multiple receptor conformation approach can improve probability to succeed in discovering novel active compounds.

4. OPIOID RECEPTORS - CASE STUDY 2

4.1 OPIOID RECEPTORS

Opium has been used for centuries for pain, diarrhea, cough and other pathologies. It is a mix of substances, including several alkaloids, such as morphine, codeine, papaverine, thebaine and noscapine which control therapeutic effects. Principally, opium owes its analgesic properties to morphine. However, clinical use of the latter has been restricted, due to a wide range of undesirable side effects. Consequently, many efforts have been directed towards the development of novel ligands with a better pharmacologic profile, leading to the discovery of heroine, methadone, meperidine, and several other opiates.[207] Among them, nalorphine was also discovered, able to counter the side effects of the morphine and, at the same time, to act as analgesic agent. It was the starting point for the development of pure antagonists, such as naloxone. The evidence that rigid and non-rigid molecules bound differently to the active site led to suppose the possible existence of multiple opioid receptors.[208] In the mid-1970s, Martin and co-workers confirmed that several opiate compounds showed different pharmacologic profiles in the chronic spinal dog and, consequently, postulated the existence of receptors named after the drugs used in the studies: μ (morphine), κ (ketocyclazocine), and σ (SKF 10,047 or N-allylnormetazocine).[209-211] The latter is no longer classified as opioid receptor. Moreover, the presence of opioid receptors in brain membranes was the prerequisite for the discovery of endogenous ligands, the enkephalins, with morphine-like activity. An observation that the mouse vas deferens had greater affinity for enkephalins than morphine led to the introduction of the δ opioid receptor.[212] Hereafter, the three subtypes are referred to as μ-OR, δ-OR and κ-OR. More recently, a further opioid receptor, the nociceptin/orphanin FQ (N/OFQ) peptide receptor (NOP or ORL-1), was discovered by molecular cloning and characterization of an orphan GPCR.[213] However, it is a non-classical opioid receptor showing a high sequence similarity with classical subtypes (~ 60%), but a very low affinity for classical opioid peptides and most morphine-like small molecules.[214]

Pharmacological data have also revealed different receptor subtypes, namely $\mu_{1\text{-}3}$, $\delta_{1,2}$ and $\kappa_{1\text{-}3}$-receptors.[215-217] However, since only three opioid receptor genes have been identified so far, the existence of these subtypes is explained by two hypotheses: receptor splice variants and receptor heterodimers.[218]

The classical opioid receptors and NOP belong to the G-protein coupled receptors, characterized by the presence of seven transmembrane domains (TM) linked through 3 extracellular and 3 intracellular loops (ECL and ICL, respectively), an extracellular N-terminus and an intracellular C-terminal tail. The sequence identity within the transmembrane helices is about 75%among μ-OR, δ-OR and κ-OR,[213] whereas no homology is found in the extracellular loops and the N-terminal tail. Moreover, two cysteine residues are conserved in all opioid receptors forming a disulphide bond between ECL2 and TM3.

As regards the signal transduction, opioid receptors are coupled to G proteins at the intracellular domain, consisting of three distinct subunits: α, β and γ. Upon agonist binding, opioid receptors undergo conformational changes which promote the exchange of guanosine dihosphate (GDP) bound to the α subunit with guanidine triphosphate (GTP). It results in dissociation of the G protein from the opioid receptor and dissociation of α-GTP from the βγ complex. Then, both subunits can activate/deactivate intracellular effectors by acting on enzymes and ion channels. Conversion of GTP into GDP leads to the reassocciation of the trimeric complex $G_{\alpha\beta\gamma}$ and cessation of signaling.[219] In details, intracellular pathways linked to G subunits concern *i*) the closure of N/P-type voltage-sensitive calcium channels, *ii*) the opening of potassium channels and *iii*) the enzymatic inhibition of the adenylyl cyclase causing block of cyclic adenosine monophosphate (cAMP) production. Opioid receptors cause a negative regulation of calcium channels located at synaptic terminals, so as to prevent neurotransmitter release. In particular, a positive shift in the voltage dependence of the channel coupled to a slowing of activation reduces synaptic transmission. On the other side, $G_{\beta\gamma}$ activates G-protein inwardly rectifying potassium channels and favors membrane hyperpolarization, resulting in decreased neuronal excitability and nociceptive transmission. In addition, inhibition of cAMP production leads to a shift of the threshold of voltage dependent ion channels towards more negative potentials, causing reduction of excitability. This inhibition also reduces neurotransmitter release by acting on cAMP-dependent protein kinase. Figure 27 displays a schematic diagram of the mechanisms described above.

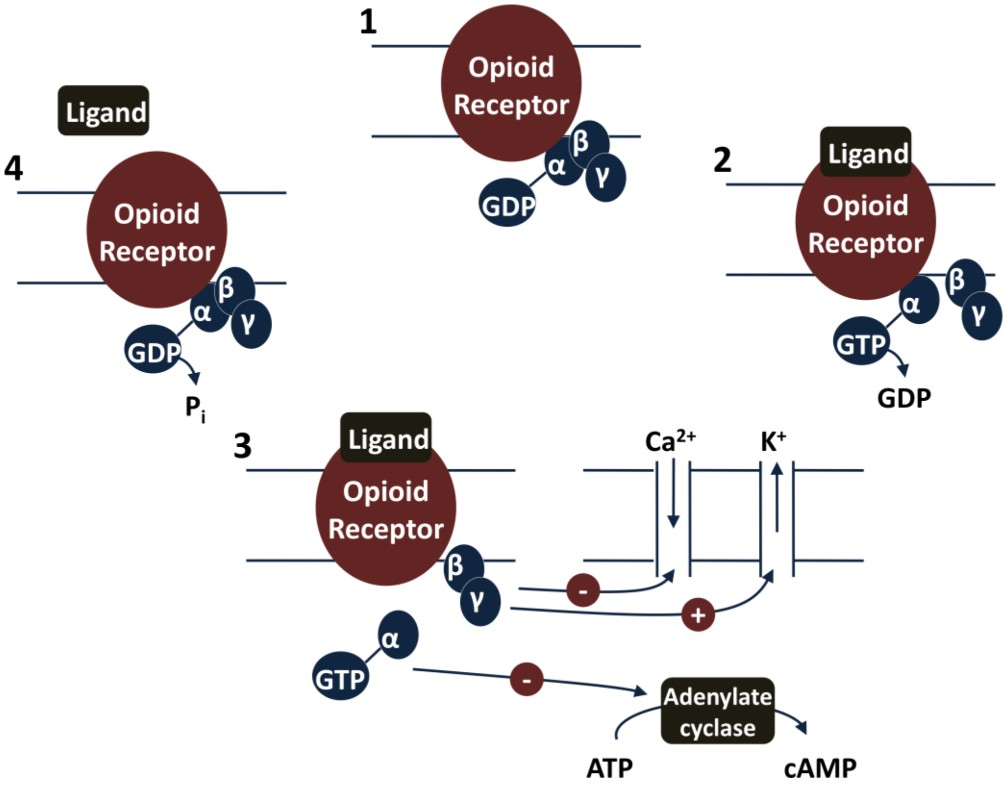


Figure 27. *Opioid receptors and signaling transduction. 1) Opioid receptor is bound to the trimeric form of the protein G, with α subunit linked to GDP. 2) Upon ligand binding, GDP is exchanged for GTP causing dissociation of α subunit from βγ dimer. 3) The two dissociated subunits interact with adenylate cyclase (negative effect), Ca^{2+} channels (negative effect on ion influx) and K$^+$ channels (positive effect on ion efflux), affecting the neurotransmitter release. 4) When GTP is converted to GDP by intrinsic GTPase activity, the system is turned off.*

The main regions containing opioid receptors include the supraspinal and spinal sites, such as the periaqueductal gray, the locus coeruleus and the dorsal horn of the spinal cord which is an important area for opioid-induced analgesia.[219] Moreover, they are located in the peripheral nervous system, for instance primary afferent neurons and dorsal root ganglia. In terms of clinical applications, opioids are used in both acute and chronic pain as effective analgesics. In this scenario, they are administered preemptively or after occurrence of noxious stimulation. The main side effects limiting their use concern the cardiovascular, respiratory and gastrointestinal systems and include also sedation, nausea and vomiting, cough suppression, pupil constriction and skeletal muscle rigidity. The mechanisms behind these effects are multiple. For instance, the inhibitory action of opioids on peristalsis depends on the enteric nerve pathway and the blockade of presynaptic release of acetylcholine.

As regards the chronic application, the most common limitation in the usage of morphine-like ligands is the development of tolerance and dependence and, therefore, loss of opioid receptor-activated function, due to G-protein uncoupling or internalization of the opioid receptors.[220] Tolerance establishes when repeated administration of opioids causes decrease of pharmacological effects. In details, opioid receptor phosphorylation by kinases promotes the binding with arrestin molecules. The steric bulk of the resulting complex prevents the coupling with G proteins and supports the internalization process reducing the number of opioid receptors on the cell surface. However, recent findings suggest that opioid receptor internalization might be a protective mechanism leading to less tolerance, since receptors would be resensitized through the recycling process to the plasma membrane.[221] In general, the development of opioid tolerance is variable depending on the pathological situations.[219] On the other hand, dependence consists of physical and/or psychological symptoms arisen from abrupt interruption of opioid administration, such as mydriasis, diarrhea, restlessness, drug seeking, etc. Administration of opioid antagonists can also cause an immediate precipitated withdrawal. However, the mechanisms underlying these phenomena are not totally understood and are still under investigation.

4.2 STRUCTURAL FEATURES

Recently, high resolution crystal structures of opioid receptors have been published, providing the first evidence for the binding mode of opioids to their counterparts.[214, 222-224] In order to facilitate receptor crystallization, some domains of the opioid receptors have been replaced with another protein, like T4 lysozyme. All the resolved structures bind antagonists; thereby, they are represented as inactive states. The DRY sequence located at the end of the helix 3 is highly conserved in GPCRs and is involved in a ionic-lock, that is a salt bridge between the arginine of this motif and Asp/Glu belonging to the cytoplasmic end of helix VI, which is thought to stabilize the inactive conformation. Although opioid receptors lack of this acidic residue, arginine of DRY motif forms a hydrogen bond with a threonine of helix VI, stabilizing, in any case, the inactive state.[223-224] In general, opioid pharmacology is explained through the message-address theory. The message sequence is aimed to produce the biological response, whereas the address domain is related to binding selectivity. Differences in opioid receptor structures can describe this phenomenon. In particular, the lower binding pocket is highly conserved within this class and recognizes the message of the ligands. Conversely, the upper binding sites show structural diversity and represent the address regions responsible for opioid selectivity. The transmembrane domains of the four opioid receptors are very similar each other, along with the β-hairpin in the ECL2 which is conserved, despite its low sequence identity. Different associations of opioid receptors with each other or other GPCRs to form dimers and oligomers have been observed and can explain changes in signaling properties. The μ-OR crystallizes as a parallel dimer, of which subtypes mainly interact throughTM5 and TM6, whereas the δ-OR forms an anti-parallel arrangement. By contrast, the dimeric form of the κ-OR involves interactions of TM1, TM2 and helix 8. It is important to highlight that the variance of dimeric interfaces should be related to different crystallographic conditions, different T4L arrangements or other reasons. Therefore, conclusions about the physiological relevance of oligomer may be extracted by the analysis of crystal structures with caution.

The recent published μ-OR has been crystallized in complex with the irreversible morphinan antagonist β-FNA involving Lys233$^{5.39}$ (superscripts indicate Ballesteros-Weinstein numbers) for the covalent attachment (PDB ID:4DKL). This ligand is accommodated in a binding pocket exposed to the solvent, probably reflecting the general rapid dissociation half-lives of potent opioids. Figure 28 describes the binding mode of β-FNA to the μ-OR. First of all, the positively charged amine engages a ionic interaction with Asp147$^{3.32}$ and two water molecules generate a hydrogen-bonding network between the phenolic hydroxyl of the morphine group and His297$^{6.52}$. Moreover, Tyr148$^{3.33}$ establishes an hydrogen bond with the furanic oxygen, whereas the cyclopropane reaches the hydrophobic pocket consisting of Tyr326$^{7.43}$ and Trp293$^{6.48}$. In addition, Val300$^{6.55}$ and Ile296$^{6.51}$ are involved in hydrophobic contacts with the morphine scaffold.

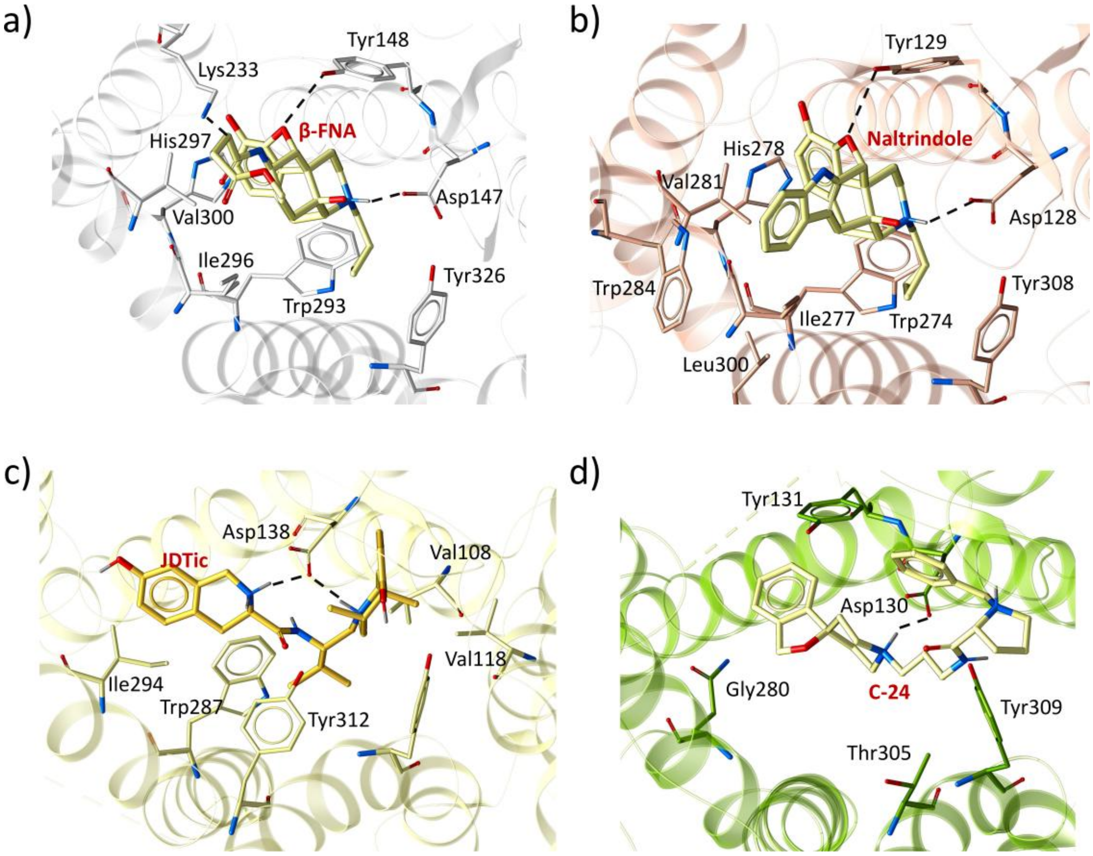


Figure 28. *Schematic representation of interactions between a) β-FNA and the μ-OR, b) Naltrindole and the δ-OR, c) JDTic and the κ-OR and d) C-24 and the χ-OR.*

The crystal structure of δ-OR includes the selective antagonist Naltrindole which is located in an exposed binding pocket (PDB ID:4EJ4). Similarly to the μ-OR, hydrogen bond contacts are conserved, involving His278$^{6.52}$ and Tyr129$^{3.33}$, along with the charge-charge interaction between the positive nitrogen and Asp128$^{3.32}$. Moreover, Naltrindole establishes hydrophobic contacts through Tyr308$^{7.43}$, Trp274$^{6.48}$, Ile277$^{6.51}$, Val281$^{6.55}$, Leu300$^{7.35}$ and Trp284$^{6.58}$. A comparison between μ-OR and δ-OR highlights that the residues within the binding pockets are highly conserved, with the exception of three amino acids. In details, Glu229, Lys303$^{6.58}$ and Trp318$^{7.35}$ correspond to aspartate, tryptophan and leucine in the δ-OR, respectively. In particular, Leu318$^{7.35}$ is in contact with the indole group of Naltrindole and is responsible for the binding selectivity of this opioid antagonist for the δ-OR. In fact, the corresponding amino acids at the same position in the μ-OR and κ-OR are incompatible with Naltrindole binding mode. In other words, the indole group corresponds to the address of this ligand, whereas the morphine scaffold represents the message. Details of the Naltrindole binding mode are shown in Figure 28b.

As regards the κ-OR, JDTic has been used as selective antagonist for the crystal structure (PDB ID:4DJH). Compared to the other subtypes, it shows a displacement of the extracellular half of TM1 which may be related to crystallization conditions. JDTic is accommodated within the binding pocket, where the positively charged nitrogen atoms of piperidine and isoquinoline reach the negative carboxylate group of Asp138$^{3.32}$. The distal hydroxyl groups play an important role for κ-OR affinity. In fact, SAR studies suggest water-mediated hydrogen-bonding networks between these two functional groups and the protein (see Figure 28c for detailed description of contacts). The interactions that are thought to contribute to κ-OR selectivity involve Val108$^{2.53}$, Val118$^{2.63}$, Ile294$^{6.55}$ and Tyr312$^{7.35}$. Moreover, the isopropyl group of JDTic interacts with Trp287$^{6.48}$ through hydrophobic contacts.

Finally, ORL-1 has been cocrystallized with the peptide mimetic antagonist C-24 which mimics the first four amino-terminal residues of the selective peptide antagonist UFP-101[225] and, more in general, represents the message domain of endogenous peptides (PDB ID:4EA3). C-24 interacts through several hydrophobic and electrostatic interactions, as shown in Figure 28d. The protonated nitrogen atom of the ligand interacts with Asp130$^{3.32}$ through the conserved ionic contact. Moreover residues between helices III, V and VI define a hydrophobic pocket in which C-24 is buried through its benzofuran/piperidine rings. Tyr131$^{3.33}$ included in this hydrophobic pocket is involved in a π-stacking interaction with the phenylalanine of the ligand. The reduced affinity of ORL-1 for morphine-based ligands is linked to different residue positions in the binding pocket, such as Ala216$^{5.39}$, Gly280$^{6.52}$ and Thr305$^{7.39}$.

4.3 OPIOID LIGANDS

According to the chemical structures, ligands that bind opioid receptors are differentiated in two classes: peptides and alkaloids.

Met-enkephalin and Leu-enkephalin were the first endogeneous peptides, discovered by Hughes and co-workers in mammalian brain.[226] Thereafter, β-endorphins and dynorphins were isolated from brain, spinal cord and other tissues.[227] In addition, other opioid peptides were discovered and termed endomorphins.[228] The common feature of these ligands is the presence of the tetrapeptide sequence Tyr-Gly-Gly-Phe at their N-terminus, with the exception of endormophins (Tyr-Pro-Trp/Phe-Phe). The structure of peptides includes two components: the message and the address sequences, that is the N-terminal tri- or tetrapeptide and the remaining C-terminal fragment, respectively.[229] In this case, the C-terminal address sequence stabilizes a specific conformation accessible to the N-terminal message sequence. The latter is represented by tyrosine and phenyalanine residues. In details, the amino and phenolic groups of tyrosine and the aromatic ring of phenylalanine are necessary and are optimally separated through glycine residues. Opioid peptides are cleaved from three distinct opioid precursors: proopiomelanocortin for β-endorphins, proenkephalins for Met- and Leu-enkephalins, and prodynorphin for dynorphins. The precursors for endomorphins are yet to be identified. These peptides bind to any of the known opioid receptors with different affinity. Also, they have negligible affinity for ORL-1 receptor. The latter interacts with nociceptin/orphanin FQ that is derived from pro-nociceptin. Compared to the typical opioids described above, nociceptin N-terminal amino acid sequence varies only at position 1, with a phenylalanine instead of the tyrosine.[230]

The second class of opioid ligands include alkaloids, that are mainly classified in agonists and antagonists. The main difference is that agonists bind the receptor conformation in active form, whereas antagonists stabilize the inactive form and interfere with the binding of agonists. The most known opiate alkaloid is the morphine, which is a 7,8-didehydro-4,5-epoxy-17-methyl-(5α,6α)-morphinan-3,6-diol. In the last century, a high number of analogues has been synthesized and proposed as analgesic agents. However the morphine is still used in clinical practice. As regards the antagonists, the most frequently used alkaloids are Naloxone and Naltrexone. In order to introduce novel opioid ligands, two approaches have been proposed including the modification of known morphine analogues and endogenous ligands. Hence, simplification or changes of morphine skeleton has led to the development of new classes of opioid ligands, as reported in Figure 29. In particular, morphinans, benzomorphans, 4-phenylpiperidines, 4-anilinopiperidines, N-benzylpiperazines and methadone-like compounds have been designed and synthesized. Moreover, oripavine compounds include a further ring to the morphine scaffold.

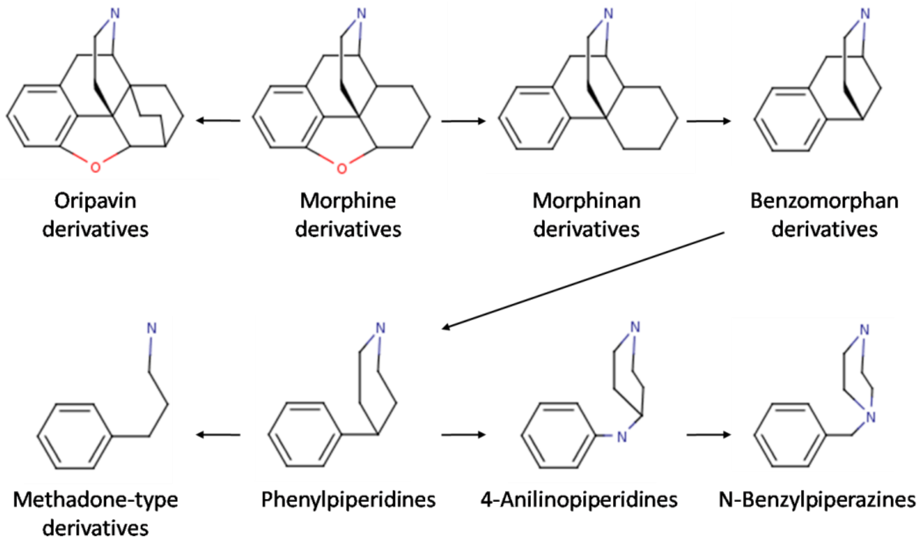


Figure 29. *Opioid derivatives obtained through the modification/simplification of the morphine scaffold.*

Some examples of the well known compounds from each categories are listed as follows: Buprenorphine as oripavine derivative, Codeine as representative of morphine molecules, Levorphanol, Pentazocine, Meperidine and Phentanyl as morphinan, benzomorphan, phenylpiperidine and anilinopiperazine derivative, respectively and Methadone and Propoxyphene belonging to methadone class. The common feature of these ligands is a positively charged nitrogen that interacts with a negatively charged counterpart of the binding site (the highly conserved aspartate residue), as described in the previous section. Moreover, many opioid ligands present a phenolic hydroxyl group separated from the positive charge through six carbon atoms, mimicking the amino-terminal tyrosine of endogenous peptides.

4.4 AIM OF THE PROJECT

Opioid receptors represent the main target for the treatment of pain, cough, diarrhea, and several other diseases. Specifically, each subtypes mediates the action of endogenous and exogenous ligands on specific physiological processes. Therefore, for centuries, they have represented a hot topic of physiological and pharmacological studies. The main limitation in using opiates is that most of the known opiates display side effects which limit their therapeutic use.

The recent publication of the crystal structures for all four opioid receptors has provided a number of insights concerning the binding mode of opioids, so as to well understand the key interactions affecting the action of these drugs. In the context of drug discovery, it represents an incomparable opportunity to explore the structural features to be exploited to develop less toxic therapeutics.

In consideration of the importance of protein flexibility in drug design, the opioid receptor crystal structures offer also the possibility to investigate local and/or global fluctuations that can affect a proper accommodation of ligands within the binding site, as well as conformational variants can couple to different functional pathways. Moreover, protein ensembles can improve the Virtual Screening predictivity in the search for novel compounds.

Exploring the receptor conformations which opioid ligands bind and stabilize is fundamental to design novel active compounds and is at the basis of this work. In particular we used computational tools to define multiple opioid receptor conformations, aimed at exploiting protein flexibility in perspective of a Virtual Screening campaign.

The starting point of this project was a validation study of the opioid crystal structures per se, in order to assess their ability to discriminate known active from inactive compounds. Subsequently, opioid receptor homology models were developed, in order to introduce slightly diverse backbone and side chain orientations. For each opioid receptor, three models were defined by using the remaining isoforms as templates, so resulting in a conformational ensemble consisting of four protein structures (a crystal structure plus three models). Due to the low or even missing affinity of most opioid ligands for NOP receptor, we limited the use of this subtype to the homology modeling and focused our research efforts concerning the validation of resulting models on the μ-OR, δ-OR and κ-OR. For docking study, three different ligand data sets were defined by selecting compounds with experimentally tested antagonist activity against their corresponding opioid receptor subtypes, from the ChEMBL database.[231] In spite of the structural diversity of the three protein ensembles compared to the initial corresponding crystal structures, retrospective Virtual Screening results showed a low recognition performance. Then, we used ALiBERO approach (Automated Ligand-guided Backbone Ensemble Receptor Optimization) to refine the homology models. Through an iterative procedure, ALiBERO led to the definition of three different conformational ensembles, one per opioid subtype, showing improved recognition capability. Optimization of the ALiBERO procedure and the validation of the resulting models are now in progress and are not included in this chapter. However, the preliminary results reported below are a strong evidence of the success of ALiBERO protocol in dealing with local and global flexibility in a drug discovery context.

4.5 RESULTS AND DISCUSSION

The three opioid receptor crystal structures (μ-OR, δ-OR and κ-OR) were used to perform a ligand docking, with ligands data sets from the ChEMBL database. In particular, these collections consisted of active/inactive ratio of 117/350 (μ-OR), 107/228 (δ-OR) and 149/355 (κ-OR). The activity cutoff was chosen based on the availability of experimentally tested compounds: a pK_i and pIC_{50} of 8 was used to select active compounds, whereas activity values less than 6 identified the inactive chemicals. In line with the pharmacological class of the cocrystallized ligands, only antagonist compounds were taken into account. Moreover, we mainly focused our research on alkaloids and derivatives (Figure 29), and discarded large peptides due to their high number of rotatable bonds. Ligand preparation, along with docking studies were carried out with ICM software[143] and details of these procedures are described in the following section. Docking results were analyzed through visual inspection and figures of merit were computed, in order to evaluate the recognition capability of each crystal structure. Figure 30 shows the resulting ROC curves.

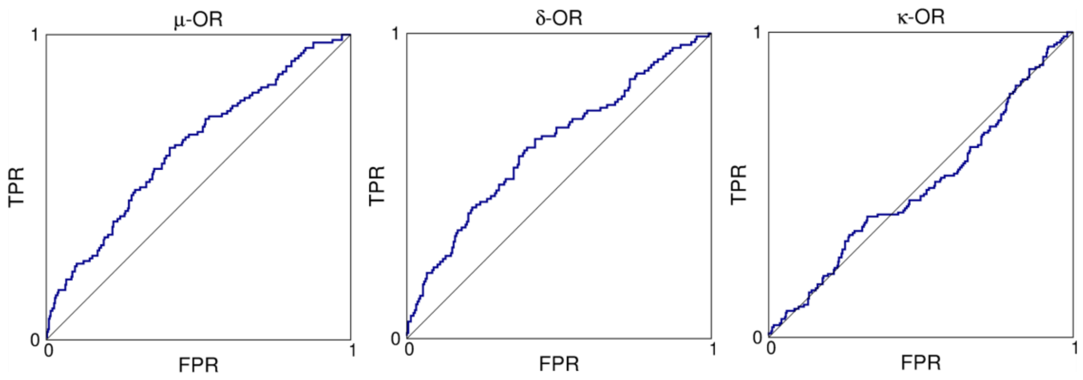


Figure 30. *ROC curves of μ, δ and κ opioid receptor crystal structures.*

A poor performance was observed for the three docking runs, with an Area Under the ROC curve (AUC) above 0.60 for μ and δ opioid receptors and a lower AUC value for the κ-OR (< 0.50). The latter showed a ROC curve approximating the random picking of compounds. Moreover, the slope of the leftmost part of the ROC curve highlighted a negligible initial enrichment.

A visual inspection of docking results highlighted that most of the morphine-like active compounds showed consistent binding modes with conserved interactions, such as the ionic contact between the positively charge nitrogen atom of ligands and the highly conserved aspartate of the opioid receptors. It can be explained by considering that protein structures can have “memory” of their related cocrystallized ligands, resulting in a better score for similar chemicals and a lower score for compounds with different scaffolds. However, ligands interacted with the receptor in a deep but open binding site which allowed both opioid receptor active and inactive compounds to reach a proper geometry and form favorable interactions. Moreover, some inactive compounds are structurally similar to the actives, resulting in a more challenging early recognition.

A limited resolution of protein structures represented another possible cause of low recognition capability. The opioid receptor crystal structures show a low resolution resulting in disordered protein domains and approximate placements of heavy-atoms. High β-factor values of some residues in proximity of the binding site confirm uncertainty about their atom positions. They include the highly conserved residues Asp128$^{3.32}$ and Tyr129$^{3.33}$ belonging to δ-OR, involved in the binding mode of Naltrindole and more in general in opioid ligand binding, or some hydrophobic side chains of the κ-OR binding site (β-factor values higher than 70 in both systems). The low recognition capability of each crystal structure confirmed that, generally, single conformations lack in protein plasticity required to well accommodate and discriminate chemically diverse ligands. Moreover, it is clear that availability of a good conformational ensemble is essential in compound screening, especially for targets characterized by flexible binding pockets and recognizing many different endogenous and exogenous compound chemotypes.

In the light of these preliminary results, we built homology models based on the available crystal structures, aimed to introduce structural diversity in terms of backbone and side chain orientations in the recognition study. In particular, three models were defined for each subtype. Details of this procedure are reported in the next section. A relevant difference between μ-OR and δ-OR compared to κ subtype is the orientation of Tyr$^{7.43}$ and Asp$^{3.32}$ as highlighted in Figure 31. In details, a slight displacement of this tyrosine in the κ-OR leads to a larger binding pocket in respect to the other subtypes. The different volume could be related to the features of the cocrystallized JDTic which is a larger compound compared to the morphine-like cognate ligands of the μ and δ crystal structures. The ORL-1 crystal structure represents an exception; in fact, tyrosine and aspartate side chains define a binding pocket similar to the μ and δ receptors in terms of volume, although the cognate ligand is a large peptide-like compound. This apparent inconsistency is explained by considering that the peptide-like ligand is accommodated in a less deep cavity respect to JDTic. In other words, we can speculate that the side chain displacements of $Tyr^{7.43}$ and $Asp^{3.32}$ allow favorable binding of bulky ligands within an internal cavity of the active site.

The analysis of ligand binding pockets by the ICM PocketFinder algorithm confirmed that κ opioid receptor crystal structure displayed a larger cavity (510.80 $Å^3$ vs. 288.10 and 321 $Å^3$ for μ-OR and δ-OR, respectively). This structural variability has been exploited in the homology modeling in the perspective of an improved recognition capability of known active from inactive chemicals.

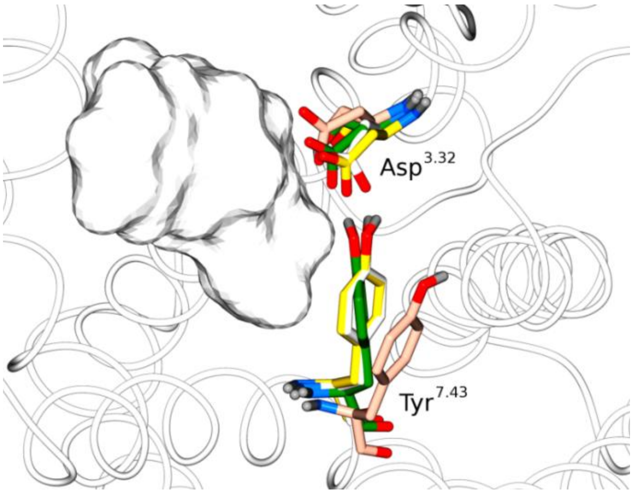


Figure 31. *Different orientation of $Asp^{3.32}$ and $Tyr^{7.43}$ in the opioid receptor crystal structures. Side chains are colored differently based on the protein subtypes: white, yellow, pink and green correspond to μ-OR, δ-OR, κ-OR and ORL-1, respectively. White surface describes the cavity that ligands occupy within the binding site.*

The resulting conformational ensembles included three homology models plus the target crystal structure for each opioid receptor subtype. Hereafter, we refer to the μ-OR, δ-OR, and κ-OR ensembles as HM1, HM2 and HM3, respectively (from Homology Modeling).

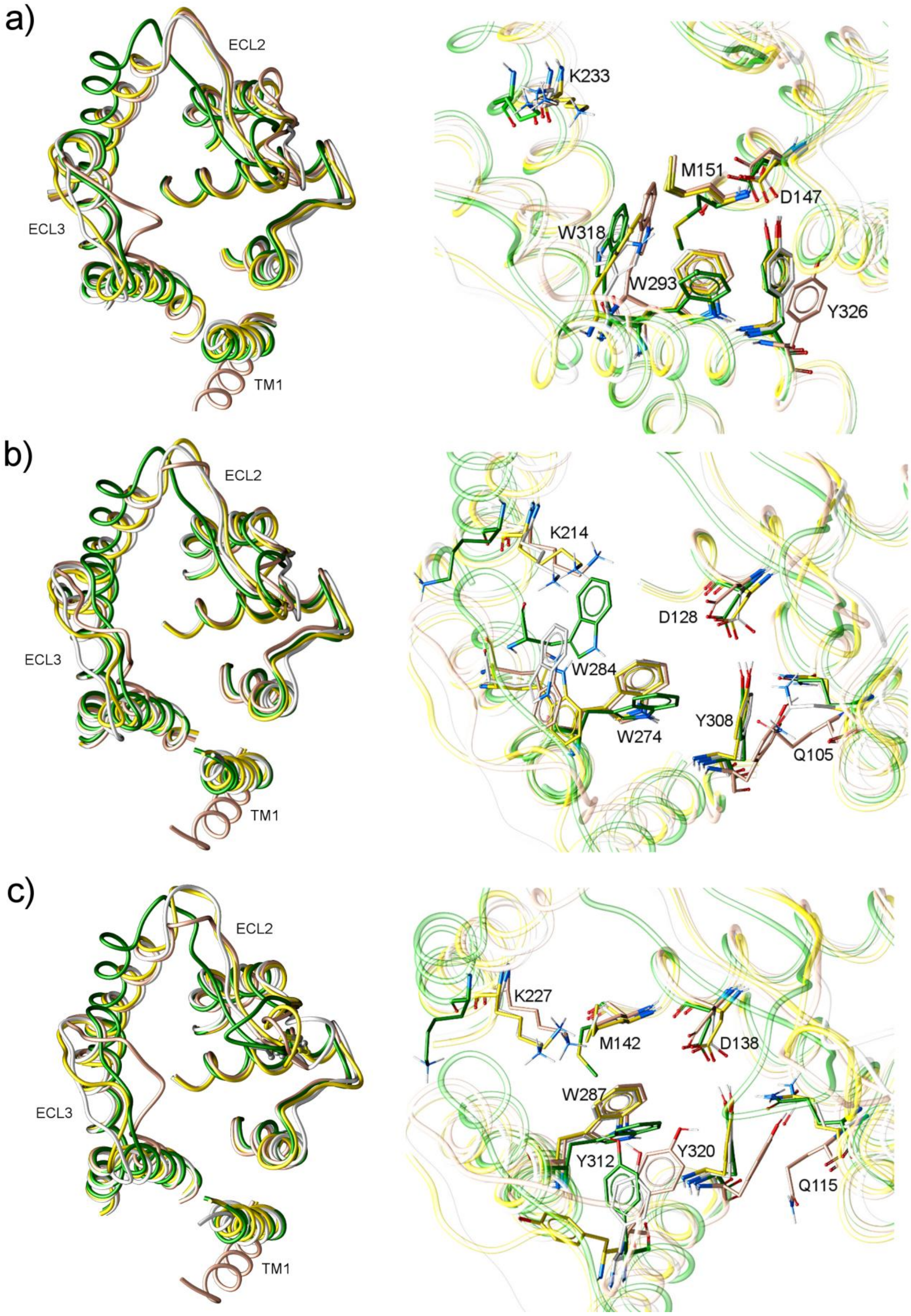


Figure 32. *a) μ-OR ensemble (HM1), b) δ-OR ensemble (HM2) and c) κ-OR ensemble (HM2) include crystal structure and three homology models. Labels indicate the protein domains (left) and residues (right) with the most relevant differences in terms of backbone and side chain orientations, respectively. White, yellow, pink and green representations correspond to crystal structures and homology models of μ-OR, δ-OR, κ-OR and ORL-1, respectively.*

Comparison of the models with the target protein showed that the main structural diversity concerned the second and third extracellular loops which are not directly involved in ligand binding. Moreover, displacement of the first transmembrane domain was observed in protein structures obtained from the κ-OR template. A common difference within HM1, HM2 and HM3 concerned the orientations of $\mathrm{Asp}^{3.32}$ and $\mathrm{Tyr}^{7.43}$. Also, aromatic side chains of tryptophan residues within the active sites were differently oriented and affected the binding mode of opioid ligands. In addition, lysine side chain which is involved in the covalent binding with β-FNA in the μ-OR protein complex, and methionine residues showed diverse geometries within each ensembles. Figure 32a, 32b and 32c summarize the differences of backbone and side chain geometries for each ensemble.

A standard screening run with the ChEMBL data sets described before was performed independently on each ensemble structure. In order to obtain a consistent definition of the binding pocket, residues in the range of 4 Å from all the ligands of the ensemble were taken into account. In this way, common binding pockets were defined within each ensemble, independently from the protein conformation. We evaluated the recognition capability of single ensemble structures and also combined docking results according to a multiple receptor conformation approach. Among the different procedures to combine docking results we used the "best score" approach: we selected the best score for each ligand from the five independent docking runs, resulting in a unique list of docked compounds with relative best score which was used to define the ROC curves. This post-processing step was carried out with ICM tables allowing to store, sort, remove duplicates and, in this specific case, handle large databases. The results of this retrospective Virtual Screening are reported in Table 3.

Ensemble	Subtype	Docking performance (AUC)
HM1	μ-OR*	63.25
	δ-OR	62.27
	κ-OR	44.99
	χ-OR	48.41
	Combination	**58.05**
HM2	μ-OR	55.68
	δ-OR*	63.39
	κ-OR	49.94
	χ-OR	51.72
	Combination	**58.51**
HM3	μ-OR	50.89
	δ-OR	61.42
	κ-OR*	49.42
	χ-OR	49.78
	Combination	**56.74**

Table 3. *Docking performance for HM1, HM2 and HM3 ensembles. Asterisk indicates the crystal structure.*

These results showed that the μ-OR and δ-OR crystal structures better recognized known active compounds compared to the related homology models (HM1 and HM2), whereas κ-OR crystal structure in HM3 and κ-based models in HM1 and HM2 exhibited unsatisfying performance. Although the κ-OR side chain rearrangements, as introduced before, define a larger binding pocket able to accommodate ligands of different size, this large cavity allowed to both active and inactive chemicals to reach favorable binding poses, resulting in poor early recognition ability. Analysis of docking results concerning the κ-based homology model in HM1, confirmed that most of the best ranked compounds were flexible bulky ligands, similar to JDTic and most of them were inactive ligands. In spite of the poor performance, most of morphine-like compounds were properly located within the binding pockets with a consistent binding mode compared to the cocrystallized ligands. It is important to underline that even in presence of an optimal conformational ensemble, docking simulations can fail because of well known limitations and approximations of sampling and scoring steps, as extensively reported in literature.[232-234]

Although protein plasticity was taken into account in ligand docking, this outcome highlighted that the homology models were ill-suited to discriminate compounds and, hence, more permissive fluctuations had to be considered in the definition of protein conformations. Then, the HM1, HM2 and HM3were refined and evaluated by using ALiBERO approach which samples conformational space of side chains and protein backbone within the binding site by means of Elastic Network-Normal Mode Analysis (EN-NMA) and Monte Carlo sampling. [97] Details of this procedure are reported in the next section. ALiBERO approach iteratively defines conformational ensemble with an improved recognition capability against known active and inactive compounds. Starting with HM1, HM2 and HM3, multiple conformations were generated, ligand docking with the selected ChEMBL data sets described before was carried out, and the best performing conformations were combined to define novel protein ensembles. For each protein structure and conformational combination, ROC curves were built and the NSQ_AUC fitness function was also computed. The latter emphasizes the early hit enrichment of screening results and, at the same time, retains contribution for overall selectivity and sensitivity of the model.[150] The protein ensemble consisting of five conformations with the highest NSQ_AUC were used as starting point to sample novel conformations with the EN-NMA method and this procedure was repeated five times to optimize the initial homology models. For the sake of simplicity, we refer to the ALiBERO ensembles defined from HM1 (μ-OR), HM2 (δ-OR) and HM3 (κ-OR), as AL1, AL2 and AL3, respectively. The final AL1, AL2 and AL3 consisted of five conformations, reported in Figure 33 along with the resulting ROC curves.

ALiBERO results showed an improved recognition capability compared to the original homology models and initial crystal structures. In particular, AL1 exhibited AUC and NSQ_AUC values above 0.80 and 0.60 respectively. Also, good results were obtained for AL2 and AL3, with AUC > 0.70 and NSQ_AUC > 0.50.

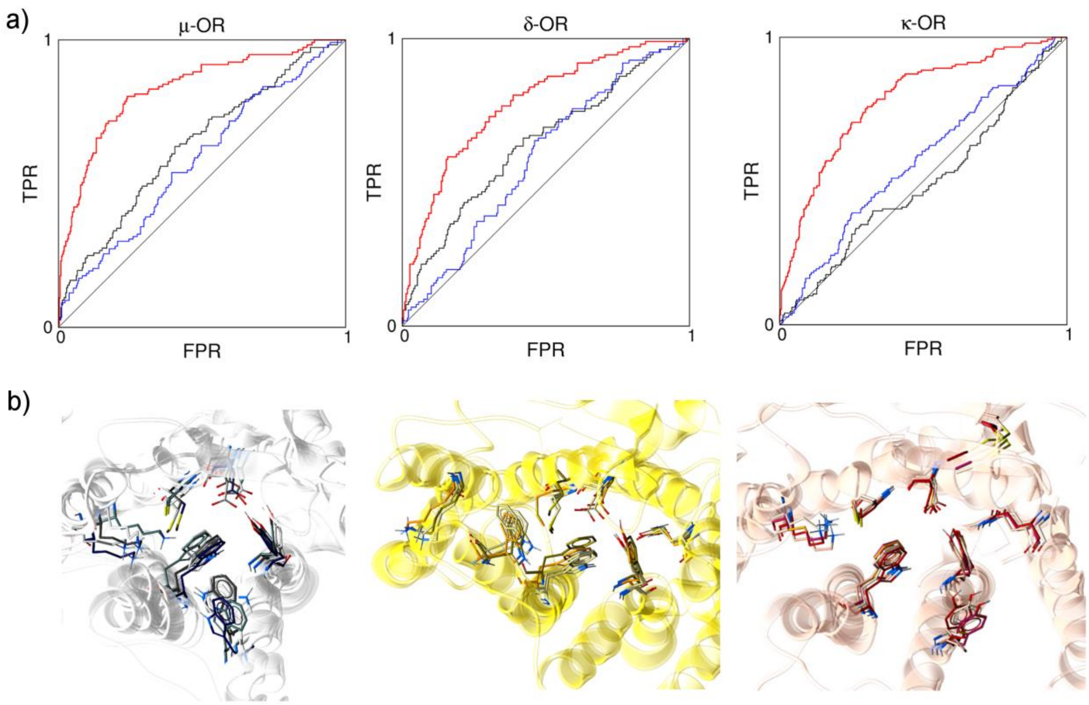


Figure 33. *a) ROC curves of AL1, AL2 and AL3 are represented in red, whereas black and blue ROC curves correspond to the docking results obtained from the crystal structure and the homology models, respectively. b) Different side chain orientations within the binding pocket of AL1, AL2 and AL3.*

The final conformations within AL1, AL2 and AL3 were visual inspected. In details, AL1 consisted of μ-OR conformations with different rearrangements of tryptophan, methionine, tyrosine side chains within the binding pockets. Similar differences were observed for AL2 and AL3. We also found that AL2 conformations were characterized by a significant difference of lysine side chain orientation compared to the other ensembles, resulting in a larger cavity to accommodate bulky ligands. The docked binding modes were analyzed, in order to evaluate if the highly conserved interactions of the most common opioid chemotypes were kept. For example, morphine-like scaffolds were carefully visual inspected and compared to the native poses of the cocrystallized ligands. In general, the majority of docked ligands in AL1, AL2 and AL3 systems were accommodated in the same binding pocket of β-FNA, Naltrindole and JDTic. On the other side, some binding poses were also found in more external cavities, especially when molecules were characterized by large size. For instance, some compounds with bulky functional groups bound to the morphine moiety were not properly accommodate and lost some conserved interactions with the receptor.

Overall, these preliminary results indicated that the binding pockets defined by using ALiBERO protocol displayed good screening capabilities with satisfying performance in the binding mode prediction.

Eventually, the final ensembles will be assessed for their ability to discriminate active from inactive compounds belonging to external unrelated data sets compared to the training set used for the ALiBERO protocol.

4.6 METHODS

4.6.1 LIGAND DOCKING SETUP

In the light of the high chemical diversity of opioid ligands, an homogenous selection of antagonist compounds from the ChEMBL database was carried out. In particular, actives and inactives were clustered based on their chemical structures. First of all, molecules with morphine-like moieties were separated from the remaining ChEMBL antagonists and classified, according to their scaffold, in: oripavines, morphine, morphinane and benzomorphane derivatives, that is 6, 5, 4 and 3 fused ring systems, respectively. The rest of ChEMBL antagonists were classified based on other scaffolds characterizing further opioid ligands, such as phenylpiperidines, 4-anilinopiperidines, etc., as reported in Figure 29. Taking in consideration this classification, μ-OR, δ-OR and κ-OR datasets were defined. Moreover, molecules with a number of rotatable bonds larger than 12 were discarded, in order to consider a reasonable number of degrees of freedom during ligand docking. ICM software was used to define the three dimensional structures. In details, the MMFF types and charges, hydrogen atoms, tautomeric forms were assigned, whereas stereochemistry was retained from the 2D structures. In order to obtain the proper geometry of morphine-like molecules which generally exhibit a three dimensional T-shaped configuration, we improved the ICM tool used for 3D conversions.

Opioid receptors were taken from the Pocketome database, which is a large encyclopedia of experimentally solved conformational ensembles of druggable binding sites in proteins, grouped by location and consistent chain/cofactor composition.[235] When this study was performed, the Pocketome entry for each opioid receptor subtype included one crystal structure. As regards the binding pocket definition, residues within 4.0 Å from the ligands were selected in order to define the boundaries of docking search and the size of the grids. Ligand binding site at the receptor was represented as 0.5 Å spacing potential grid maps describing van der Waals potentials for hydrogens and heavy-atoms, electrostatics, hydrophobicity and hydrogen bonding, as implemented in ICM docking software.[236] The binding energy was defined through the standard empirical scoring function as described in Chapter 2 and docking results were analyzed with ICM tools.

4.6.2 HOMOLOGY MODELING

Ensembles of 3D models of each of μ-OR, δ-OR and κ-OR were built by homology using a hand-curated sequence alignment and the ICM homology modeling routine.[237-238] All opioid receptor structures available when we started this study were used (PDB ID: 4DKL, 4EJ4, 4DJH and 4EA3 corresponding to μ-OR, δ-OR, κ-OR and χ-OR, respectively). Briefly, each model was initially built as a straight ideal geometry amino acid chain. Known disulfide bonds were assigned. For backbone regions conserved between the template and the target sequence, the coordinates of the template were assigned to the model atoms. Short residue insertions and deletions were refined by exhaustive conformational sampling. For longer insertions and poorly conserved loops that are less than 12 residues long, conformational lists were built by searching through a comprehensive library of PDB fragments of similar length and termini topology. Then, energy-based conformational sampling was performed to ensure steric compatibility and favorable packing of each loop with its environment. Each of the resulting models was energy minimized first by thorough conformational sampling of the residue side-chains in internal coordinates with fixed backbone, and then by gradient minimization of all atoms in the presence of distance restraints maintaining the conformational integrity of the model and conserved residue contacts.

4.6.3 ALiBERO SETUP

ALiBERO was used to refine the homology models, in order to obtain novel conformational ensembles in perspective of more accurate docking results. This method is based on the LiBERO framework[239] and is an iterative algorithm consisting of two main steps: *i*) generation of multiple protein conformations and *ii*) flexible-ligand static-receptor small-scale docking on each of the conformers. Then, the best performing children pockets are selected for next generation. This procedure is repeated until a termination condition has been reached, such as the number of iterations or a threshold for a fitness function. In this study, we started with four protein structures for the HM1, HM2 and HM3. In particular, an ALiBERO run consisting of 5 generations was carried out. For each iteration, 100 conformers were generated by using EN-NMA at 300 K and Monte Carlo refinement, in order to promote slight side chain and backbone movements in the range of 1 Å. On every generation, a maximum number of five complementary pocket conformations was retained. The first ten complexes with the best scored active compounds were also subject to Monte Carlo refinement, in order to simulate the induced-by-ligand changes in the side chains. Then, the refined complexes were recombined resulting in optimized pocket ensembles. ALiBERO uses different metrics to evaluate the ability of the pockets to recognize active from inactive compounds. In this case, we used the Normalized SQuare root AUC (NSQ_AUC) which is an innovative fitness function, as described in the Chapter 2. In case the optimized pockets displayed better NSQ_AUC values compared to the previous generation, they were considered successful and represented the "parents" of the next generation.

4.7 CONCLUSIONS

In this study, we performed ligand docking on the opioid receptor crystal structures, in order to investigate the recognition capability of these recent structures. The resulting poor docking performance offered the opportunity to investigate opioid receptor flexibility aimed to properly accommodate ligands within the binding pockets and, then, improve the recognition capability of known active compounds. However, the homology models generated for each subtype still displayed a low discrimination ability. Several computational methods are available to study protein plasticity in ligand docking and, more in general, in drug discovery context. Among them, we selected the EN-NMA and Monte Carlo algorithm to exhaustively sample the protein conformational space and define novel opioid receptor conformations with an improved recognition capability. In details, we used ALiBERO protocol based on these computational approaches to combine the protein conformations, so as to define protein ensembles which maximized the discrimination of known actives from decoys. An interesting improvement of docking performance was obtained by using the ALiBERO protein ensembles. Moreover, these opioid receptor models well predicted atomic contacts with the small molecules. This result confirmed that multiple receptor conformations, defined through a carefully selected computational approach, can introduce a structural variability suitable to discriminate between actives and inactives belonging to diverse chemical classes. More in general, this study highlighted that ALiBERO protocol is a successful tool to introduce protein flexibility in a screening campaign.

It is important to emphasize that ALiBERO results depend on the quality of the initial structures, ligands and parameters used for modeling the system. Therefore, a bad combination of these conditions may result in overfitting producing models with a poor predictivity outside the context of the training set. In order to confirm the quality of the collected conformational ensembles, different ALiBERO runs and comparison of the resulting docking performance will be evaluated. The final ensembles will be assessed for their ability to discriminate active from inactive compounds belonging to external unrelated data sets compared to the training set used for the ALiBERO protocol.

Eventually, the validated ensembles may be used for a selectivity study, in order to investigate the role of protein flexibility to discriminate known chemicals not only within each conformational ensemble, but also among the three opioid receptor subtypes μ, δ and κ. This aspect is very important since it is related to side effects and limited use of the most common opioid therapeutics.

5. CONCLUSIONS AND PERSPECTIVES

Protein receptor flexibility plays a key role in ligand binding. The lock-and-key model introduced by Fischer is based on the assertion that a protein receptor exists as a single conformation. However, with the advent of X-ray crystallography along with sophisticated experimental methods, it has been demonstrated that proteins are flexible entities undergoing local and global fluctuations upon ligand binding. As a consequence, the lock-and-key model was supplanted by the induced fit and conformational selection theories which better describe protein plasticity. According to these theories, the role of protein flexibility has become increasingly relevant in drug discovery and several computational approaches have been developed for this purpose.

The projects presented in this thesis are two example of computational applications taking into account protein flexibility in Virtual Screening. Side chain fluctuations and wide motions of protein domains have been included in screening campaign, in line with the induced fit and conformational selection theories. In particular, we carried out conformational sampling through two different computational tools: MD-based methods and Elastic Network-Normal Mode Analysis with Monte Carlo sampling. Both methods led to exhaustive exploration of conformational space, resulting in the generation of diverse conformations. In both projects, we demonstrated that using single protein structures results in a low performance compared to a multiple conformational approach. To the best of our knowledge, they represent the first applications of multiple receptor conformational approaches in screening campaign concerning LDHA and opioid receptors.

The analysis of the hit compounds from the LDHA Virtual Screening highlighted also an interesting chemical novelty compared to the known inhibitors. The present strategy was used to study a specific biological system characterized by a flexible loop affecting the active site rearrangement. However, it might be applied as general approach to include protein flexibility in Ensemble-based Virtual Screening. The hit compounds obtained from this screening campaign will be optimized, in order to identify novel lead compounds with an improved LDHA inhibitory activity.

As regards the opioid receptor project, ALiBERO approach turned out to be an accurate protocol to refine homology models. This project highlighted that the a careful selection of computational tools is required to define multiple conformations and that local rearrangements introduced through homology modeling might be not sufficient to promote the right accommodations of ligands within the binding site and the early recognition of known active compounds. The promising results suggest that ALiBERO protocol might represent a general tool to assess homology models and crystal structures.

As next step, the opioid receptor conformational ensembles will be optimized and validated, in perspective of a screening campaign. The protein flexibility introduced in this study will be also exploited to study the selectivity of opioid ligands for specific opioid receptor subtypes, which is an important aspect of this class of chemicals to reduce their side effects.

6. ACKNOWLEDGMENT

Vorrei ringraziare innanzitutto il Prof. Maurizio Recanatini che sin dalla tesi di laurea mi ha dato l'opportunità di scoprire il mondo della chimica farmaceutica computazionale. A lui va tutta la mia gratitudine, per aver creduto in me, per avermi dato fiducia affidandomi talvolta compiti di cui non mi sentivo all' altezza. Porterò con me i suoi preziosi insegnamenti professionali e non, e ne farò buon uso nelle nuove esperienze che mi aspettano. Ringrazio il Prof. Andrea Cavalli per le stimolanti discussioni scientifiche e per avermi dato importanti suggerimenti in alcuni momenti decisivi di questo percorso.

Un ringraziamento più che speciale va a Matteo, guida "indispensabile" di questo cammino. Sempre presente nei momenti positivi e negativi, sempre pronto ad aiutarmi quando ce n'è stato bisogno e a spronarmi a credere in me nei momenti più difficili. Ringrazio in particolar modo lui per avermi mostrato cos'è la ricerca e cosa vuol dire avere passione per il proprio lavoro. Sono sicura che qualsiasi esperienza lavorativa intraprenderò, porterò dietro il suo modo di lavorare come esempio per dare il massimo. Ringrazio gli altri ragazzi del QSARLab, chi di passaggio, chi sempre con me durante questi 3 anni. Federico come "distruttore" ufficiale delle mie immagini di VMD, e soprattutto per avermi mostrato il vero modo di approcciarsi al drug discovery. Giampaolo, compagno di università e di dottorato, per il tanto tempo trascorso insieme a divertirci, lavorare e stressarci. Esisterà mai un nuovo compagno di viaggio così? A Maria, la mia prima e unica laureanda. Se oggi è un mostro di bravura non è certo merito mio, ma ringrazio quel "mobile loop" per averle dato la possibilità di entrare in questo mondo. La ringrazio perché assistendola in parte durante la sua tesi ho avuto modo di mettere alla prova le mie conoscenze e di trasmettere il mio sapere.

Ringrazio Luisa perché è stata una persona importante in questo percorso, soprattutto il primo anno di dottorato, per avermi spronato ad andare in America, per essere stata una cara amica dentro e fuori il laboratorio. Infine ringrazio tutti i laureandi che hanno trascorso un periodo in laboratorio, perché anche le più semplici domande di aiuto o i più semplici consigli sono stati un modo per testare la preparazione e trasmettere il piacere per ciò che ho fatto.

Un ringraziamento diverso, ma altrettanto speciale va alle persone che hanno condiviso con me l'esperienza alla UCSD. Prima di tutto, ringrazio il Prof. Ruben Abagyan, perché mi ha dato la possibiltà di trascorrere un periodo presso il suo gruppo, perché mi ha mostrato cos'è la determinazione nel lavoro, cosa vuol dire credere in ciò che si fa. Un

grazie particolare va a Irina, perché mi ha sostenuto nel progetto svolto, mi ha svelato tanti segreti sul magico ICM, perché mi ha messo in contatto con Eugene e per essersi rivelata una splendida persona anche al di fuori delle questioni lavorative. Ringrazio Fiona, per il supporto morale prima di tutto e professionale durante i 7 mesi a San Diego. Infine, ringrazio le mie compagne di ufficio, Clarisse e Claudia, perché hanno avuto la pazienza di sopportare i miei frequenti nervosismi davanti al computer e per i numerosi suggerimenti durante lo svolgimento del progetto.

Mi auguro di trovare lungo il percorso lavorativo persone così speciali, dai professori ai più giovani, e non posso far altro che ringraziare ancora tutti per questo tempo trascorso insieme e per aver reso questo dottorato un’esperienza di vita unica.

7. BIBLIOGRAPHY

1. Fischer, E., Einfluss der Configuration auf die Wirkung der Enzyme. *Berichte der deutschen chemischen Gesellschaft* **1894,** *27* (3), 2985-2993.

2. Henzler-Wildman, K. A.; Thai, V.; Lei, M.; Ott, M.; Wolf-Watz, M.; Fenn, T.; Pozharski, E.; Wilson, M. A.; Petsko, G. A.; Karplus, M.; Hubner, C. G.; Kern, D., Intrinsic motions along an enzymatic reaction trajectory. *Nature* **2007,** *450* (7171), 838-44.

3. Vogt, A. D.; Pozzi, N.; Chen, Z.; Di Cera, E., Essential role of conformational selection in ligand binding. *Biophysical chemistry* **2013**.

4. Hill, T. L., *Free energy transduction in biology: the steady-state kinetic and thermodynamic formalism*. Academic Press: 1977.

5. Koshland, D. E., The Key–Lock Theory and the Induced Fit Theory. *Angewandte Chemie International Edition in English* **1995,** *33* (23-24), 2375-2378.

6. Koshland, D. E., Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proceedings of the National Academy of Sciences* **1958,** *44* (2), 98-104.

7. Frauenfelder, H.; Parak, F.; Young, R. D., Conformational substates in proteins. *Annual review of biophysics and biophysical chemistry* **1988,** *17*, 451-79.

8. Bosshard, H. R., Molecular recognition by induced fit: how fit is the concept? *News in physiological sciences : an international journal of physiology produced jointly by the International Union of Physiological Sciences and the American Physiological Society* **2001,** *16*, 171-3.

9. Monod, J.; Wyman, J.; Changeux, J. P., On the Nature of Allosteric Transitions: A Plausible Model. *Journal of molecular biology* **1965,** *12*, 88-118.

10. Boehr, D. D.; McElheny, D.; Dyson, H. J.; Wright, P. E., The dynamic energy landscape of dihydrofolate reductase catalysis. *Science* **2006,** *313* (5793), 1638-42.

11. Aden, J.; Verma, A.; Schug, A.; Wolf-Watz, M., Modulation of a pre-existing conformational equilibrium tunes adenylate kinase activity. *Journal of the American Chemical Society* **2012,** *134* (40), 16562-70.

12. Honaker, M. T.; Acchione, M.; Zhang, W.; Mannervik, B.; Atkins, W. M., Enzymatic detoxication, conformational selection, and the role of molten globule active sites. *The Journal of biological chemistry* **2013,** *288* (25), 18599-611.

13. Koh, C. Y.; Kim, J. E.; Shibata, S.; Ranade, R. M.; Yu, M.; Liu, J.; Gillespie, J. R.; Buckner, F. S.; Verlinde, C. L.; Fan, E.; Hol, W. G., Distinct states of methionyl-tRNA synthetase indicate inhibitor binding by conformational selection. *Structure* **2012,** *20* (10), 1681-91.

14. Grant, B. J.; Gorfe, A. A.; McCammon, J. A., Large conformational changes in proteins: signaling and other functions. *Current opinion in structural biology* **2010,** *20* (2), 142-7.

15. Ganguly, D.; Zhang, W.; Chen, J., Synergistic folding of two intrinsically disordered proteins: searching for conformational selection. *Molecular bioSystems* **2012,** *8* (1), 198-209.

16. Tsai, C. D.; Ma, B.; Kumar, S.; Wolfson, H.; Nussinov, R., Protein folding: binding of conformationally fluctuating building blocks via population selection. *Critical reviews in biochemistry and molecular biology* **2001,** *36* (5), 399-433.

17. Gronenborn, A. M.; Clore, G. M.; Blazy, B.; Baudras, A., Conformational selection of syn-cAMP upon binding to the cAMP: receptor protein. *FEBS letters* **1981,** *136* (1), 160-4.

18. Birdsall, B.; Feeney, J.; Roberts, G. C.; Burgen, A. S., The use of saturation transfer NMR experiments to monitor the conformational selection accompanying ligand-protein interactions. *FEBS letters* **1980,** *120* (1), 107-9.

19. Boehr, D. D.; Nussinov, R.; Wright, P. E., The role of dynamic conformational ensembles in biomolecular recognition. *Nature chemical biology* **2009,** *5* (11), 789-96.

20. Csermely, P.; Palotai, R.; Nussinov, R., Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. *Trends in biochemical sciences* **2010,** *35* (10), 539-46.

21. Tsai, C. J.; Ma, B.; Nussinov, R., Folding and binding cascades: shifts in energy landscapes. *Proceedings of the National Academy of Sciences of the United States of America* **1999,** *96* (18), 9970-2.

22. Frauenfelder, H.; Sligar, S. G.; Wolynes, P. G., The energy landscapes and motions of proteins. *Science* **1991,** *254* (5038), 1598-603.

23. Rejto, P. A.; Freer, S. T., Protein conformational substates from X-ray crystallography. *Progress in biophysics and molecular biology* **1996,** *66* (2), 167-96.

24. Onuchic, J. N., Contacting the protein folding funnel with NMR. *Proceedings of the National Academy of Sciences of the United States of America* **1997,** *94* (14), 7129-31.

25. Freire, E., Statistical thermodynamic linkage between conformational and binding equilibria. *Advances in protein chemistry* **1998,** *51*, 255-79.

26. Weikl, T. R.; von Deuster, C., Selected-fit versus induced-fit protein binding: kinetic differences and mutational analysis. *Proteins* **2009,** *75* (1), 104-10.

27. Grunberg, R.; Leckner, J.; Nilges, M., Complementarity of structure ensembles in protein-protein binding. *Structure* **2004,** *12* (12), 2125-36.

28. Wlodarski, T.; Zagrovic, B., Conformational selection and induced fit mechanism underlie specificity in noncovalent interactions with ubiquitin. *Proceedings of the National Academy of Sciences of the United States of America* **2009,** *106* (46), 19346-51.

29. Antal, M. A.; Bode, C.; Csermely, P., Perturbation waves in proteins and protein networks: applications of percolation and game theories in signaling and drug design. *Current protein & peptide science* **2009,** *10* (2), 161-72.

30. Bailor, M. H.; Sun, X.; Al-Hashimi, H. M., Topology links RNA secondary structure with global conformation, dynamics, and adaptation. *Science* **2010,** *327* (5962), 202-6.

31. Solomatin, S. V.; Greenfeld, M.; Chu, S.; Herschlag, D., Multiple native states reveal persistent ruggedness of an RNA folding landscape. *Nature* **2010,** *463* (7281), 681-4.

32. Iwahara, J.; Clore, G. M., Detecting transient intermediates in macromolecular binding by paramagnetic NMR. *Nature* **2006,** *440* (7088), 1227-30.

33. Tang, C.; Iwahara, J.; Clore, G. M., Visualization of transient encounter complexes in protein-protein association. *Nature* **2006,** *444* (7117), 383-6.

34. Hatzakis, N. S., Single molecule insights on conformational selection and induced fit mechanism. *Biophysical chemistry* **2013**.

35. Bilderback, T.; Fulmer, T.; Mantulin, W. W.; Glaser, M., Substrate binding causes movement in the ATP binding domain of Escherichia coli adenylate kinase. *Biochemistry* **1996,** *35* (19), 6100-6.

36. Kalinin, S.; Peulen, T.; Sindbert, S.; Rothwell, P. J.; Berger, S.; Restle, T.; Goody, R. S.; Gohlke, H.; Seidel, C. A., A toolkit and benchmark study for FRET-restrained high-precision structural modeling. *Nature methods* **2012,** *9* (12), 1218-25.

37. Lu, H. P.; Xun, L.; Xie, X. S., Single-molecule enzymatic dynamics. *Science* **1998,** *282* (5395), 1877-82.

38. Hatzakis, N. S.; Wei, L.; Jorgensen, S. K.; Kunding, A. H.; Bolinger, P. Y.; Ehrlich, N.; Makarov, I.; Skjot, M.; Svendsen, A.; Hedegard, P.; Stamou, D., Single enzyme studies reveal the existence of discrete functional states for monomeric enzymes and how they are "selected" upon allosteric regulation. *Journal of the American Chemical Society* **2012,** *134* (22), 9296-302.

39. Teague, S. J., Implications of protein flexibility for drug discovery. *Nature reviews. Drug discovery* **2003,** *2* (7), 527-41.

40. Carlson, H. A., Protein flexibility and drug design: how to hit a moving target. *Current opinion in chemical biology* **2002,** *6* (4), 447-52.

41. DePristo, M. A.; de Bakker, P. I.; Blundell, T. L., Heterogeneity and inaccuracy in protein structures solved by X-ray crystallography. *Structure* **2004,** *12* (5), 831-8.

42. Keseru, G. M.; Kolossvary, I., Fully flexible low-mode docking: application to induced fit in HIV integrase. *Journal of the American Chemical Society* **2001,** *123* (50), 12708-9.

43. Mangoni, M.; Roccatano, D.; Di Nola, A., Docking of flexible ligands to flexible receptors in solution by molecular dynamics simulation. *Proteins* **1999,** *35* (2), 153-62.

44. Sinko, W.; Lindert, S.; McCammon, J. A., Accounting for receptor flexibility and enhanced sampling methods in computer-aided drug design. *Chemical biology & drug design* **2013,** *81* (1), 41-9.

45. Thorpe, M. F.; Lei, M.; Rader, A. J.; Jacobs, D. J.; Kuhn, L. A., Protein flexibility and dynamics using constraint theory. *Journal of molecular graphics & modelling* **2001,** *19* (1), 60-9.

46. Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E., How fast-folding proteins fold. *Science* **2011,** *334* (6055), 517-20.

47. Shan, Y.; Kim, E. T.; Eastwood, M. P.; Dror, R. O.; Seeliger, M. A.; Shaw, D. E., How does a drug molecule find its target binding site? *Journal of the American Chemical Society* **2011,** *133* (24), 9181-3.

48. Lindorff-Larsen, K.; Best, R. B.; Depristo, M. A.; Dobson, C. M.; Vendruscolo, M., Simultaneous determination of protein structure and dynamics. *Nature* **2005,** *433* (7022), 128-32.

49. Andrusier, N.; Mashiach, E.; Nussinov, R.; Wolfson, H. J., Principles of flexible protein-protein docking. *Proteins* **2008,** *73* (2), 271-89.

50. Jiang, F.; Kim, S. H., "Soft docking": matching of molecular surface cubes. *Journal of molecular biology* **1991,** *219* (1), 79-102.

51. Jones, G.; Willett, P.; Glen, R. C., Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *Journal of molecular biology* **1995,** *245* (1), 43-53.

52. Leach, A. R., Ligand docking to proteins with discrete side-chain flexibility. *Journal of molecular biology* **1994,** *235* (1), 345-56.

53. Schaffer, L.; Verkhivker, G. M., Predicting structural effects in HIV-1 protease mutant complexes with flexible ligand docking and protein side-chain optimization. *Proteins* **1998,** *33* (2), 295-310.

54. Carlson, H. A.; McCammon, J. A., Accommodating Protein Flexibility in Computational Drug Design. *Molecular Pharmacology* **2000,** *57* (2), 213-218.

55. Sherman, W.; Day, T.; Jacobson, M. P.; Friesner, R. A.; Farid, R., Novel procedure for modeling ligand/receptor induced fit effects. *Journal of medicinal chemistry* **2006,** *49* (2), 534-53.

56. Sherman, W.; Beard, H. S.; Farid, R., Use of an Induced Fit Receptor Structure in Virtual Screening. *Chemical biology & drug design* **2006,** *67* (1), 83-84.

57. Cavasotto, C. N.; Abagyan, R. A., Protein flexibility in ligand docking and virtual screening to protein kinases. *Journal of molecular biology* **2004,** *337* (1), 209-25.

58. Meiler, J.; Baker, D., ROSETTALIGAND: protein-small molecule docking with full side-chain flexibility. *Proteins* **2006,** *65* (3), 538-48.

59. Davis, I. W.; Baker, D., RosettaLigand docking with full ligand and receptor flexibility. *Journal of molecular biology* **2009,** *385* (2), 381-92.

60. Lemmon, G.; Meiler, J., Rosetta Ligand docking with flexible XML protocols. *Methods in molecular biology* **2012,** *819*, 143-55.

61. Claussen, H.; Buning, C.; Rarey, M.; Lengauer, T., FlexE: efficient molecular docking considering protein structure variations. *Journal of molecular biology* **2001,** *308* (2), 377-95.

62. Bottegoni, G.; Kufareva, I.; Totrov, M.; Abagyan, R., Four-dimensional docking: a fast and accurate account of discrete receptor flexibility in ligand docking. *Journal of medicinal chemistry* **2009,** *52* (2), 397-406.

63. Knegtel, R. M.; Kuntz, I. D.; Oshiro, C. M., Molecular docking to ensembles of protein structures. *Journal of molecular biology* **1997,** *266* (2), 424-40.

64. Craig, I. R.; Essex, J. W.; Spiegel, K., Ensemble docking into multiple crystallographically derived protein structures: an evaluation based on the statistical analysis of enrichments. *Journal of chemical information and modeling* **2010,** *50* (4), 511-24.

65. Rueda, M.; Bottegoni, G.; Abagyan, R., Recipes for the selection of experimental protein conformations for virtual screening. *Journal of chemical information and modeling* **2010,** *50* (1), 186-93.

66. Novoa, E. M.; Pouplana, L. R. d.; Barril, X.; Orozco, M., Ensemble Docking from Homology Models. *Journal of Chemical Theory and Computation* **2010,** *6* (8), 2547-2557.

67. Lin, J. H.; Perryman, A. L.; Schames, J. R.; McCammon, J. A., Computational drug design accommodating receptor flexibility: the relaxed complex scheme. *Journal of the American Chemical Society* **2002,** *124* (20), 5632-3.

68. Lin, J. H.; Perryman, A. L.; Schames, J. R.; McCammon, J. A., The relaxed complex method: Accommodating receptor flexibility for drug design with an improved scoring scheme. *Biopolymers* **2003,** *68* (1), 47-62.

69. Nichols, S. E.; Baron, R.; Ivetac, A.; McCammon, J. A., Predictive power of molecular dynamics receptor structures in virtual screening. *Journal of chemical information and modeling* **2011,** *51* (6), 1439-46.

70. Carlson, H. A.; Masukawa, K. M.; Rubins, K.; Bushman, F. D.; Jorgensen, W. L.; Lins, R. D.; Briggs, J. M.; McCammon, J. A., Developing a dynamic pharmacophore model for HIV-1 integrase. *Journal of medicinal chemistry* **2000,** *43* (11), 2100-14.

71. Alonso, H.; Bliznyuk, A. A.; Gready, J. E., Combining docking and molecular dynamic simulations in drug design. *Medicinal research reviews* **2006,** *26* (5), 531-68.

72. Shaw, D. E.; Dror, R. O.; Salmon, J. K.; Grossman, J. P.; Mackenzie, K. M.; Bank, J. A.; Young, C.; Deneroff, M. M.; Batson, B.; Bowers, K. J.; Chow, E.; Eastwood, M. P.; Ierardi, D. J.; Klepeis, J. L.; Kuskin, J. S.; Larson, R. H.; Lindorff-Larsen, K.; Maragakis, P.; Moraes, M. A.; Piana, S.; Shan, Y.; Towles, B., Millisecond-scale molecular dynamics simulations on Anton. In *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis*, ACM: Portland, Oregon, **2009;** 1-11.

73. Torrie, G. M.; Valleau, J. P., Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics* **1977,** *23* (2), 187-199.

74. Barducci, A.; Bonomi, M.; Parrinello, M., Metadynamics. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2011,** *1* (5), 826-843.

75. Sugita, Y.; Okamoto, Y., Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters* **1999,** *314* (1–2), 141-151.

76. Hamelberg, D.; Mongan, J.; McCammon, J. A., Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. *The Journal of chemical physics* **2004,** *120* (24), 11919-29.

77. Zhou, T.; Caflisch, A., Free Energy Guided Sampling. *Journal of Chemical Theory and Computation* **2012,** *8* (6), 2134-2140.

78. Krivov, S. V.; Karplus, M., One-Dimensional Free-Energy Profiles of Complex Systems: Progress Variables that Preserve the Barriers. *The Journal of Physical Chemistry B* **2006,** *110* (25), 12689-12698.

79. Prinz, J. H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Schutte, C.; Noe, F., Markov models of molecular kinetics: generation and validation. *The Journal of chemical physics* **2011,** *134* (17), 174105.

80. Brooks, B.; Karplus, M., Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor. *Proceedings of the National Academy of Sciences of the United States of America* **1983,** *80* (21), 6571-5.

81. Wilson, E. B.; Decius, J. C.; Cross, P. C., *Molecular Vibrations: The Theory of Infrared and Raman Vibrational Spectra*. Dover Publications: **1955**.

82. Tirion, M. M., Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. *Physical review letters* **1996,** *77* (9), 1905-1908.

83. May, A.; Zacharias, M., Protein-ligand docking accounting for receptor side chain and global flexibility in normal modes: evaluation on kinase inhibitor cross docking. *Journal of medicinal chemistry* **2008,** *51* (12), 3499-506.

84. Hinsen, K., Analysis of domain motions by approximate normal mode calculations. *Proteins* **1998,** *33* (3), 417-29.

85. Cavasotto, C. N.; Kovacs, J. A.; Abagyan, R. A., Representing receptor flexibility in ligand docking through relevant normal modes. *Journal of the American Chemical Society* **2005,** *127* (26), 9632-40.

86. Rueda, M.; Bottegoni, G.; Abagyan, R., Consistent improvement of cross-docking results using binding site ensembles generated with elastic network normal modes. *Journal of chemical information and modeling* **2009,** *49* (3), 716-25.

87. Barril, X.; Fradera, X., Incorporating protein flexibility into docking and structure-based drug design. *Expert Opinion on Drug Discovery* **2006,** *1* (4), 335-349.

88. Yuriev, E.; Ramsland, P. A., Latest developments in molecular docking: 2010-2011 in review. *Journal of molecular recognition : JMR* **2013,** *26* (5), 215-39.

89. Durrant, J. D.; McCammon, J. A., Computer-aided drug-discovery techniques that account for receptor flexibility. *Current opinion in pharmacology* **2010,** *10* (6), 770-4.

90. Feixas, F.; Lindert, S.; Sinko, W.; McCammon, J. A., Exploring the role of receptor flexibility in structure-based drug discovery. *Biophysical chemistry* **2013**.

91. Amaro, R. E.; Baron, R.; McCammon, J. A., An improved relaxed complex scheme for receptor flexibility in computer-aided drug design. *Journal of computer-aided molecular design* **2008,** *22* (9), 693-705.

92. Shao, J.; Tanner, S. W.; Thompson, N.; Cheatham, T. E., Clustering Molecular Dynamics Trajectories: 1. Characterizing the Performance of Different Clustering Algorithms. *Journal of Chemical Theory and Computation* **2007,** *3* (6), 2312-2334.

93. Hayward, S.; de Groot, B. L., Normal modes and essential dynamics. *Methods in molecular biology* **2008,** *443*, 89-106.

94. Balsera, M. A.; Wriggers, W.; Oono, Y.; Schulten, K., Principal Component Analysis and Long Time Protein Dynamics. *The Journal of Physical Chemistry* **1996,** *100* (7), 2567-2572.

95. Campbell, Z. T.; Baldwin, T. O.; Miyashita, O., Analysis of the bacterial luciferase mobile loop by replica-exchange molecular dynamics. *Biophysical journal* **2010,** *99* (12), 4012-9.

96. Truchon, J. F.; Bayly, C. I., Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. *Journal of chemical information and modeling* **2007,** *47* (2), 488-508.

97. Rueda, M.; Totrov, M.; Abagyan, R., ALiBERO: evolving a team of complementary pocket conformations rather than a single leader. *Journal of chemical information and modeling* **2012,** *52* (10), 2705-14.

98. Irwin, J. J.; Shoichet, B. K., ZINC--a free database of commercially available compounds for virtual screening. *Journal of chemical information and modeling* **2005,** *45* (1), 177-82.

99. Lipinski, C. A., Lead- and drug-like compounds: the rule-of-five revolution. *Drug Discovery Today: Technologies* **2004,** *1* (4), 337-341.

100. Baell, J. B.; Holloway, G. A., New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *Journal of medicinal chemistry* **2010,** *53* (7), 2719-40.

101. McGovern, S. L.; Caselli, E.; Grigorieff, N.; Shoichet, B. K., A common mechanism underlying promiscuous inhibitors from virtual and high-throughput screening. *Journal of medicinal chemistry* **2002,** *45* (8), 1712-22.

102. Gohlke, H.; Kiel, C.; Case, D. A., Insights into protein-protein binding by binding free energy calculation and free energy decomposition for the Ras-Raf and Ras-RalGDS complexes. *Journal of molecular biology* **2003,** *330* (4), 891-913.

103. Massova, I.; Kollman, P., Combined molecular mechanical and continuum solvent approach (MM-PBSA/GBSA) to predict ligand binding. *Perspectives in Drug Discovery and Design* **2000,** *18* (1), 113-135.

104. Rastelli, G.; Del Rio, A.; Degliesposti, G.; Sgobba, M., Fast and accurate predictions of binding free energies using MM-PBSA and MM-GBSA. *J Comput Chem* **2010,** *31* (4), 797-810.

105. Buonfiglio, R.; Ferraro, M.; Falchi, F.; Cavalli, A.; Masetti, M.; Recanatini, M., Collecting and assessing human lactate dehydrogenase-A conformations for structure-based virtual screening. *Journal of chemical information and modeling* **2013,** *53* (11), 2792-7.

106. Born, M.; Oppenheimer, R., Zur Quantentheorie der Molekeln. *Annalen der Physik* **1927,** *389* (20), 457-484.

107. Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C., Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics* **1977,** *23* (3), 327-341.

108. Verlet, L., Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Physical Review* **1967,** *159* (1), 98-103.

109. Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R., Molecular dynamics with coupling to an external bath. *The Journal of chemical physics* **1984,** *81* (8), 3684-3690.

110. Feller, S. E.; Zhang, Y.; Pastor, R. W.; Brooks, B. R., Constant pressure molecular dynamics simulation: The Langevin piston method. *The Journal of chemical physics* **1995,** *103* (11), 4613-4621.

111. Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E., Equation of State Calculations by Fast Computing Machines. *The Journal of chemical physics* **1953,** *21* (6), 1087-1092.

112. Rao, M.; Pangali, C.; Berne, B. J., On the force bias Monte Carlo simulation of water: methodology, optimization and comparison with molecular dynamics. *Molecular Physics* **1979,** *37* (6), 1773-1798.

113. Earl, D. J.; Deem, M. W., Parallel tempering: theory, applications, and new perspectives. *Physical chemistry chemical physics : PCCP* **2005,** *7* (23), 3910-6.

114. Swendsen, R. H.; Wang, J. S., Replica Monte Carlo simulation of spin glasses. *Physical review letters* **1986,** *57* (21), 2607-2609.

115. Hansmann, U. H. E., Parallel tempering algorithm for conformational studies of biological molecules. *Chemical Physics Letters* **1997,** *281* (1–3), 140-150.

116. Kofke, D. A., On the acceptance probability of replica-exchange Monte Carlo trials. *The Journal of chemical physics* **2002,** *117* (15), 6911-6914.

117. Kofke, D. A., Erratum: "On the acceptance probability of replica-exchange Monte Carlo trials" [J. Chem. Phys. 117, 6911 (2002)]. *The Journal of chemical physics* **2004,** *120* (22), 10852-10852.

118. Rathore, N.; Chopra, M.; de Pablo, J. J., Optimal allocation of replicas in parallel tempering simulations. *The Journal of chemical physics* **2005,** *122* (2), 024111.

119. Kone, A.; Kofke, D. A., Selection of temperature intervals for parallel-tempering simulations. *The Journal of chemical physics* **2005,** *122* (20), 206101.

120. Cheng, X.; Cui, G.; Hornak, V.; Simmerling, C., Modified replica exchange simulation methods for local structure refinement. *The journal of physical chemistry. B* **2005,** *109* (16), 8220-30.

121. Jang, S.; Shin, S.; Pak, Y., Replica-Exchange Method Using the Generalized Effective Potential. *Physical review letters* **2003,** *91* (5), 058305.

122. Okur, A.; Wickstrom, L.; Layten, M.; Geney, R.; Song, K.; Hornak, V.; Simmerling, C., Improved Efficiency of Replica Exchange Simulations through Use of a Hybrid Explicit/Implicit Solvation Model. *Journal of Chemical Theory and Computation* **2006,** *2* (2), 420-433.

123. Go, N., A theorem on amplitudes of thermal atomic fluctuations in large molecules assuming specific conformations calculated by normal mode analysis. *Biophysical chemistry* **1990,** *35* (1), 105-12.

124. Jain, A. K.; Murty, M. N.; Flynn, P. J., Data clustering: a review. *ACM Comput. Surv.* **1999,** *31* (3), 264-323.

125. Torda, A. E.; Gunsteren, W. F. v., Algorithms for clustering molecular dynamics configurations. *J. Comput. Chem.* **1994,** *15* (12), 1331-1340.

126. MacQueen, J. In *Some methods for classification and analysis of multivariate observations*, Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, Berkeley, Calif., 1967; University of California Press: Berkeley, Calif., **1967**; 281-297.

127. Downs, G. M.; Barnard, J. M., Clustering Methods and Their Uses in Computational Chemistry. In *Reviews in Computational Chemistry*, John Wiley & Sons, Inc.: 2003; pp 1-40.

128. Berry, M. J.; Linoff, G., *Data Mining Techniques: For Marketing, Sales, and Customer Support*. John Wiley \&amp; Sons, Inc.: **1997**; 444.

129. Davies, D. L.; Bouldin, D. W., A Cluster Separation Measure. *IEEE Trans. Pattern Anal. Mach. Intell.* **1979,** *1* (2), 224-227.

130. Mitchell, T. M., *Machine Learning*. McGraw-Hill, Inc.: **1997**; 432.

131. Tibshirani, R.; Walther, G.; Hastie, T., Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **2001,** *63* (2), 411-423.

132. Shepard, R., The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika* **1962,** *27* (2), 125-140.

133. Shepard, R., The analysis of proximities: Multidimensional scaling with an unknown distance function. II. *Psychometrika* **1962,** *27* (3), 219-246.

134. Barabasi, A. L.; Oltvai, Z. N., Network biology: understanding the cell's functional organization. *Nature reviews. Genetics* **2004,** *5* (2), 101-13.

135. Erdos, P.; Renyi, A., {On the evolution of random graphs}. *Publ. Math. Inst. Hung. Acad. Sci* **1960,** *5*, 17-61.

136. Gfeller, D.; De Los Rios, P.; Caflisch, A.; Rao, F., Complex network analysis of free-energy landscapes. *Proceedings of the National Academy of Sciences of the United States of America* **2007,** *104* (6), 1817-22.

137. Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J., Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature reviews. Drug discovery* **2004,** *3* (11), 935-49.

138. Leach, A., *Molecular Modelling: Principles and Applications (2nd Edition)*. Prentice Hall: **2001**.

139. Goldberg, D. E., *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley Longman Publishing Co., Inc.: **1989**; 372.

140. Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G., A fast flexible docking method using an incremental construction algorithm. *Journal of molecular biology* **1996,** *261* (3), 470-89.

141. Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S., Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *Journal of medicinal chemistry* **2004,** *47* (7), 1739-49.

142. Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L., Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *Journal of medicinal chemistry* **2004,** *47* (7), 1750-9.

143. Abagyan, R.; Totrov, M.; Kuznetsov, D., ICM—A new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation. *Journal of Computational Chemistry* **1994,** *15* (5), 488-506.

144. Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P., Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *Journal of computer-aided molecular design* **1997,** *11* (5), 425-45.

145. Nemethy, G.; Gibson, K. D.; Palmer, K. A.; Yoon, C. N.; Paterlini, G.; Zagari, A.; Rumsey, S.; Scheraga, H. A., Energy parameters in polypeptides. 10. Improved geometrical parameters and nonbonded interactions for use in the ECEPP/3 algorithm, with application to proline-containing peptides. *The Journal of Physical Chemistry* **1992,** *96* (15), 6472-6484.

146. Neves, M. A.; Totrov, M.; Abagyan, R., Docking and scoring with ICM: the benchmarking results and strategies for improvement. *Journal of computer-aided molecular design* **2012,** *26* (6), 675-86.

147. Triballeau, N.; Acher, F.; Brabet, I.; Pin, J. P.; Bertrand, H. O., Virtual screening workflow development guided by the "receiver operating characteristic" curve approach. Application to high-throughput docking on metabotropic glutamate receptor subtype 4. *Journal of medicinal chemistry* **2005,** *48* (7), 2534-47.

148. Bamber, D., The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology* **1975,** *12* (4), 387-415.

149. Kirchmair, J.; Markt, P.; Distinto, S.; Wolber, G.; Langer, T., Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy selection--what can we learn from earlier mistakes? *Journal of computer-aided molecular design* **2008,** *22* (3-4), 213-28.

150. Katritch, V.; Kufareva, I.; Abagyan, R., Structure based prediction of subtype-selectivity for adenosine receptor antagonists. *Neuropharmacology* **2011,** *60* (1), 108-15.

151. Sheridan, R. P.; Singh, S. B.; Fluder, E. M.; Kearsley, S. K., Protocols for bridging the peptide to nonpeptide gap in topological similarity searches. *Journal of chemical information and computer sciences* **2001,** *41* (5), 1395-406.

152. Rolland, F.; Winderickx, J.; Thevelein, J. M., Glucose-sensing and -signalling mechanisms in yeast. *FEMS yeast research* **2002,** *2* (2), 183-201.

153. Bryant, C.; Voller, A.; Smith, M. J., The Incorporation of Radioactivity from (14c)Glucose into the Soluble Metabolic Intermediates of Malaria Parasites. *The American journal of tropical medicine and hygiene* **1964,** *13*, 515-9.

154. Warburg, O., On the origin of cancer cells. *Science* **1956,** *123* (3191), 309-14.

155. Zu, X. L.; Guppy, M., Cancer metabolism: facts, fantasy, and fiction. *Biochemical and biophysical research communications* **2004,** *313* (3), 459-65.

156. Griguer, C. E.; Oliva, C. R.; Gillespie, G. Y., Glucose metabolism heterogeneity in human and mouse malignant glioma cell lines. *Journal of neuro-oncology* **2005,** *74* (2), 123-33.

157. Miccheli, A.; Tomassini, A.; Puccetti, C.; Valerio, M.; Peluso, G.; Tuccillo, F.; Calvani, M.; Manetti, C.; Conti, F., Metabolic profiling by 13C-NMR spectroscopy: [1,2-13C2]glucose reveals a heterogeneous metabolism in human leukemia T cells. *Biochimie* **2006,** *88* (5), 437-48.

158. DeBerardinis, R. J.; Cheng, T., Q's next: the diverse functions of glutamine in metabolism, cell biology and cancer. *Oncogene* **2010,** *29* (3), 313-24.

159. Pfeiffer, T.; Schuster, S.; Bonhoeffer, S., Cooperation and competition in the evolution of ATP-producing pathways. *Science* **2001,** *292* (5516), 504-7.

160. Zheng, J., Energy metabolism of cancer: Glycolysis versus oxidative phosphorylation (Review). *Oncology letters* **2012,** *4* (6), 1151-1157.

161. Vaupel, P., Metabolic microenvironment of tumor cells: a key factor in malignant progression. *Experimental oncology* **2010,** *32* (3), 125-7.

162. Chen, Z.; Lu, W.; Garcia-Prieto, C.; Huang, P., The Warburg effect and its cancer therapeutic implications. *Journal of bioenergetics and biomembranes* **2007,** *39* (3), 267-74.

163. Carew, J. S.; Nawrocki, S. T.; Xu, R. H.; Dunner, K.; McConkey, D. J.; Wierda, W. G.; Keating, M. J.; Huang, P., Increased mitochondrial biogenesis in primary

leukemia cells: the role of endogenous nitric oxide and impact on sensitivity to fludarabine. *Leukemia* **2004,** *18* (12), 1934-40.

164. Carew, J. S.; Zhou, Y.; Albitar, M.; Carew, J. D.; Keating, M. J.; Huang, P., Mitochondrial DNA mutations in primary leukemia cells after chemotherapy: clinical significance and therapeutic implications. *Leukemia* **2003,** *17* (8), 1437-47.

165. Semenza, G. L., Targeting HIF-1 for cancer therapy. *Nature reviews. Cancer* **2003,** *3* (10), 721-32.

166. Rodríguez-Enríquez, S.; Carreño-Fuentes, L.; Gallardo-Pérez, J. C.; Saavedra, E.; Quezada, H.; Vega, A.; Marín-Hernández, A.; Olín-Sandoval, V.; Torres-Márquez, M. E.; Moreno-Sánchez, R., Oxidative phosphorylation is impaired by prolonged hypoxia in breast and possibly in cervix carcinoma. *The International Journal of Biochemistry & Cell Biology* **2010,** *42* (10), 1744-1751.

167. Upadhyay, M.; Samal, J.; Kandpal, M.; Singh, O. V.; Vivekanandan, P., The Warburg effect: insights from the past decade. *Pharmacology & therapeutics* **2013,** *137* (3), 318-30.

168. Lunt, S. Y.; Vander Heiden, M. G., Aerobic glycolysis: meeting the metabolic requirements of cell proliferation. *Annual review of cell and developmental biology* **2011,** *27*, 441-64.

169. Fantin, V. R.; St-Pierre, J.; Leder, P., Attenuation of LDH-A expression uncovers a link between glycolysis, mitochondrial physiology, and tumor maintenance. *Cancer cell* **2006,** *9* (6), 425-34.

170. Papaconstantinou, J.; Colowick, S. P., The role of glycolysis in the growth of tumor cells. I. Effects of oxamic acid on the metabolism of Ehrlich ascites tumor cells in vitro. *The Journal of biological chemistry* **1961,** *236*, 278-84.

171. Deck, L. M.; Royer, R. E.; Chamblee, B. B.; Hernandez, V. M.; Malone, R. R.; Torres, J. E.; Hunsaker, L. A.; Piper, R. C.; Makler, M. T.; Vander Jagt, D. L., Selective inhibitors of human lactate dehydrogenases and lactate dehydrogenase from the malarial parasite Plasmodium falciparum. *Journal of medicinal chemistry* **1998,** *41* (20), 3879-87.

172. Cameron, A.; Read, J.; Tranter, R.; Winter, V. J.; Sessions, R. B.; Brady, R. L.; Vivas, L.; Easton, A.; Kendrick, H.; Croft, S. L.; Barros, D.; Lavandera, J. L.; Martin, J. J.; Risco, F.; Garcia-Ochoa, S.; Gamo, F. J.; Sanz, L.; Leon, L.; Ruiz, J. R.; Gabarro, R.; Mallo, A.; Gomez de las Heras, F., Identification and activity of a series of azole-based compounds with lactate dehydrogenase-directed anti-malarial activity. *The Journal of biological chemistry* **2004,** *279* (30), 31429-39.

173. Granchi, C.; Bertini, S.; Macchia, M.; Minutolo, F., Inhibitors of lactate dehydrogenase isoforms and their therapeutic potentials. *Current medicinal chemistry* **2010,** *17* (7), 672-97.

174. Manerba, M.; Vettraino, M.; Fiume, L.; Di Stefano, G.; Sartini, A.; Giacomini, E.; Buonfiglio, R.; Roberti, M.; Recanatini, M., Galloflavin (CAS 568-80-9): a novel inhibitor of lactate dehydrogenase. *ChemMedChem* **2012,** *7* (2), 311-7.

175. Billiard, J.; Dennison, J. B.; Briand, J.; Annan, R. S.; Chai, D.; Colon, M.; Dodson, C. S.; Gilbert, S. A.; Greshock, J.; Jing, J.; Lu, H.; McSurdy-Freed, J. E.; Orband-Miller, L. A.; Mills, G. B.; Quinn, C. J.; Schneck, J. L.; Scott, G. F.; Shaw, A. N.; Waitt, G. M.; Wooster, R. F.; Duffy, K. J., Quinoline 3-sulfonamides inhibit lactate

dehydrogenase A and reverse aerobic glycolysis in cancer cells. *Cancer & metabolism* **2013,** *1* (1), 19.

176. Dragovich, P. S.; Fauber, B. P.; Corson, L. B.; Ding, C. Z.; Eigenbrot, C.; Ge, H.; Giannetti, A. M.; Hunsaker, T.; Labadie, S.; Liu, Y.; Malek, S.; Pan, B.; Peterson, D.; Pitts, K.; Purkey, H. E.; Sideris, S.; Ultsch, M.; VanderPorten, E.; Wei, B.; Xu, Q.; Yen, I.; Yue, Q.; Zhang, H.; Zhang, X., Identification of substituted 2-thio-6-oxo-1,6-dihydropyrimidines as inhibitors of human lactate dehydrogenase. *Bioorganic & medicinal chemistry letters* **2013,** *23* (11), 3186-94.

177. Kohlmann, A.; Zech, S. G.; Li, F.; Zhou, T.; Squillace, R. M.; Commodore, L.; Greenfield, M. T.; Lu, X.; Miller, D. P.; Huang, W. S.; Qi, J.; Thomas, R. M.; Wang, Y.; Zhang, S.; Dodd, R.; Liu, S.; Xu, R.; Xu, Y.; Miret, J. J.; Rivera, V.; Clackson, T.; Shakespeare, W. C.; Zhu, X.; Dalgarno, D. C., Fragment growing and linking lead to novel nanomolar lactate dehydrogenase inhibitors. *Journal of medicinal chemistry* **2013,** *56* (3), 1023-40.

178. Ward, R. A.; Brassington, C.; Breeze, A. L.; Caputo, A.; Critchlow, S.; Davies, G.; Goodwin, L.; Hassall, G.; Greenwood, R.; Holdgate, G. A.; Mrosek, M.; Norman, R. A.; Pearson, S.; Tart, J.; Tucker, J. A.; Vogtherr, M.; Whittaker, D.; Wingfield, J.; Winter, J.; Hudson, K., Design and synthesis of novel lactate dehydrogenase A inhibitors by fragment-based lead generation. *Journal of medicinal chemistry* **2012,** *55* (7), 3285-306.

179. Rao, S. T.; Rossmann, M. G., Comparison of super-secondary structures in proteins. *Journal of molecular biology* **1973,** *76* (2), 241-56.

180. Schmidt, R. K.; Gready, J. E., Molecular Dynamics Simulations of L-Lactate Dehydrogenase: Conformation of a Mobile Loop and Influence of the Tetrameric Protein Environment. *J Mol Model* **1999,** *5* (9), 153-168.

181. Wang, X.-C.; Jiang, L.; Zhou, H.-M., Minimal Functional Unit of Lactate Dehydrogenase. *J Protein Chem* **1997,** *16* (3), 227-231.

182. Gerstein, M.; Chothia, C., Analysis of protein loop closure. Two types of hinges produce one motion in lactate dehydrogenase. *Journal of molecular biology* **1991,** *220* (1), 133-49.

183. Read, J. A.; Winter, V. J.; Eszes, C. M.; Sessions, R. B.; Brady, R. L., Structural basis for altered activity of M- and H-isozyme forms of human lactate dehydrogenase. *Proteins* **2001,** *43* (2), 175-85.

184. Dunn, C. R.; Wilks, H. M.; Halsall, D. J.; Atkinson, T.; Clarke, A. R.; Muirhead, H.; Holbrook, J. J., Design and synthesis of new enzymes based on the lactate dehydrogenase framework. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **1991,** *332* (1263), 177-84.

185. Clarke, A. R.; Wigley, D. B.; Chia, W. N.; Barstow, D.; Atkinson, T.; Holbrook, J. J., Site-directed mutagenesis reveals role of mobile arginine residue in lactate dehydrogenase catalysis. *Nature* **1986,** *324* (6098), 699-702.

186. Deng, H.; Zhadin, N.; Callender, R., Dynamics of protein ligand binding on multiple time scales: NADH binding to lactate dehydrogenase. *Biochemistry* **2001,** *40* (13), 3767-73.

187. Pineda, J. R.; Callender, R.; Schwartz, S. D., Ligand binding and protein dynamics in lactate dehydrogenase. *Biophysical journal* **2007,** *93* (5), 1474-83.

188. Qiu, L.; Gulotta, M.; Callender, R., Lactate dehydrogenase undergoes a substantial structural change to bind its substrate. *Biophysical journal* **2007,** *93* (5), 1677-86.

189. Deng, H.; Brewer, S.; Vu, D. M.; Clinch, K.; Callender, R.; Dyer, R. B., On the pathway of forming enzymatically productive ligand-protein complexes in lactate dehydrogenase. *Biophysical journal* **2008,** *95* (2), 804-13.

190. Burgner, J. W.; Ray, W. J., Acceleration of the NAD-cyanide adduct reaction by lactate dehydrogenase: equilibrium binding effect as a measure of the activation of bound NAD. *Biochemistry* **1984,** *23* (16), 3620-3626.

191. Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K., BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic acids research* **2007,** *35* (Database issue), D198-201.

192. 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. *Journal of medicinal chemistry* **2012,** *55* (14), 6582-6594.

193. Halgren, T. A., Identifying and characterizing binding sites and assessing druggability. *Journal of chemical information and modeling* **2009,** *49* (2), 377-89.

194. Fauber, B. P.; Dragovich, P. S.; Chen, J.; Corson, L. B.; Ding, C. Z.; Eigenbrot, C.; Giannetti, A. M.; Hunsaker, T.; Labadie, S.; Liu, Y.; Liu, Y.; Malek, S.; Peterson, D.; Pitts, K.; Sideris, S.; Ultsch, M.; VanderPorten, E.; Wang, J.; Wei, B.; Yen, I.; Yue, Q., Identification of 2-amino-5-aryl-pyrazines as inhibitors of human lactate dehydrogenase. *Bioorganic & medicinal chemistry letters* **2013,** *23* (20), 5533-5539.

195. 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. *Journal of molecular graphics & modelling* **2010,** *29* (2), 157-70.

196. Canvas, version 1.8, Schrödinger, LLC, New York, NY, 2013.

197. Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K., Scalable molecular dynamics with NAMD. *J Comput Chem* **2005,** *26* (16), 1781-802.

198. Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E., Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* **2010,** *78* (8), 1950-8.

199. Bonomi, M.; Branduardi, D.; Bussi, G.; Camilloni, C.; Provasi, D.; Raiteri, P.; Donadio, D.; Marinelli, F.; Pietrucci, F.; Broglia, R. A.; Parrinello, M., PLUMED: A portable plugin for free-energy calculations with molecular dynamics. *Computer Physics Communications* **2009,** *180* (10), 1961-1972.

200. Smoot, M. E.; Ono, K.; Ruscheinski, J.; Wang, P. L.; Ideker, T., Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* **2011,** *27* (3), 431-2.

201. Fruchterman, T. M. J.; Reingold, E. M., Graph drawing by force-directed placement. *Softw. Pract. Exper.* **1991,** *21* (11), 1129-1164.

202. Bader, G. D.; Betel, D.; Hogue, C. W., BIND: the Biomolecular Interaction Network Database. *Nucleic acids research* **2003,** *31* (1), 248-50.

203. Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L., Evaluation and Reparametrization of the OPLS-AA Force Field for Proteins via Comparison with Accurate Quantum Chemical Calculations on Peptides†. *The Journal of Physical Chemistry B* **2001,** *105* (28), 6474-6487.

204. www.asinex.com.

205. Rishton, G. M., Nonleadlikeness and leadlikeness in biochemical screening. *Drug discovery today* **2003,** *8* (2), 86-96.

206. Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S., Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): evaluation of performance. *Journal of chemical information and computer sciences* **2004,** *44* (5), 1708-18.

207. Brownstein, M. J., A brief history of opiates, opioid peptides, and opioid receptors. *Proceedings of the National Academy of Sciences of the United States of America* **1993,** *90* (12), 5391-3.

208. Portoghese, P. S., A new concept on the mode of interaction of narcotic analgesics with receptors. *Journal of medicinal chemistry* **1965,** *8* (5), 609-16.

209. Gilbert, P. E.; Martin, W. R., Sigma effects of nalorphine in the chronic spinal dog. *Drug and alcohol dependence* **1976,** *1* (6), 373-6.

210. Gilbert, P. E.; Martin, W. R., The effects of morphine and nalorphine-like drugs in the nondependent, morphine-dependent and cyclazocine-dependent chronic spinal dog. *The Journal of pharmacology and experimental therapeutics* **1976,** *198* (1), 66-82.

211. Martin, W. R.; Eades, C. G.; Thompson, J. A.; Huppler, R. E.; Gilbert, P. E., The effects of morphine- and nalorphine- like drugs in the nondependent and morphine-dependent chronic spinal dog. *The Journal of pharmacology and experimental therapeutics* **1976,** *197* (3), 517-32.

212. Lord, J. A.; Waterfield, A. A.; Hughes, J.; Kosterlitz, H. W., Endogenous opioid peptides: multiple agonists and receptors. *Nature* **1977,** *267* (5611), 495-9.

213. Mollereau, C.; Parmentier, M.; Mailleux, P.; Butour, J. L.; Moisand, C.; Chalon, P.; Caput, D.; Vassart, G.; Meunier, J. C., ORL1, a novel member of the opioid receptor family. Cloning, functional expression and localization. *FEBS letters* **1994,** *341* (1), 33-8.

214. Thompson, A. A.; Liu, W.; Chun, E.; Katritch, V.; Wu, H.; Vardy, E.; Huang, X. P.; Trapella, C.; Guerrini, R.; Calo, G.; Roth, B. L.; Cherezov, V.; Stevens, R. C., Structure of the nociceptin/orphanin FQ receptor in complex with a peptide mimetic. *Nature* **2012,** *485* (7398), 395-9.

215. Pasternak, G. W., Multiple opiate receptors: deja vu all over again. *Neuropharmacology* **2004,** *47 Suppl 1*, 312-23.

216. Mattia, A.; Vanderah, T.; Mosberg, H. I.; Porreca, F., Lack of antinociceptive cross-tolerance between [D-Pen2, D-Pen5]enkephalin and [D-Ala2]deltorphin II in mice: evidence for delta receptor subtypes. *Journal of Pharmacology and Experimental Therapeutics* **1991,** *258* (2), 583-587.

217. Attali, B.; Gouarderes, C.; Mazarguil, H.; Audigier, Y.; Cros, J., Evidence for multiple "Kappa" binding sites by use of opioid peptides in the guinea-pig lumbo-sacral spinal cord. *Neuropeptides* **1982,** *3* (1), 53-64.

218. Law, P. Y.; Reggio, P. H.; Loh, H. H., Opioid receptors: toward separation of analgesic from undesirable effects. *Trends in biochemical sciences* **2013,** *38* (6), 275-82.

219. Zollner, C.; Stein, C., Opioids. *Handbook of experimental pharmacology* **2007,** (177), 31-63.

220. Johnson, S. M.; Fleming, W. W., Mechanisms of cellular adaptive sensitivity changes: applications to opioid tolerance and dependence. *Pharmacological reviews* **1989,** *41* (4), 435-88.

221. Koch, T.; Widera, A.; Bartzsch, K.; Schulz, S.; Brandenburg, L. O.; Wundrack, N.; Beyer, A.; Grecksch, G.; Hollt, V., Receptor endocytosis counteracts the development of opioid tolerance. *Mol Pharmacol* **2005,** *67* (1), 280-7.

222. Granier, S.; Manglik, A.; Kruse, A. C.; Kobilka, T. S.; Thian, F. S.; Weis, W. I.; Kobilka, B. K., Structure of the delta-opioid receptor bound to naltrindole. *Nature* **2012,** *485* (7398), 400-4.

223. Manglik, A.; Kruse, A. C.; Kobilka, T. S.; Thian, F. S.; Mathiesen, J. M.; Sunahara, R. K.; Pardo, L.; Weis, W. I.; Kobilka, B. K.; Granier, S., Crystal structure of the micro-opioid receptor bound to a morphinan antagonist. *Nature* **2012,** *485* (7398), 321-6.

224. Wu, H.; Wacker, D.; Mileni, M.; Katritch, V.; Han, G. W.; Vardy, E.; Liu, W.; Thompson, A. A.; Huang, X. P.; Carroll, F. I.; Mascarella, S. W.; Westkaemper, R. B.; Mosier, P. D.; Roth, B. L.; Cherezov, V.; Stevens, R. C., Structure of the human kappa-opioid receptor in complex with JDTic. *Nature* **2012,** *485* (7398), 327-32.

225. Goto, Y.; Arai-Otsuki, S.; Tachibana, Y.; Ichikawa, D.; Ozaki, S.; Takahashi, H.; Iwasawa, Y.; Okamoto, O.; Okuda, S.; Ohta, H.; Sagara, T., Identification of a novel spiropiperidine opioid receptor-like 1 antagonist class by a focused library approach featuring 3D-pharmacophore similarity. *Journal of medicinal chemistry* **2006,** *49* (3), 847-9.

226. Hughes, J.; Smith, T. W.; Kosterlitz, H. W.; Fothergill, L. A.; Morgan, B. A.; Morris, H. R., Identification of two related pentapeptides from the brain with potent opiate agonist activity. *Nature* **1975,** *258* (5536), 577-80.

227. Simantov, R.; Snyder, S. H., Morphine-like peptides, leucine enkephalin and methionine enkephalin: interactions with the opiate receptor. *Mol Pharmacol* **1976,** *12* (6), 987-98.

228. Zadina, J. E.; Hackler, L.; Ge, L. J.; Kastin, A. J., A potent and selective endogenous agonist for the mu-opiate receptor. *Nature* **1997,** *386* (6624), 499-502.

229. Yamazaki, T.; Ro, S.; Goodman, M.; Chung, N. N.; Schiller, P. W., A topochemical approach to explain morphiceptin bioactivity. *Journal of medicinal chemistry* **1993,** *36* (6), 708-19.

230. Meunier, J. C.; Mollereau, C.; Toll, L.; Suaudeau, C.; Moisand, C.; Alvinerie, P.; Butour, J. L.; Guillemot, J. C.; Ferrara, P.; Monsarrat, B.; et al., Isolation and structure of the endogenous agonist of opioid receptor-like ORL1 receptor. *Nature* **1995,** *377* (6549), 532-5.

231. Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P.,

ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic acids research* **2012,** *40* (Database issue), D1100-7.

232. Cheng, T.; Li, X.; Li, Y.; Liu, Z.; Wang, R., Comparative assessment of scoring functions on a diverse test set. *Journal of chemical information and modeling* **2009,** *49* (4), 1079-93.

233. Ferrara, P.; Gohlke, H.; Price, D. J.; Klebe, G.; Brooks, C. L., 3rd, Assessing scoring functions for protein-ligand interactions. *Journal of medicinal chemistry* **2004,** *47* (12), 3032-47.

234. Klebe, G., Virtual ligand screening: strategies, perspectives and limitations. *Drug discovery today* **2006,** *11* (13-14), 580-94.

235. Kufareva, I.; Ilatovskiy, A. V.; Abagyan, R., Pocketome: an encyclopedia of small-molecule binding sites in 4D. *Nucleic acids research* **2012,** *40* (Database issue), D535-40.

236. Totrov, M.; Abagyan, R., Flexible protein-ligand docking by global energy optimization in internal coordinates. *Proteins* **1997,** *Suppl 1,* 215-20.

237. Abagyan, R.; Batalov, S.; Cardozo, T.; Totrov, M.; Webber, J.; Zhou, Y., Homology modeling with internal coordinate mechanics: deformation zone mapping and improvements of models via conformational search. *Proteins* **1997,** *Suppl 1,* 29-37.

238. Cardozo, T.; Batalov, S.; Abagyan, R., Estimating local backbone structural deviation in homology models. *Computers & chemistry* **2000,** *24* (1), 13-31.

239. Katritch, V.; Rueda, M.; Lam, P. C.; Yeager, M.; Abagyan, R., GPCR 3D homology models for ligand screening: lessons learned from blind predictions of adenosine A2a receptor complex. *Proteins* **2010,** *78* (1), 197-211.