Università degli Studi di Padova

Dipartimento di *Scienze Chimiche*

SCUOLA DI DOTTORATO DI RICERCA IN SCIENZE MOLECOLARI

INDIRIZZO: Scienze Farmaceutiche - XXVIII CICLO

NOVEL IN SILICO APPROACHES TO DEPICT THE PROTEIN-LIGAND RECOGNITION EVENTS

Direttore della Scuola : Ch.mo Prof. Antonino Polimeno

Coordinatore d’indirizzo: Ch.mo Prof. Alessandro Dolmella

Supervisore :Ch.mo Prof. Stefano Moro

Dottorando : Alberto Cuzzolin

To my Family . . .

LIST OF CONTENTS

Abstract

The discovery and commercialization of a new drug is a long and expensive process. Such process is divided into different phases during which the phisico-chemical and therapeutic properties of the compounds are determined. In particular, the aim of the first phase is to verify whether the compound recognises and interacts efficiently with the target protein.

In the last decade, several computational tools have been developed and used to support experimentalists. For this purpose, the scientist have to deal with high complex systems that are difficult to study in whole; thus, the methods and algorithms developers have to strongly simplify the system treatment. Moreover, the time required to obtain the results depends on the computational resources (hardware) available. Fortunately, the technological progress have increased the computing power at low cost, resulting in new and more complex techniques development.

During this Ph.D. project we were focused on the development and even the improvement of *in silico* methods, which allowed to answer certain questions by saving time and money. Furthermore, these methods were implemented in software presenting a Graphical Unit Interface (GUI) with the aim to enhance the user-friendliness.

The computational techniques often require a high understanding of the methodology theoretical aspects and also a good informatics proficiency, like different type files handling and hardware management. For this reason, our developed software were organized as pipelines to automatize the entire process and to make this tools useful also for non-expert users.

Finally, these methodologies were applied in several research projects demonstrating their usefulness by elucidating, for the first time, interesting aspects of the ligand-protein recognition pathway.

Sommario

La scoperta e la commercializzazione di un nuovo farmaco è un processo lungo e dispendioso, che si articola in diverse fasi durante le quali vengono determinate le proprietà fisiche, chimiche e terapeutiche dei composti investigati. In particolare, nella prima fase di questo processo si cerca di verificare che il composto riconosca e interagisca efficacemente con la proteina bersaglio.

A tale scopo, negli ultimi decenni numerosi strumenti computazionali sono stati sviluppati e utilizzati per supportare i ricercatori che si adoperano nella parte sperimentale. I problemi affrontati presentano un alto livello di complessità, che sarebbero difficili da studiare *in toto,* perciò gli sviluppatori di metodi e algoritmi devono necessariamente adottare notevoli semplificazioni. Inoltre, le risorse di calcolo (*hardware*) determinano le tempistiche con le quali è possibile ottenere il risultato richiesto. In tal senso, lo sviluppo tecnologico ha portato a un importante aumento della potenza di calcolo a costi accessibili, stimolando l'interesse per lo sviluppo di tecniche sempre più complesse.

Durante questo progetto di dottorato ci si è focalizzati sullo sviluppo e il miglioramento di metodi *in silico,* che permettono di rispondere ad alcuni interrogativei a costi e tempistiche di molto ridotte.

Inoltre, tali metodi sono stati implementati in software dotati di interfaccia grafica (GUI) al fine di poter aiutare l'utente nel loro utilizzo.

Le tecniche computazionali spesso richiedono un'elevata conoscenza teorica delle metodologie e anche una certa competenza informatica, come la gestione di diversi tipologie di file e delle risorse hardware da impiegare. Per questo motivo i software da noi sviluppati sono stati organizzati in *pipelines,* in modo da automatizzare l'intero processo e rendere questi strumenti fruibili anhce a persone non esperte.

Infine, l'utilità di queste nuove metodologie è stata comprovata in progetti in cui questi strumenti hanno permesso di delucidare aspetti interessanti e fino ad ora non ancora accessibili nell'ambito del riconoscimento proteina-ligando.

INTRODUCTION

1.1 The drug discovery process

The drug discovery process aims to identify molecules with specific therapeutic effects that can be introduced into the market.
Historically, the drugs discovered were identified from natural active products and serendipitous events. For example morphine and digoxin are drugs extracted from opium and *Digitalis lanata* respectively, whereas the penicillin unearthed by Fleming is an example of fortuity discovery.

Since the beginning of the contemporary medicinal chemistry, it was clear that there was a relationship between chemical properties and biological response. The previous understanding induces the researchers to synthetize multiple molecules, similar to known active compounds, in order to pharmacologically test them and finally to identify plausible new candidates for a specific target.

Nowadays the pharmaceutical companies need more multidisciplinary and high-level of planning to accomplish successfully the entire process[1]. Furthermore the market is very competitive and therefore demands for high value-added compounds, which needs to have improved characteristics in order to be fully beneficial for the healthcare. The modern discovery process consists of multiple consecutive steps that can be grouped into two main stages: the Preclinical and Clinical stages.

The early-stage of the discovery usually starts with the identification and characterization of a target that can be involved in the treatment of a specific disease. The validation of the target is not trivial and can be achieved by evaluating the signaling system downstream with different techniques: generation of drug-resistance mutant, knockdown or overexpression of the presumed target. Afterwards by using a High-Throughput Screening (HTS), the companies attempt to identify active compounds (hit compounds). The chemical libraries usually contain several hundreds or even thousand of compounds, which can be part of in-home chemical library or to be synthetized *ex novo*(Fig. 1). In the next stage the chemical properties of the identified compounds are compared with the biological response, in order to determine any relationship between the molecule structure and the activity response (SAR).

Once promising compounds are identified, they are tested to evaluate their properties in more details, such as its mechanism of action, effectiveness compared with similar drugs and the intrinsic chemical properties. Thus the synthesis of different analogs based on the hits obtained previously allow to guide the development of optimized ligand (lead compounds.) The data are analyzed and the best performed compounds are carried on for the next phase, which requires 3-4 years and arise 1.000 compounds (Fig. 1B)[2].

In the preclinical phase (Fig.1) the molecules are previously screened *in vitro* in order to evaluate their properties more in detail and consequently reduce the number of compounds to be screened *in vivo.* The remaining compounds are tested in animal models to gather information about the administration route, the Absorption, the Distribution, the Metabolism, the Excretion and the Toxicity

(ADMET). Afterwards, for the compounds with the best features, the formulation development and physiological assay are carried out. The preclinical stage is essential to determine the security profile of the entities proposed in order to respect the mandatory principle "*primum non nocere"* (first, do not harm) and this stage generally takes two years, promoting only 10 molecules (Fig. 1).

Afterwards the candidate drugs pass into the clinical phase that is focused in humans. The clinical phase can be splitted in three stages, in which different aspects are taken into account:

- **The Phase I** also called Clinical Pharmacology phase*,* has the purpose to gather information about the absorption, the metabolism, and the distribution effect in organs and tissues. In addition, the dose-dependent side effects are monitored. The results are obtained considering 20-100 volunteers that have to be healthy, in order to limit the observed side effects not relied on the administration.
- **The Phase II** or Efficacy phase, aims to elucidate the treatment efficacy, the short-term side effects and the optimization of the dose. For this stage several hundreds of patients are recruited and splitted into two groups: treated and control group. The former is treated with the drug, whereas the latter receives a false-drug in which none active compound is present. The reason for this procedure is to establish the placebo and nocebo effect experienced by the patients. These effects emerge due to the therapy expectation of the patients, where the placebo comprehends positive effect and the nocebo effect consist in side effects.
- **The Phase III** or Multicentric phase, precedes the approval of the drug and its consequent release on the market. During this step, up to thousands patients are treated with the drug in order to evaluate the benefit versus risk ratio, the uncommon and the long-term side effects and the preparation of the "Patient information leaflet", which is made at the end of the phase.

At the end of the clinical phase one compound usually reach the market, but it is still monitoring for 2-3 years in order to evaluate rare or long-term side effects.

The entire process presents a high rate of fails: 5% of the screened compounds present suitable characteristics and therefore pass into the preclinical phase; in the next step only the 2% of molecules give positive results, whereas in the clinical phase the fail ratio is over the 80%[3].

The attrition rate was investigated also by Pranita *et al.* and they estimated that only 1 out of 12 drugs entering in the Clinical phase become a new drug. The 50% of failure can be ascribed to poor bioavailability, pharmacokinetics or cause adverse events.

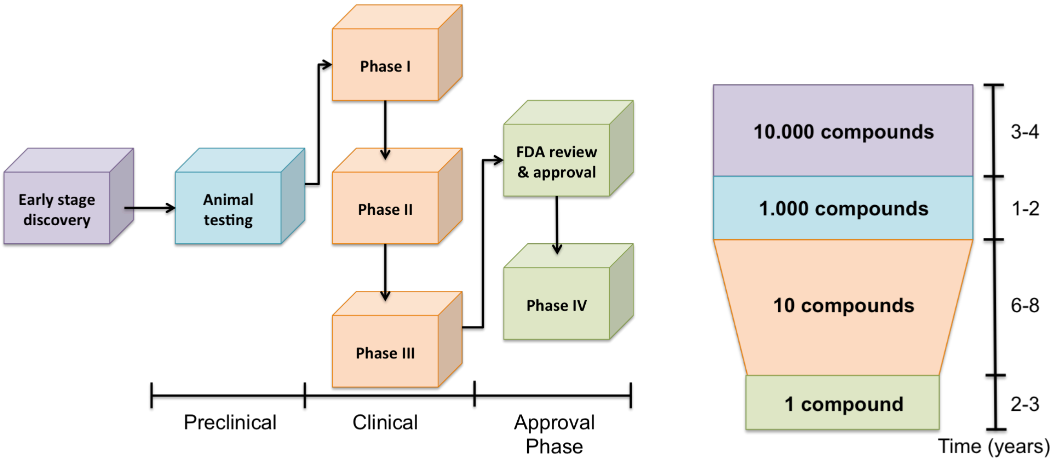


Figure 1 – Drug Discovery Process (DDP): On the left is reported the stages of the drug discovery process; on the right the timeline is shown coupled with the amount of compounds usually consider for each stage.

As it was mentioned above, the development of a new drug is a process that requires an incredible amount of time and founding, highlighted also by the *Tufts Center for the Study of Drug Development* in a recent study. Such study consisted on analyze 10 pharmaceutical companies and 106 randomly selected drugs, between 1995 and 2007[4]. From the data, it was estimated that the average cost for one medicine is 2.6 billion of dollars and can take longer than 10 years. In addition $312 million is the average out-of-pocket cost for the post-approval, Research and Development (R&D), consisting of new indications, formulations, dosage strengths and regimens, and for the monitoring of the long-term side effects required by the U.S. Food and Drug Administration (FDA).

The Fig. 2 shows how the cost of the R&D has been exponentially increased in the last decades due to the dreadful challenge level reached. At the same time, the expectation of the high-quality new chemical entities (NCEs) induces the companies to enhance their in-home technologies and protocols. In 2002, Bolten *et al.* estimated that R&D spending increased by approximatley 40%, whereas new drug approvals dramatically decreased by 50%[3]. Furthermore, they calculated that a pharmaceutical company needs to release on market at least 4 NCEs per year, each of them with an average revenue of $350 milion for not failing. In the same period Oprea *et al.* evaluated the R&D cost and efficacy in the development of new drugs. Indeed, both groups concluded that in order to satisfy the investor expectation, the companies need to reduce R&D spending and to shorten from the drug design, to its development and finally the launch of the drug into the market[5].

In the 2011, Pammolli *et al.* highlighted the crisis in R&D, by evaluating the cost of each stage in the drug discovery; they identified which one presents the worst success rate versus cost ratio[6]. In the time window considered in these studies, the pharmaceutical companies experienced an alarming crisis with the lowest NCEs that have been approved in the 2007 (19 NCEs) since 1983. Indeed this scenarios continuous up to the 2010, thereafter the NCEs approved shown a drastic increase with a peak of 41 in the 2014(Fig. 3)[7].

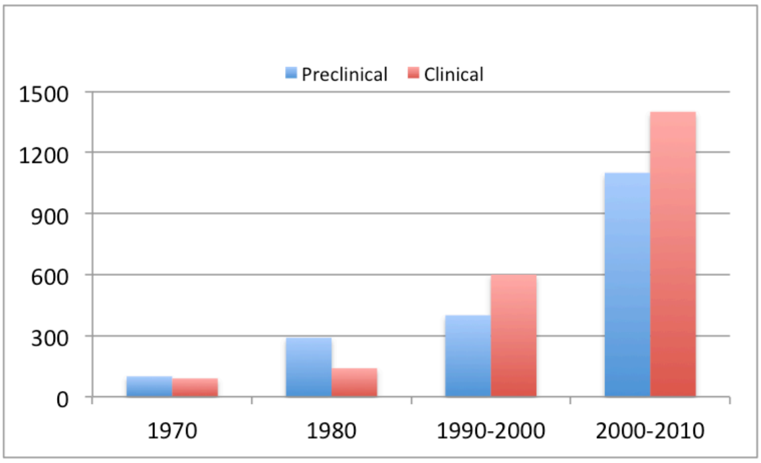


Figure 2 – Drug Discovery cost: **Comparison of the Clinical and Preclinical average costs of drugs in the market, from 1970s to 2010s**

It is clear that the drug discovery is a tedious and a time-consuming process. In addition, the revenues are constantly decreasing in last decade due to multiple reasons, which determine this dramatic scenario, like the patent expiration and consequent competition with the generic, the reduced periods of exclusivity and price constraints.

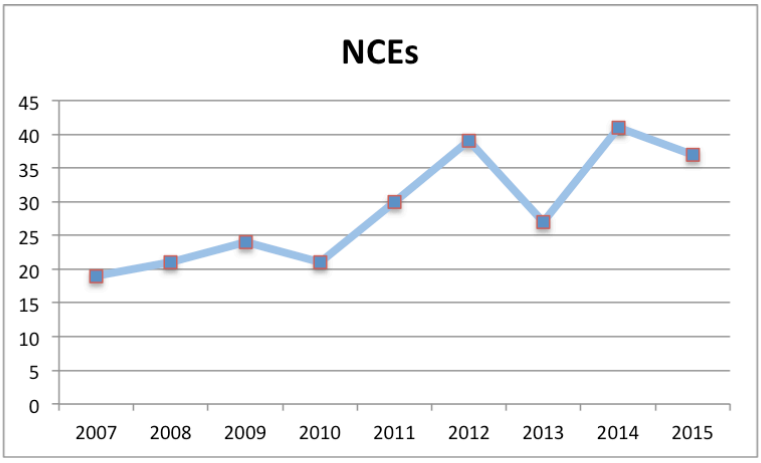


Figure 3 – NCE in the market: The number of new chemical entities introduced into the market (y-axis) are plotted against the years (x-axis)

The pharmaceutical companies are constantly improving innovative methodologies to shorten and reduce the spending in the discovery cycle. The Computational Aided Drug Design (CADD), introduced in the 1980s, is a perfect example of methodology that helped to speed up the discovery, with a limited investment. This approach depicts perfectly the current main goal of the companies that can be summed up as follow: “fail faster, fail cheaper”[8].

1.2 Computational Aided Drug Design (CADD)

Since the beginning of the 1900s with the receptor theory of “lock-and-key” formulated by E. Fisher (1894) and P. Ehrlich (1909)[9], the compounds properties become the main focus. It was clear that the complementarity between the receptor and a candidate ligand needs to be investigated deeply. Thus, according to this theory, the biological response (BR) depends on the ligand structure (S) and its physicochemical properties along all the pathway, from the administration to the receptor interaction (Fig. 4).

Undoubtedly multiple properties have to be taken into account to be able to evaluate the BR which can roughly be expressed as a function of these assumptions[1]:

$$BR = f(S)$$

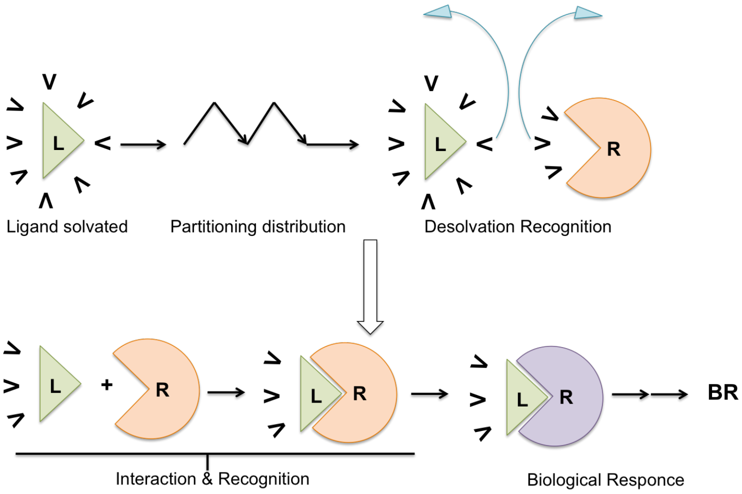


Figure 6 – Ligand pathway to Biological response: The consequent main steps followed by the ligand from the administration to the biological response elicit, is reported.

Interestingly, it is not possible to establish *a priori* which physicochemical properties have to be considered to explain the BR observed or to distinguish between active or inactive compounds.

The high amount of properties and their ample variety can discourage their determinations, especially from an experimentally point of view. Moreover certain properties are prohibitive or even impossible to be measured by experimentalists, such as the electronic structures or molecular orbital occupancies. Fortunately, this limitation could be overcome by computational scientists, which can support the understanding by analyzing part of these properties in parallel in order to gather useful information. In contrast, bulk properties such as pK and solubility are accessible only by experiments. Other properties like partition-coefficient (LogP) can be obtained by both experimentalists and computational scientists.

As it was mentioned above the physicochemical properties of the ligand structure could be implied at different stages of the process, thus CADD can be employed at different level of the drug discovery, to save time and money. However computational methods provide the best performance in the early-stage of the process, especially in the lead discovery and lead optimization (Fig. 5). Undeniably computational studies are a useful support of the experiments and different approaches are available. At this point is important to mention that the methodologies applied are dependent on the information available, such as protein structure or known active and inactive compounds.

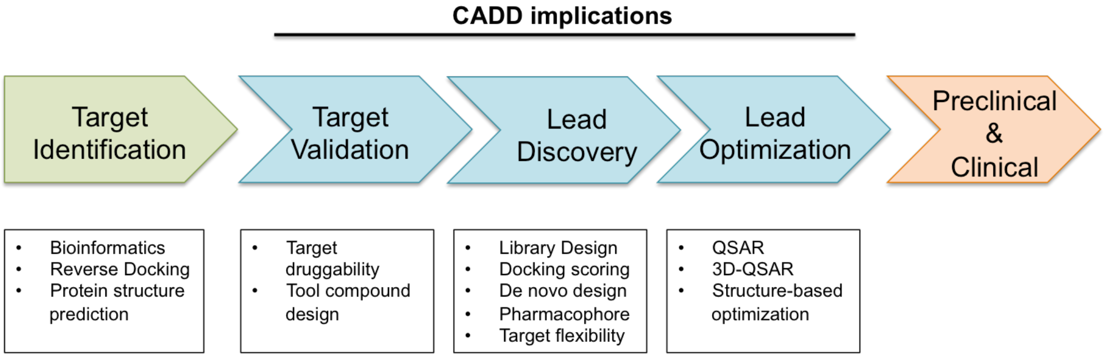


Figure 5 – CADD implications in the early stage of the drug discovery: The sub-stages of the early stage of the drug discovery are reported. The CADD implications and methodology are also reported.

1.3 Lead Discovery

In the first stage of the drug discovery process the researchers have to collect information about the target and its implication in the disease. Therefore the planning of the discovery is based on the accessible information about the target and from any previous screening.

Compounds (Ligand/Inhibitor)	Protein (Enzyme/Receptor): Unkown	Protein (Enzyme/Receptor): Known
Known	Pharmacophore Ligand Similarity QSAR	Structure Based Drug Design
Unknown	Combinatorial Chemistry Highthroughput Screening Random Screening	De Novo Design

Figure 6 – CADD applications in the drug discovery process: Four scenarios are possible based on the compounds activities and 3D structure of the protein available. Moreover the main techniques are reported for each case.

1.3.1 Lead Generation

1.3.1.a Case 1: Unknown 3D Protein Structure

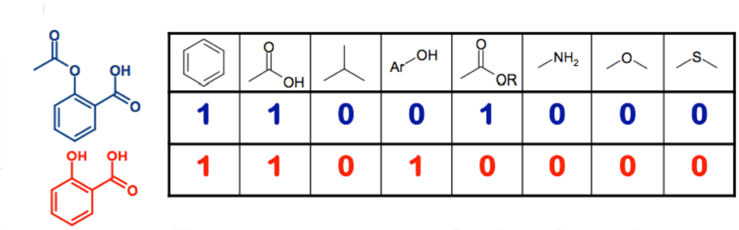
In the case of a completely new campaign the researchers can often face the scenario in which neither the protein structure nor known active compounds are available. In this inauspicious case it is necessary to quickly identify compounds that show at least a minimum activity. For this purpose, a set of compounds is collected and a random screening is performed. In the 1990s the combinatorial chemistry was introduced as a new methodology that speed up the synthesis of millions of compounds in a single process. Consequently the first HTS had to be set up in order to test this amount of molecules in a short time. Beyond this assay, the selection of the molecules to be included in the chemical library is of utmost importance to save time and money during the *in vitro* tests.

I) Chemical Library Preparation

The in-home chemical libraries often contain many analogs with a common scaffold structure causing the presence of duplicates and therefore, reducing the chemistry diversity ratio of the set to investigate. Although the application of filters reduces the amount of molecules to manage, this inevitably decreases the chemical space explored. For this reason the strictness of the selection must be well balanced in order not to miss possible candidates and, at the same time, not to screen too many molecules.

In that sense, computational scientists play a key role in the building of the chemical library by considering different physicochemical properties and similarity of the molecules to be included. The compounds of a collection can be compared to determine how much similar are the molecules with respect of each other. Hence the similarity approach gives the possibility to exclude part of them and to enhance the chemical diversity of the library.

The similarity among the compounds can be determined by computing the Jaccard-Tanimoto index[10] (eq. 1), which reports the ratio of common elements present into two different vectors. In chemical field the arrays are bitmaps representing the presence of functional groups (Fig. 7). The Molecular ACCess System (MACCS) database collects a total of 166 functional groups that are a set of the most accessible medchem chemical space[11].



1	1	0	0	1	0	0	0
1	1	0	1	0	0	0	0

Figure 7 – MACCS functional group example: Two compounds are used to show an example of how MACCS functional groups are converted into bitmap vectors.

$$(eq. 1) \qquad Tanimoto\ index = \frac{A \cap B}{A_{solo} + B_{solo} + A \cap B}$$

A value of 0.85 for the Tanimoto index is the cutoff commonly used to determine the diversity of the molecules. The application of a rational selection of the molecules to be included in the chemical library is estimated to reduce the amount of the molecules to be screened by approximately 30% without threatening the chemical space investigated. Furthermore it was exploited that

3.5-3.7 times more compounds need to be considered with the use of a random approach to cover the same chemical space of a rational approach[12].

Alternatively the physicochemical properties can be exploited to investigate which compounds present a drug-likeness profile. In 1997 Lipinski formulates the so called “rule of five” that provides a set of molecules properties that are usually observed in small active molecules[13,14]. The statements of the rules consider the molecular weight, the hydrophobicity and the H-bond features:

- Molecular weight lower than 500 Dalton
- Log P lower than 5
- H-bond donors lower than 5
- H-bond acceptors lower than 10

Other studies suggest to use even more strict cutoffs in the early hit identification, as reported by Hann *et al.*[15]. In the same time Oprea *et al.* analyzed a set of lead-drug pairs and they infer that the hit compounds often present 100 Da, one ring, two flexible bonds and a LogP/LogD unit less than optimized lead[16]. Nevertheless the compounds collection can contain hits that have substructural features that arise false positive response. Hence these hits are artifacts whom activity is independent from a specific ligand-protein interactions and these molecules are defined as “pan-assay interference compounds” (PAINS)[17,18]. The false activity shown by these compounds can be due to different events, such as aggregation, chemical decomposition, protein reactivity and fluorescence. Also in this case CADD can perform a filtering of the database based on the chemical features which induce these assay interferences.

Recently Vilar *et al.* proposed a protocol to predict drug-drug interactions (DDIs) based on database of known DDIs that are identified by a similarity search. In this protocol the MACCS functional groups are used as bitmap vectors and the Tanimoto index is used to evaluate the similarity[19].

Finally, the database collects a high chemical diversity of molecules that are characterized by drug-like physicochemical properties and without known toxic features. The compounds collected are screened and the active molecules are deeply investigated in order to identify any similarity of their physicochemical properties. Once active compounds are available for a target are available it is possible to investigate their properties so as to identify which of them could be responsible for the desired pharmacological activity[20]. Indeed, the more hits are known, the more information can be obtained and compared with the aim of recognize which parameters are really important.

II) Ligand-Based Drug Design

The discovery approach based on the known active compounds is named “Ligand-Based Drug Design” or, in alternative, “Indirect drug design”.

Ligand Based approach attempts to return plausible new candidates from a structure activity relationship (SAR).

All the techniques comprehended in this field are based on the main concept of pharmacophore, which was defined by Ehrlich as: "a molecular framework that carries the essential features responsible for a drug's biological activity"[9]. By the analysis of the ligands features, is possible to build a pharmacophore model in which the types, positions and the directions of the features can be encoded and the possible steric constrains can be also included[21]. The features represent the substructural fundamental elements like aromatic rings, hydroxyl groups and basic amines that are reported as atom types and connectivities. For each feature it is possible the conversion into the equivalent geometric object that encodes the position and consequently provide three-dimensional properties. Moreover the direction can be take into account as vectors or planes in order to describe the ligand orientation respect to the receptor (Fig. 8). Then, the pharmacophore can be optimized by applying geometric constrains to the features, such as distances, angles and torsional angles.

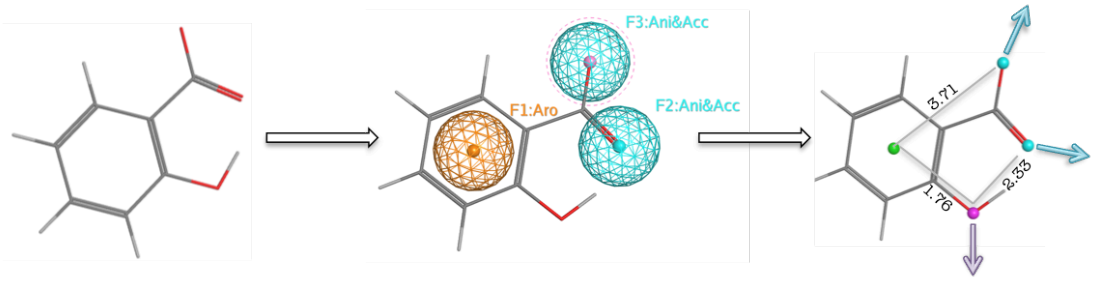


Figure 8 – Pharmacophore model: In the middle the pharmacophore features are represented by a red ball for the aromatic (F1:Aro) and two cyan for the H-bond acceptor (F2 and F3:Ani&Aro). On the right one additional feature is included as magenta sphere representing the H-bond donor and several constrains like the distances between the atoms are reported.

Normally more than one pharmacophore model is generated, thus it is necessary to identify the most reliable one. For this purpose, another database of known active and inactive compounds are used as validation set for the generated models, which are accepted only if they are able to distinguish between active and inactive ones.

Finally the accepted pharmacophore model can be used to query any compound database available to identify new hits. Indeed the conformations of the compounds present in the database could not be the suitable one to right match the three-dimensional pharmacophore model, thus a conformations search need to be performed. The conformations can be previously generated or computed "on the fly".

Due to the fact that the bio-active conformation could not be the lowest energy one, but surely is not the highest-energy conformation[22,23], it is possible to perform a simple torsional minimization. By using this technique it is common that only the conformations related to the closest local minimum are explored, therefore iterating this process by starting from several random conformations can enhance the conformation space explored.

As an example of drug design from a pharmacophore model we can mention the study presented by the group of Lopez-Rodriguez. They focused the study in the serotonin receptor (5-HT$_7$R), which is involved in the treatment of sleep disorder, believed to have a role in depression as consequence of deregulated circadian rhythm. They take advantage of the pharmacophore technique to design new derivatives of two known antagonists: Naphtholactam and Naphthosultam. Based on these two structures a pharmacophore model was built and used to identify new antagonists for the G protein-coupled receptor (GPCR). The pharmacophore model consisted of five features: a positive ionizable atom (PI), a H-bond acceptor group (HBA), and three hydrophobic regions (HYD). The quality of the model was proved by screening the compound designed and synthesized based on the pharmacophore model[24].

The brief overview provided here is not certainly exhaustive, but points out only the main concepts. For further information, there are several reviews that provide more details of the ligand-based approach to discover a lead compound[21,25,26]. The Figure 9 summarizes the classic workflow for the identification of a lead compound.

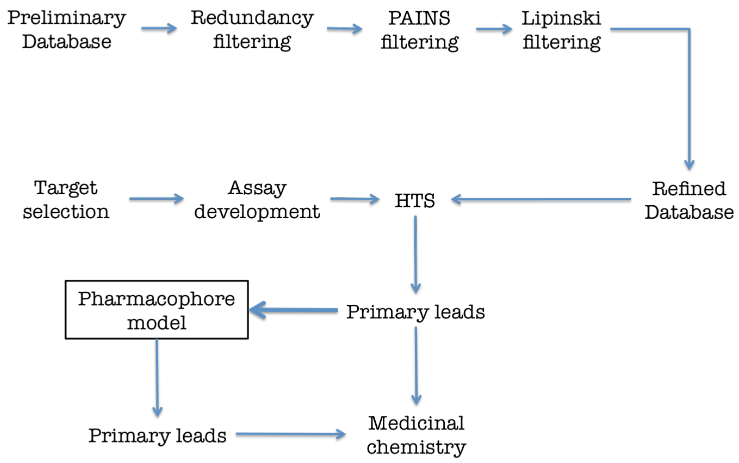


Figure 9 – Ligand Based Drug Design workflow: A classical workflow for the Ligand Based Drug Design.

1.3.1.b Case 2: Known 3D Protein Structure

The knowledge of the three-dimensional structure of the protein can provide important information that are not accessible with the Ligand-Based approach. In the previous century several Nobel Prizes were awarded to researchers for the protein structure determinations. In 1936 Debye was the first of them who was the pioneer of the x-ray diffraction for the study of molecules structures. Afterwards in 1962 Perutz was awarded for the structure determination of globular proteins and in the 1964 Hodgkin for vitamin B12 and insulin. Regarding membrane proteins, only in 1988 Deisenhofer, Huber and Michel were able to solve the photosynthetic reaction center and in the 1997 Walker provide the structure of ATP-synthase.

From 2003 to 2012 Mackinnon , Kornberg, Steitz and Yonath and Kobilka and Lefkowitz won also the Nobel Prize for protein structure determinations[27].

Considering the accessible information from the protein structures, the NIH was the first institution which collected these solved structures in the 1980s. Such proteins collection motivated the scientist to focus their effort into the structure determination of a large number of drug targets structures by using proteomic and genomic techniques[28,29]. The accessibility of this conspicuous number of protein structures was a new starting point of investigation for the drug discovery.

Nowadays the elucidation of the three-dimensional protein structures are obtained by using two techniques: NMR and X-ray crystallography. The former has proved its powerfulness in the last decade in the determination of polypeptides and small proteins[30] (up to 30 KDa). Indeed NMR can provide results in the dissolved aggregation state which better reproduce the real conditions. Nevertheless the solvent commonly used, such as chloroform and dimethylsulfoxide, poor mimic the physiological environmental conditions. One of the most important advantage of the NMR is the sampling of flexible regions that are withdrawn by the crystallographers and thus, do not contribute to the final model. In effect NMR structure determination does not lead to a single "image", but it generates an ensemble of structures possibly depicting different conformations. The second technique, X-ray crystallography, takes advantage of X-ray diffraction to determine precise location of all the atoms of the molecule within the crystal lattice[31].

The positional error of atoms can be approximate to 1/6 of the crystal resolution[32]. Despite of the overall fold of protein in crystalline environment was proved to be very similar to the solution one[33,34], loops conformation and side-chains can differ. The reason is mainly due to the effects of crystal packing and to potential function or search algorithm. In 2002 Jacobson *et al*. presented a side-chain prediction algorithm that takes into account the packing effects (*i.e.* van der Waals interactions), hence the possibility to generate high resolution crystal structures[35].

The most popular database where the solved structure are deposited, is "The Protein Data Bank"[36], which was established in 1971 containing seven structures. Owing to the development of efficient techniques, methodologies and efforts described above, to date the number of structures deposited rose steeply up to 114,000 structures. NMR and X-ray crystallography are the world-wild techniques used, that account for 9.73% and 89.15% of the structure determined respectively. The protein structures solved (106,000) contribute to represent 2222 superfamilies and 1393 unique folding[37].

Despite of the extraordinary increase of structures deposited in the last decade the gap between them and the known annotated sequences (around 550,000) is still huge[38]. Furthermore Levitt *et al.* highlighted that the structure of novelty of proteins, defined as sequence identity lower than 25% from any others, has remained constant since 1992[39]. Nevertheless new approaches such as structural genomics (SG) provided an important contribution in structural protein determination being responsible for the 50% of novel structures deposited in the first years of the 2000s[39,40].

Even though the improvement of NMR, X-ray and SG techniques have speeded up the process of the structure determination, the scenario in which none experimental structures are available is still real. Therefore computational methods to predict three-dimensional structure have even now an important role.

I) Homology Modeling

Comparative or Homology Modeling (HM) is the most common methodology applied to overcome the lack of experimental three-dimensional structure. The technique is a multi-step process, recently summarized in eight stages[41]. The HM approach is basically based on the general observation that proteins with similar sequences reveal similar structures. Hence, the first step is the identification of a suitable homologous sequence, called template (Fig. 10). The three-dimensional structure of the template needs to be known and it is used as reference to build the novel model of the target protein.

The search of the most suitable template is still challenging and thus, different approaches have been used to date. In principle, all the algorithms perform an alignment of the target sequence to solved protein one, likes the proteins present in PDB database. Basic Local Alignment Search Tool (BLAST)[42] is one of the most common algorithm used to attempt the recognition of a plausible template sequence. The algorithm performs very well in finding highly similar structures, but the far related templates usually are missed. For this reason several implementations and different approaches were developed such as PSI-BLAST[43], phylogenetic analysis[44] and fold-recognition methods[45]. These search methods can return multiple templates from the database query and thus the retrieved structures need to be evaluated.

The identification of the most suitable template can be achieved considering different parameters which can be classified into three main classes: quality of the three-dimensional structures, conformation state of the protein and similarity with the target. Despite there is not a golden rule, the latter is commonly the first aspect that is taken into account. In particular both global and specific local identity or similarity of the sequences can be considered. Nevertheless, the conformation state of the templates is of utmost importance and the selection strictly depends on the purpose of the model application. The second class comprehends several different aspects, from the bound-state (unbound or in complex with agonist and antagonist) to protein state (active or inactive state) being the latter dependently to the presence of any mutation or to the type of ligand eventually present.

The quality of the template structure is measured by different parameters that provide evaluation of both the experimental analysis and the model obtained by fitting these data: the Resolution (R), the R-factor[46], free R-factor[47] and real-space R-factor[48]. Based on these considerations is feasible to select the most suitable template for the purpose and it is even possible to include multiple templates in the model generation.

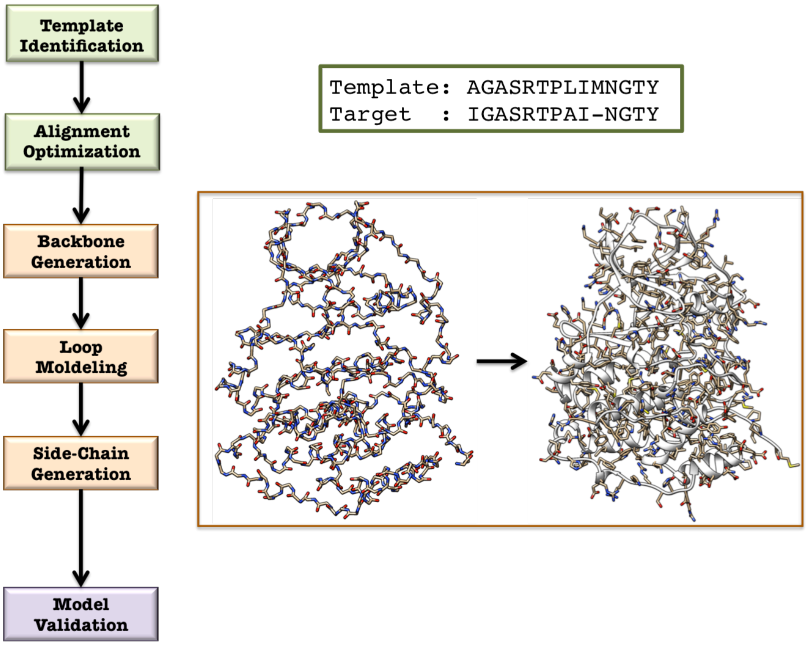


Figure 10 – Homology Modeling Process: The sequence of the stages to build and validate a model is shown.

The second step consists on the refinement of the preliminary alignment by the application of more precise methodologies and user knowledge. Nowadays the vast majority of the software are integrated with a secondary structure predictor, such as SSALN[49], that fragment the sequence and attempt to assign a secondary structure element to each of them. In addition, almost all the software are able to perform the multiple sequence alignment in order to identify the conserved regions.

Once the alignment is optimized, the next step is the model building that is a multi-step process, basically composed of three stages: backbone generation, loop modeling and side-chain modeling.

The backbone of the target structure is generated based either on the coordinates of the templates residues or based on the restrains (h-bond, torsional angles) gathered from the templates, that have to be satisfied during residues placement. The non-regular secondary structure elements in a protein structure are named loops and these regions can present deletions and insertions. For these sequence fragments, the atom coordinates need to be guessed. This can be overcome by using a knowledge-based or an energy-based approach.

At this point the backbone of the whole protein is generated and only the atoms of the side-chains remain to be predicted. Whenever the template and the target share the same residue, the atom coordinates are simply copied. On the other hand the side-chains are firstly built from scratch and optimized by using conformer libraries. Afterwards, the model geometry is optimized through minimization approaches or Molecular Dynamics (MD) simulations.

At the end, the model is subjected to different analysis in order to analyze the quality of the structure generated, such process is called the validation step. This evaluation is performed considering the bio-physical properties and divergence from the templates. For the former aspect both the phi-psi plot (Ramachandran plot) is analyzed and several web-server application such as Qmean server[50], can be considered. For the latter, it is commonly computed the Root Mean Square Deviation (RMSD) based on the alpha-carbon between the model and the templates.

From the first homology model published by Greer in 1980[51], implementations were developed to enhance the reliability of the model generated by these techniques, that are tested by the "Critical Assessment of protein Structure Prediction" (CASP) every two years since the 1994[52].

II) Target Binding Site Identification

Keeping in mind the ligand-protein interaction, it is essential the identification of the high-affinity binding site. Hence, whenever this information cannot be directly obtained by the visual inspection, neither from a related protein, the putative binding site must be predicted.

The scouting of plausible binding sites are performed by using one or multiple of the following analysis class: geometric-based, energy-based and conformations sampling. The geometric approach attempt to identify concave cavities by rolling probes through the surface and thus, their shapes and sizes are evaluated. An alternative method aims to probe a validated pharmacophore and evaluate its fitting based on the properties and constrains of the pharmacophore objects.

The energy-based approach is closely related with the previous one, but in addition different energies are calculated such as van der Waals, electrostatic, hydrogen-bonding, hydrophobic and solvent interactions.

Remembering that the protein structure considered is a single static structure, it is even possible to sample different conformations by using MD simulations. From the conformations explored during the time, the most different structures can be tested. The classic mechanics used in the MD simulation get trap the protein in local energy minima impeding an exhaustive sampling of the protein conformations. Hence, several methodologies were implemented to overcome this problem: tempered accelerated MD[53], replica exchange MD[54] and meta-dynamics[55] simulations.

III) Ligand-Protein Recognition

Once enough information about the protein structure and its binding site properties are collected, it is possible to start the molecular recognition of a ligand to the protein target.

The molecular recognition theory has constantly been changed during last century, starting from the "lock-and-key"[56], continuing with the "induce-fit"[57] and arriving finally to the "conformation selection"[58] models. However, it is important to mention that all of them are still considered.

Regarding as the key aspects of the ligand-protein binding event, this can be described as an equilibrium between the protein and ligand free in the bulk and the complex of them (Fig. 11). Firstly both the protein and the ligand interact with the solvent, thus the formation of the ligand-protein complex requires the desolvation of both molecules. For charged and polar parts of the molecule, this process arises a penalty that is only partially balanced by the electrostatic interactions and hydrogen-bonds formed. Conversely desolvation of non polar fragments produce a favorable gain in the entropy, that was defined as "Hydrophobic effect" by Kauzmann[59].

Currently, the common knowledge indicates that hydrophobic effect stabilizes the biomolecular complex, whereas the electrostatic interactions and hydrogen bonds provide the specificity for a certain target[60]. Apart from the solvent entropy changes, also the solute entropy has to be considered, such as translational, rotational, vibrational and dihedral restriction.

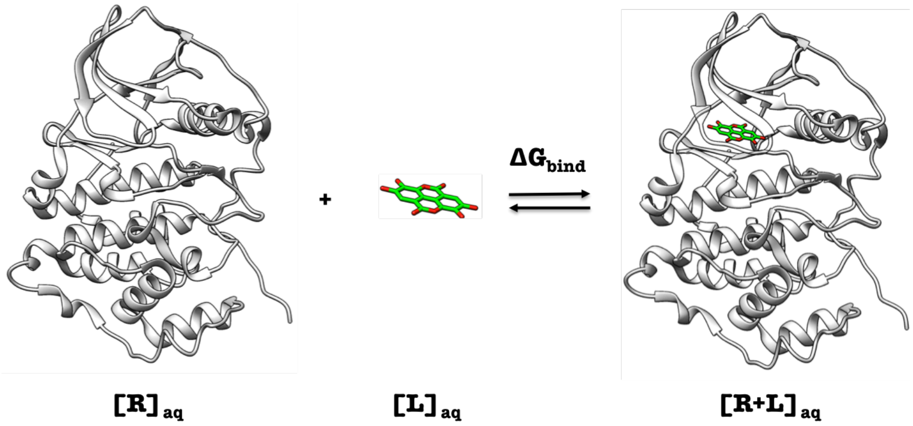


Figure 11 – Ligand-Protein binding equilibrium: protein and ligand in aqueous and in complex are reported, to describe the equilibrium that is established in a binding event.

IV) Molecular Docking

One of the most common tool used to predict ligand-protein complex is named molecular docking. This method can provide the correct binding mode by comparing the shape and chemical complementarity between the ligand and the protein. The description of the complex [RL] (eq. 2) can be obtained by considering some factors such as electrostatic, steric complementarity, hydrogen-bonding and ligand and protein strains, if these are flexible.

$$(eq.2) \qquad [RL]_{aq} \leftrightarrows [R]_{aq} + [L]_{aq}$$

More challenging is the prediction of the binding affinity and the consequent possibility to rank the compounds tested. The knowledge of the equilibrium established (eq. 3,4,5) is necessary to compute the binding affinity, hence entropy factors have also to be taken into account[61].

$$(eq. 3) \qquad K_a = \frac{K_{on}}{K_{off}} = \frac{[RL[}{[R][L]}$$

$$(eq. 4) \qquad K_d = {}^{1}/_{K_a}$$

$$(eq. 5) \qquad \Delta G_{binding} = -RTlnK_a$$

Where K_a is the affinity constant, represented by K_{on} and K_{off}, K_d is the reciprocal of the K_a, meaning the dissociation constant. In fact eq. 5 R is the gas constant and T represents the temperature.

The enthalpy and entropy calculation, necessary to obtain the free energy of binding ($\Delta G_{binding}$), are commonly computed separately. The first contribution is gathered directly from the molecular mechanics and the second one can be obtained with different methods (eq. 6,7).

$$(eq. 6) \qquad \Delta G_{binding} = \Delta H - T\Delta S$$

$$(eq. 7) \qquad \Delta G_{binding} = \Delta E_{MM} + \Delta G_{sol} - T\Delta S$$

Where H is the enthalpy, T is the temperature, S is the entropy, E_{MM} the energy obtained from the molecular mechanics and G_{sol} is the free energy value of the solvent.

The computation of the entropy effect is too high-time consuming, thus in the hit and lead identification it is usually not considered. However docking simulation depends on two aspects: the ligand placement into the binding site and evaluation of the complex generated. Due to the fact that usually the active conformation is not known, during the docking simulation it is necessary to explore the flexibility of the partners enrolled in the recognition. The time required to investigate the conformational space highly depends on the degree of freedom of the molecules, thus the normal approach keeps the protein fixed and only the ligand conformations are sampled.

Different algorithms have been developed in order to exhaustively explore the conformational space and they can be classified into three classes: systematic, stochastic and deterministic search.

The systematic search attempts to explore all the degree of freedom in a molecule, generating a combinatorial explosion. To deal with this problem several approaches were implemented, like termination criteria and incremental construction algorithm. The latter approach was implemented in several docking software and they worked into two different manners: in the first case, multiple fragments are docked and later linked together, in the second one, the ligand is divided into a core fragment, which is docked into the binding site and then, flexible parts are attached the placed core and their conformations are explored.

On the other hand, the stochastic search operates random changes, usually a single degree of freedom per iteration, and new conformations are evaluated based on a pre-defined probability function. However, the major concerning about this approach is the uncertainty of convergence, thus to overcome this problem, multiple runs are commonly performed. Two of the most famous algorithms of this class are Monte Carlo methods and genetic algorithms.

Regarding as deterministic search, this kind of simulations comprehends methodologies that produce exactly the same results when starting state and parameters are the same. The most popular and common used approaches are Molecular Dynamics simulation and energy minimization. As it was mentioned above MD is often unable to cross high-energy barriers, especially in short simulation time, thus only closed local minima of the energy surface are explored.

The second aspect in a docking simulation is the evaluation and ranking of ligand conformations based on designed scoring functions. Indeed these functions need to be enough reliable and fast to be able to screen a large amount of molecules. For this reason the techniques developed with the aim to accurately predict the free energy value are not suitable in this phase of the drug discovery. The worldwide scoring functions make assumptions and simplifications in order to speed up the process and they can be classified into three types: force field-based, empirical and knowledge-based.

The first class, called also Molecular Mechanics-based scoring function, estimates the binding free energy as the sum of receptor-ligand interaction energy and internal ligand energy (steric strain). These computed energies are usually the Coulombic formula for the electrostatic interactions and the van der Waals contribution defined by a Lennard-Jones potential function. The parameters of the latter potential has been integrated differently as 'harder', 12-6 potential, or 'softer', 8-4 potential (eq. 8).

$$(eq.8) \qquad E_{vdW} + E_{electrostatic} = \sum_{prot}\sum_{lig}[\left(\frac{A_{ij}}{d_{ij}^{a}} + \frac{B_{ij}}{d_{ij}^{b}}\right) + \varepsilon\frac{q_i q_j}{d_{ij}}]$$

Where *d* is the distance between the atoms, *A* and *B* potential parameters, *q* the atoms charge and *ε* the dielectric constant.

These calculations are computationally high costly, thus they require the introduction of cut-off distances for the treatment of the non-bonded interactions, being this distance an arbitrary value that infers the accuracy of the binding evaluation. Several software have integrated an additional parameter that takes into account a hydrogen bonding term, such as Gold[62] and Autodock[63] (eq. 9).

$$(eq.9) \qquad E_{vdW} + E_{H-bond} + E_{electrostatic} =$$

$$= \sum_{prot}\sum_{lig}[\left(\frac{A_{ij}}{d_{ij}^{12}} + \frac{B_{ij}}{d_{ij}^{6}}\right) + E(t)\left(\frac{C_{ij}}{d_{ij}^{12}} - \frac{D_{ij}}{d_{ij}^{10}}\right) + \varepsilon\frac{q_i q_j}{d_{ij}}]$$

Where *E(t)* is the angular weight factor and the other parameters are also present in eq. 8.

An alternative class of docking function contains empirical functions that attempt to reproduce experimental data, like binding and conformations energies. These functions are a weighted sum of a set of interaction terms some of whom have a counterpart in the force field-based approach. In addition, these functions added supplementary penalty terms that represent the amount of rotatable ligand bonds. The weights of the parameters are commonly obtained from a regression analysis, as it shown here for Chemscore function (eq.10).

$$\text{(eq. 10)}\quad \Delta G_{binding} =$$
$$= \Delta G_{H-bond} \sum_{H-bond} f(\Delta R, \Delta\alpha)$$
$$+ \Delta G_{metal} \sum_{metal} f(\Delta R, \Delta\alpha) + \Delta G_{lipo} \sum_{lipo} f(\Delta R)$$
$$+ \Delta G_{rotor} \sum_{rotor} f(P_{nl}, P'_{nl}) + \Delta G_0$$

Finally the knowledge-based scoring functions use a set of defined number of atom-type interactions to predict the complex structure by modeling relatively simple atomic interaction-pair potentials. As an example Potential of Mean Force (PMF)[64] evaluates the system energy changes as a function of specific reaction coordinates (eq 11).

$$(eq.11)\quad PMF = \sum_{prot} \sum_{lig} A_{ij}(d_{ij})\; A_{ij}(d_{ij}) = -k_B T ln[f^{j}_{Vol_{corr}}(r)\, \frac{\rho^{ij}_{seg}(r)}{\rho^{ij}_{bulk}}]$$

Nowadays, most of the efforts are focused in the development of force field-based scoring functions by implementing empirical terms that partially consider entropy contributions and their are called semi-empirical scoring functions.

At the end, the representation of the molecules is a crucial aspect to be considered that infer both in the accuracy and in the speeding of the process. Because of the ligand flexibility is usually explored, the molecule need to be treated in its full atoms representation whereas, the receptor can be considered into three ways: atomic, surface or grid representation.

The atomic representation was the first implemented and nowadays, it is still one of the most commonly used. In this approach the pair-wise atomic interactions are evaluated determining a non trivial computational complexity.

Concerning the surface-based representation is especially implemented for protein-protein docking tools. Their algorithm operate alignment of points on the surfaces and refining their position by minimizing the angle between the surfaces.

The last type of representation was introduced by Goodford[65] in which the receptor's energy contribution (usually electrostatic and van der Waals) are stored into grid points. This approach completely neglects the treatment of the protein atoms during the docking simulation, speeding up the calculation.

Apart from the ligand and protein, in several cases an additional actor has a crucial role in the ligand-protein recognition: the water molecules. In effect these molecules can mediate an interaction between the ligand and the protein, thus it is arguable whether the water molecules have to be taken into account in a docking simulation. The treatment of the waters is not simple and can also negative affects the posing of ligand compound due to the rigidity-induced of the protein side-chain that directly interact with the water molecules. Although most of the docking softwares, simply skipped the treatment of the water molecules by default, it is possible to analyze multiple crystallographic structures in order to evaluate the frequency of the presence of specific molecule. In this case the waters can be explicitly integrated as additional partner of the system on study.

Despite this alternative strategy, several software attempted to generalize the treatment of the waters by applying different algorithms. For example Slide accepts the pre-placement of the water molecules by the user, but they can be replaced by ligand features if the resulting pose is favorable. Differently FlexX evaluates the ligand placement and can place water molecules "on the fly" to enhance the quality of the complex generated. Another possibility is provided by GOLD which allows the user to switch on/off the treatment of the pre-placed water molecules and these are able to spin around their main axes with the aim to generate more favorable interactions.

V) Virtual High-Throughput Screening

The docking protocol is the core of the structure-based virtual high-throughput screening (SB-vHTS), that aims to identify putative hits out of huge amount of compounds.

As described above, docking protocols have advantages and disadvantages, and in addition, the accuracy is highly dependent on the case study. For this reason, docking software have integrated different search algorithms and scoring functions that respectively infer the conformational space explored and the rank of the pose generated. Due to their diversity, the comparison of the docking protocols is really tricky, however, different approaches were attempted to fulfill this important aspect[66]. In the last years two approaches were mostly accepted and applied to achieve the protocol docking selection task: a) capability to reproduce X-ray ligand pose; b) capability to accurately rank a set of known active compounds.

In the first case, the success of a program is measured by the RMSD between the X-ray ligand conformation and the predicted pose of each docking protocol. Afterwards, the computed values are statistically treated in different ways, making possible for example to evaluate the ratio of poses that have an RMSD lower than a certain value, which is unequivocally arbitrary.

The second approach attempts to evaluate the docking protocols performance by the capability to identify known active compounds out of a large set of inactive molecules. The calculated value is the enrichment factor which is defined as the ratio of the active compounds present in a x% of the ordered list of docking results. The performance of the docking protocols were deeply investigated in the last decade and such results were collected in several reviews. One f these revies was presented by Wandzikl in 2006, who collects 11 studies in which different software were compared with one of the approaches described above (Tab. 1)[67].

Table 1 - The table collects different jobs in which several docking software were compared. The number of the protein targets and the performance evaluation used are reported. Apporach A RMSD computed to the co-crystalized ligand; approach B enrichment factor of active compounds out of a set of inactive compounds.

Program Compared	Number of explored protein targets		Ref.
	Approach A	Approach B	
FlexX, DOCK, GOLD, LigandFit, Glide	69	-	68
GOLD, FlexX, Glide, Surflex	282	-	69
Glide, GOLD, FLexX, DOCK	-	9	70
Glide, FRED, FlexX	-	7	71
AutoDock, DOCK, FLexX, GOLD, ICM	37	11	72
Glide, GOLD, ICM	200	3	73
DOCK, FlexX, FRED, Glide, GOLD, Slide, Surflex, QXP	100	1	74
FlexX, GOLD, ICM, LigandFit, DOCK, QXP	11	-	75
DOCK, FLexX, GOLD, CDOCKER	41	-	76
DOCK, DockVision, Glide, GOLD	-	5	77
GOLD, QXP	-	1	78

In addition, another important aspect that inevitably has to be considered is the time necessary to dock a compound, since the more molecules are included in the library, the more is the time required to screen them. For this reason, in the hit-lead identification stage usually it is usually preferred to use a fast docking protocol, even if this could threat the accuracy of the complex quality generated. The most important information expected from vHTS is the “early recognition”, meaning that the active compounds have to be placed in the top-ranked result. Due to the fact that a pharmaceutical company usually needs to test more than 1 million of compounds to identify possible active ones, it is expected that vHTS reduces this amount to few thousands.

In theory, the placement of the active compounds into the ordered list can have the same significance only if the portion of the database experimentally tested present all the active compounds. Despite this pragmatic aspect, the vHTS reliability highly depends on the ability to rank the best compounds on the top of the ordered list. Hence the simplifications commonly used in the docking protocols generate false-positives and false-negatives in a vHTS simulation.

Nowadays, there is not a unique way to evaluate the VS performance. However two of the most famous methods are the enrichment factor (EF) and the Receiver Operating Characteristic (ROC) curve. In both the approaches, it is necessary to set up a training set, in particular a library of compounds that has to be populated with known active compounds (screened with the same assay) and decoys (compounds presumably inactive for the examined target).

In the EF it is simply computed how many active compounds are found within a fraction of the ordered list. Therefore, the EF is dependent on the fraction considered and it returns a number that can assume a value from 0 (worst performance) to 1 (best performance). The value of 0.5 is the minimum expected one indicating that the methodology is able to identify active compounds better than a random picking (reference). The result is commonly provided as a plot in which the x-axis reported the fraction of the database screened and in the y-axis shows the ratio of the active compounds found to the total of the actives present in the training set (Fig. 12)[76].

The major advantages of this evaluation is firstly that it does not weight equally all the compounds and secondly, the fact that it provides information about how much is enriched the fraction of compound wished to be screened.

The second way to evaluate the accuracy of a VS is the abovementioned ROC curve. The methodology classifies the compounds into four types, based on their activity and their rank in the ordered list: *i)* true positive (TP); *ii)* false negatives (FN); *iii)* false positives (FP); *iv)* true negatives (TN). The analysis returns a plot, in which for all possible threshold levels, the sensitivity is reported on the y-axis and the (1-specificity) is reported on x-axis (Fig. 13). The former is also called True Positive Rate (TPR) and the latter is also known as False Positive Rate (FPR) (eq. 12,13).

$$(eq.12) \qquad TPR = \frac{TP}{(TP + FN)}$$

$$(eq.13) \qquad FPR = \frac{FP}{(TN + FP)}$$

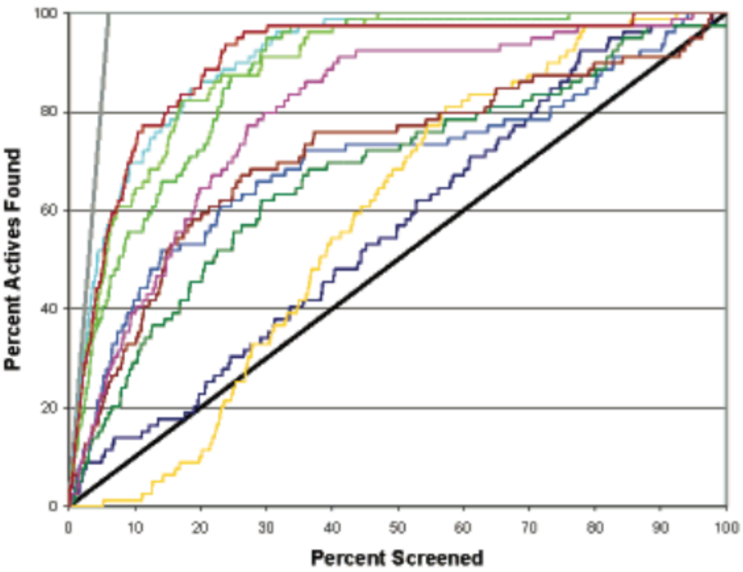


Figure 12 – Enrichment factor plot: An enrichment factor is reported for different docking softwares: Dock4-Energy (dark blue), Dockit-PMF (blue), FlexX-TotalScore (cyan), Fl0-Mcdock + FreE (green), Fred-ScreenScore (dark green), Glide-GScore (light green), Gold-Fitness (gold), Ligfit-OFF-Ligscore2 (dark orange), MOE-Uncorrected (magenta), MVP (red). In addition an ideal (gray) and a random (black) performances are reported.

An ideal classification generates a curve that immediately reach the top possible value and then continues as a line parallel to the x-axis. On the other hand, a random classification results in a diagonal line starting from the origin to the top right corner. Consequently, an acceptable ROC result provide a curve between the previous cases described above (Fig. 13). The ROC approach has an important advantage respect to EF, that is less sensible to the ”saturation effect”. The latter is defined as the dependency of the performance respect to the ratio of active compounds included into the training set (R_a). In particular the EF usually shows a wide divergence in the medium range of the fraction considered, while ROC only presents a slight divergence in the small-medium range (Fig. 14)[79]. The Fig. 14 shows two different studies elaborated by Hevener for the ROC[80] and by Warren for the EF plot[81].

The reported methodologies helped to evaluate the performance of the different docking protocols in term of capability to identify active compounds out of a range of decoys. However they do not provide any information about the absolute efficacy difference of active compounds identified. In effect, in these statistical considerations the scores are used only to rank them while the specific values are not investigated.

In conclusion, during the hit-lead generation, researchers have to face a variety of different starting points which are evaluated to correctly plan the discovery process. As presented above the limiting factor is usually the amount of the compounds that need to be screened, thus several approximations have to be applied in order to reduce the time that would be required for a deep investigation. Due to these considerations a lead generation process can be considered an acceptable success if any hit identified present a concentration for 50% inhibition (IC_{50}) of 10 micromolar[82] .

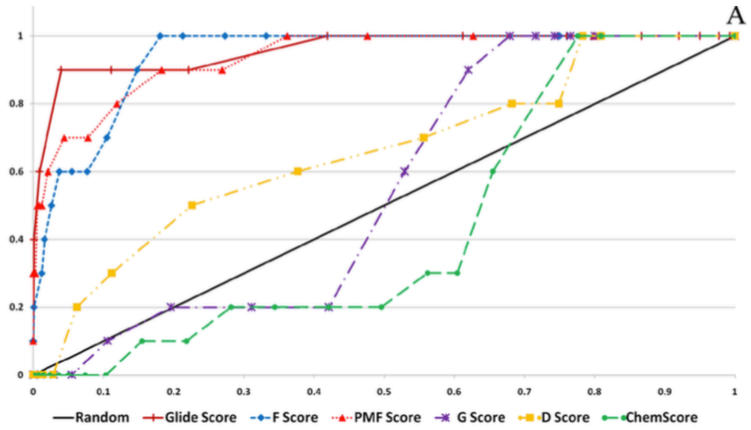


Figure 13 – Receiver operating characteristics (ROC) plot: ROC curves are reported for different docking approaches: Glide Score (red), F Score (blue), PMF Score (dark red), G Score (violet), D Score (yellow) and ChemScore (green). In addition a straight black line represents a random picking.

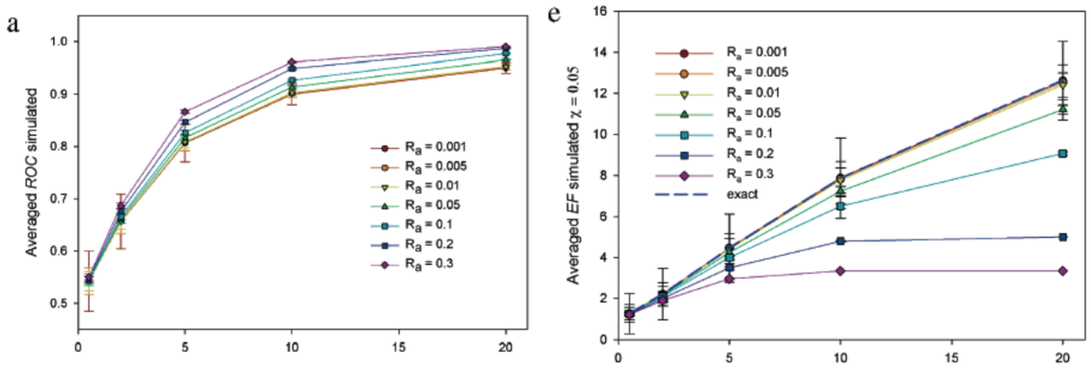


Figure 14 – ROC and Enrichment factor performances: For ROC and EF the performance is computed based on different ratio of actives present in the training set

1.3.2 Lead Optimization

The leads identified during the first stage of the early-phase drug discovery are investigated in more detail in order to understand how they can be optimized. The process can be considered a success whether the affinity value of any lead is reduced by 2/3 folds: IC_{50} in the nm range. Similarly to the lead generation phase, there are two main ways that a researcher can follow to optimize a lead: ligand-based and structure-based approach.

1.3.2.a Case 1: Unknown 3D Protein Structure

At this stage the ligand-based approach should have a disposal of a set of compounds and their related activities. In addition an active compounds can be modified and tested to investigate the influence of a different substitution to the activity (SAR). Since the pharmacophore approach that can infer to identify active compounds by evaluating the presence of specific chemical features, Quantitative Structure-Activity Relationship (QSAR) model[83] attempts to establish a relationship between physicochemical properties and the biological activity. QSAR has been used as a tool to predict and suggests new leads and also to rationalize the chemical modification of a congeneric set of compounds [84–86].

The major premise of QSAR is based on the hypothesis that similar structural or physicochemical properties elicit similar activity[87,88]. The model is generated from the active compounds available and their activity values. The descriptors, which can be both structural and physicochemical properties, are computed and analyzed in order to select the most suitable to describe the dependence between them and the biological response. Hence, the selected variables need to be related with the activity, but they should not represent similar biological or chemical parameters. For this reason whether multiple candidate parameters are suitable, it is necessary to withdraw any of them in order not to overestimates any property.

Several approaches can be considered to deal with the descriptors selection, such as genetic algorithms[89,90], principle component analysis (PCA)[91], artificial neural networks[92] and *k*-nearest neighbor[93]. Once the descriptor set is determined, a mathematical function is required with the aim to describe the relationship between the activity index (dependent variable) and the descriptors (independent variables). The mathematical models applied can be classified mainly into two categories: linear and non-linear models. In the first class the most common methods are the partial least squares (PLS)[94] and the multiple linear regression (MLR)[83]. On the other hand, whether the model need to be considered in a non-linear way, it is possible to take advantage from a machine learning methods like artificial neuronal networks[95] or support vector machines[96]. Afterwards the generated model has to be tested in a validation procedure that requires the correctly prediction of the activity of either a single or a set of compounds. The commonly validation types in this field are internal and external validation. The former foresees to exclude one (test compound) from the compound from the current training set (test compound), which is used to estimate the activity of the test compound. This procedure is iterated for all the compounds available. The two alternatives most used to this method are the "leave-on-out cross"[97] and the k-fold cross validation[98].

The success of the QSAR model is clearly high dependent on the selection of the descriptors, thus many efforts were done to improve these crucial elements. The most important ones are the molecular field descriptors that are derived from the interaction of probes and molecules. This approach was used by several groups that developed a variety of methods that are still the most used in QSAR studies; the most famous are Comparative Molecular Field Analysis (CoMFA)[99], Comparative Molecular Similarity Indices (CoMSIA)[100] and Comparative Molecular Moment Analysis (CoMMA)[101].

CoMFA was the first 3D QSAR method in which the shape-dependent steric and electrostatic properties of a molecule are used to correlate the biological activity. The method starts with the alignment of the molecules based either on a ligand crystal conformation or in the minimum-energy one. The latter is considered when none reference conformations are available, and without any doubt, this could return erroneous results because in this way it is assumed that this minimum energy conformation correspond to the bioactive conformer.

Then, the molecules are inserted into a grid defined by points in which the electrostatic (Coulombic potential) and the van der Waals contribution (steric Lennard-Jones potential) are computed. The use of the previous potentials determines unrealistic high energy values and, in addition, they neglect hydrophobicity and hydrogen-bond contributions. For this reason, an arbitrary cutoff value is integrated in the computation like force-field based scoring functions for docking protocols.

Meanwhile, CoMSIA is an improvement of the previous method, in which the probe used to compute the grid points contributions take also into account both, the hydrophobicity and hydrogen-bond donor and acceptor terms. An additional difference with CoMFA methodology is the use of a standard probe with a radius of 1 Å, and physicochemical properties like charge, hydrophobicity and hydrogen-bond equal to 1. Hence, the molecules properties calculation returns a value representing the similarity index as comparison with the standard probe. Moreover, the properties are computed thanks to a Gaussian function that smooth the binding interaction result, avoiding the necessity to introduce a cutoff as it was seen for CoMFA.

The last method, CoMMA, handles spatial moment descriptors that are used to obtain a unique value as result of the fitting of such parameters.

Although these approaches were widely used in the past, they still have an important role in the current drug discovery process.

1.3.2.b Case 2: Known 3D Protein Structure

The optimization of hits and leads with Structure Based (SB) relies its usefulness on the three-dimensional knowledge of the protein structure, plus any information obtained by previous VS and docking approaches. From these systems, it is possible to evaluate which are the most important residues, proving the hypothesis by using wet mutagenesis experiments and computational techniques. The parallel use of these approaches can provide useful and complementary information to understand the physico-chemical and pharmacological implication of mutations in a biological macromolecule[102,103]. The computational approach was used by Cristiani *et al.* to evaluate conformational changes in three variants of Factor VII (FVII)[104]. In this work, MD simulations were performed for both the wild type and the mutant structures. Then, the RMSD of alpha-carbon per each residue were computed versus the starting conformations. The results were reported as graphical representations (RainbowRMSD) in which the residues, the time of the simulation and the related RMSD were reported (Fig. 15). The analysis clearly shows how the mutation Arg79Gln infers the global flexibility of the protein, even in far region from the local mutation. The abovementioned analysis gave further information that can be taken into account with the aim to improve the physico-chemical complementarity between the ligand and the target protein. From a large amount of compounds, the top ranked ones of the VS simulation are analyzed and screened *in vitro*.

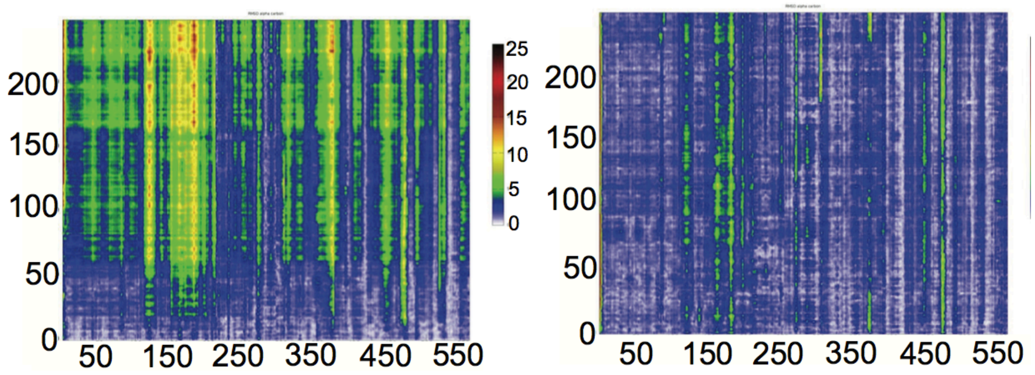


Figure 15 – RainbowRMSD graphs: The graphs represent the RMSD (colorimetric scale) of each alpha-carbon of all the protein residues (x-axis) along the simulation time (y-axis)

The reduced number of compounds to be studied in the following stage allows researchers to add more sophisticated methodologies for their investigation like MD related approaches, which are described below.

In the lead generation process, the docking protocols and in particular scoring functions need to use several approximations in order to reduce the time required to each calculation. Despite docking have proved to enrich hits better than random screening, this methodology arises false positives and false negatives. Moreover, docking tool is completely ineffective to grade compounds according to the binding affinities[105]. For this reason, the consideration of the aspects of the recognition events neglected by these techniques can improve the accuracy and the reliability of the prediction. In effect, a docking simulation is commonly performed with rigid protein, none or few water molecules are explicitly considered and entropy contribution, such as solvent-related terms, is ignored.

Thus, different methodologies were applied in order to predict how these aspects infer the complex generation. One of the most important tools available is the MD simulation that allows the exploration of the motion of both protein and ligand along the time. Molecular Dynamics simulation investigates the atoms motion, exploring their movement around an equilibrium position or even larger fluctuations. The basic theoretical assumption is that the energy of the system is a function of the atomic coordinates and their evolution along the time can be described by a Newton's equation.

$$(eq. 14) \qquad F_i = -\frac{\partial V_i}{\partial x_i} = m_i a_i$$

Hence, the force acting on each atom of the system is related to the potential energy V, with respect to atom position x, then by solving the equation based on a force it is possible to determine atom motions as a function of the time.

In 1977 Karplus *et al*. performed the first MD simulation of bovine pancreatic trypsin inhibitor. Despite of this simulation was performed in vacuum, these researchers introduced to the community a new challenging approach in medicinal chemistry. Thus, thanks to the development of new algorithms and the increase of the performance of the computational resources, MD is able to treat explicit water molecules and consequently, the solvent effect can be directly estimated from the simulation. In 2002 Karplus *et al*. investigated the role of the solvent in protein atomic fluctuations by performing different MD simulations, in which a combination of temperatures were applied to the protein and the solvent as it is shown in Table 2[106]. The average Mean Square Fluctuations (MSF) was computed for the backbone atoms and all heavy atoms of the protein in order to evaluate the dependence of the temperature applied. The Table 2 shows that the protein fluctuations highly depend on the solvent motion and this induces the internal motion of the protein. In the same work they explained that at 80 K the protein could be considered freeze and even if the temperature of the protein is set to 300 K, the internal motional barrier is too high and thus, dominates the dynamic of the protein.

Table 2 - The table reported the average mean square deviation of backbone and heavy atoms in different MD simulations. The temperature of the solute and solvent are varied

Average mean square fluctuations		
Temperature[1]	**Backbone (Å^2)**	**Heavy Atoms (Å^2)**
P300/S300	0.23	0.36
P180/S300	0.18	0.28
P300/S180	0.09	0.13
P180/S180	0.09	0.13
[1]Tempereature in Kelvin. ‘P’ refers to protein and ‘S’ to solvent		

In the last years, the inferring of the solvent contribution to the ligand interactions and binding to a protein was investigated and different approaches were proposed, such as “WaterMap” tool by Abel *et al*[107]. This approach starts from a MD simulation that is used to generate the positions of water sites and for each of them, the free energy of the water displacement is computed by inhomogeneous solvation theory. The displacement of unfavorable water molecules by the ligand can allow the establishment of interactions of complementary groups with the protein, and these are known to be the principal driving force for the recognition events[69]. This innovative tool also provide information about the so called dry region of an active site of the protein, in which the presence of water molecules is so unfavorable that a void is formed. These regions return a value under a specified threshold when WaterMap is applied, gathering important hot spot for ligand placement. Then, Wang *et al*. implemented an additional attribute to WaterMap function that takes into account when ligand atoms occupy these regions[108]. The relevance of the presence of specific water molecules was also investigated by Sabbadin *et al*. The MD simulation of a complex is compared with another Apo-form simulation and the fluctuations of the water molecules (RMSF) is monitored. The post-processing analysis returns a bidimensional graph, Water Fluid Dynamics (WFD), which make easier the identification of the protein hot-spots. Nowadays, Molecular Dynamics is extensively used in combination with molecular docking due to the fact that MD can neutralize the defects, or at least, reduce the inefficacy beyond the scoring function of the docking. In effect, it is important to remember that the ligand protein recognition can be explained only thanks to an ensemble property such as binding free energy, whereas docking simulation focus its attention to a single snapshot of the protein.

A post-processing methodology was proposed by Sabbadin *et al.*[109] in which the different docking poses were subjected to a 60 ns of MD simulation and three aspects were analyzed: *i)* evolution of interaction fingerprints (dynamic IEFs)*; ii)* the ligand fluctuation (RMSD)*; iii)* the cumulative sum of interaction energy normalized by the ligand atom coordinates deviation.

Among all the above aspect, the latter provide the most interesting information by evaluating the electrostatic and hydrophobic contributions along all the simulation (Fig. 16). The result is presented as a plot in which the x-axis reported the time and y-axis reported the sum of the energies for the current frame plus all the previous. In addition, these values are normalized on the RMSD of the ligand heavy atoms. The slope tendency of the curve describes both the interaction strength and the positional deviation of the ligand poses. Thus, a strong binder could generate a straight line with a very negative slope value.

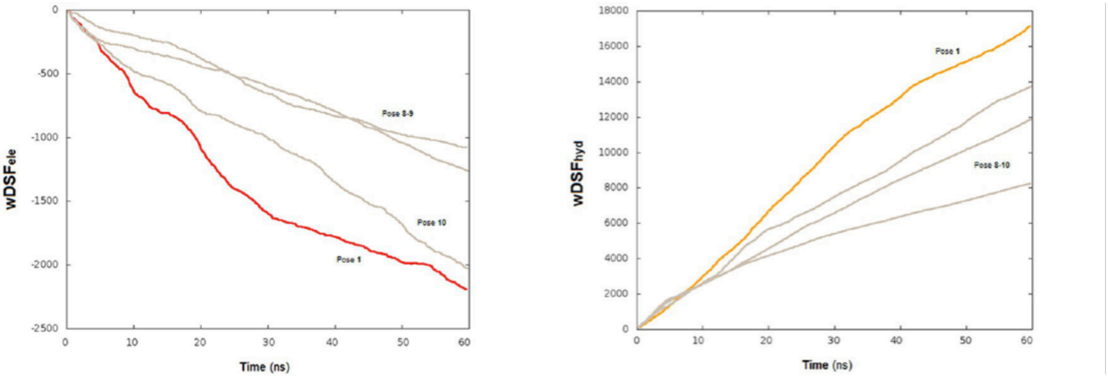


Figure 16 – Cumulative energy plots: The sum of cumulative energy normalized by the RMSD of the ligand are reported. On the left, the electrostatic contribution is shown, whereas on the right the hydrophobic contribution is presented.

MD simulation combined to molecular docking provides useful information and a wise strategy to recognize the most stable complex with respect to ligand conformation. Although several alternatives were introduced to enhance the prediction accuracy of the docking-based approaches, the scoring functions are still poorly or even erroneously able to evaluate binding free energy. For this reason, the binding free energy (eq. 15) estimation has been deeply attempted and it is still the major focusing of researchers. In these years, several MD-based methodologies were developed in order to improve the computational accuracy in this field.

$$(eq. 15) \qquad \Delta G_{bind} = E_{MM} - T\Delta S_{solute} + \Delta G_{solvent}$$

The first term is derived from the Force Field, which generally are described by the following equation:

$$(eq.16)\ E = \sum_{bonds} K_r (r - r_{eq})^2 + \sum_{angles} K_\vartheta (\vartheta - \vartheta_{eq})^2 + \sum_{dihedrals} \frac{V_n}{2} [1 + \cos(n\phi - \gamma)] + \sum_{i<j} \frac{[A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\varepsilon R_{ij}}]$$

In the eq. 16 the last term is the sum of the van der Waals and electrostatic contributions while the other terms present a constant (K_r, K_ϑ, V_n) multiply for the difference from the equilibrium stat of the relative term. The other terms of the eq. 15 can be obtained by using high-time computational methodologies. The most famous rigorous approaches that partially address the issue abovementioned are: free energy perturbation[110] (FEP), thermodynamic integration[111] (TI), ligand interaction energy approach[112,113] (LIE), λ-dynamics[114], ligand interactions scanning[115] and MM-(PB/GB)SA[116].

The FEP and TI are the most rigorous ways and the most used approaches to predict the binding free energy. In these methodologies, the difference in the binding free energy between two similar states is computing by mutating one state to the other one through multiple intermediated states ("Computational Alchemy"). The thermodynamic cycle perturbation method (Fig. 17) allows to precisely calculate the relative binding free energy properly due to the fact that it is a state of function (eq. 17). This process is done for both the complex and the ligand free in the bulk solvent. Hence, in this alchemical simulation the potential energy function from C_1 is slightly converted to C_2 during a MD simulation.

Without any doubt, the accuracy of the result obtained is highly dependent on the relevant configurations considered[117]; thus, the post-docking approach described above could be used to retrieve the most suitable starting point.

Alternative methods such as MM-(GB/PB)SA takes into account the bulk solvent effects and its energy can be divided into two terms, nonpolar and polar effects that can be calculated separately.

$$\begin{tikzcd}
M & + & C_1 \arrow[r, "1"] \arrow[d, "3"] & C_1M \arrow[d, "4"'] \\
M & + & C_2 \arrow[r, "2"] & C_2M
\end{tikzcd}$$

Figure 17 – Computational alchemic cycle: The Alchemic cycle for the conversion from one ligand to another one is summarized. M, C_1 C_2 represent respectively the protein target and the two ligand in the bulk solvent, while the C_1M and C_2M are the complex generated.

$$(eq.17)\qquad \Delta\Delta G = \Delta G_2^\circ - \Delta G_1^\circ = \Delta G_4^\circ - \Delta G_3^\circ$$

The nonpolar effect is obtained by computing the area of the solute and the energy value is estimated as the cost to generate a cavity in the solvent for guesting the solute. This energy quantity considers the water molecules that are freeze around the solute plus the van der Waals interaction energy between solute and solvent.

On the other hand, the polar solvation is computing with continuum solvent approximations provided by either the Poisson-Boltzmann[118] equation or the generalized Born[119,120] equation.

The application of these calculations allow to treat the third term of the eq. 15 (the solvent energy term), as the sum of the abovementioned solvation effects: G_{PB} or G_{GB} and G_{SA} (eq. 18).

$$(eq. 18) \qquad G = E_{MM} - TS_{solute} + G_{PB/GB} + G_{SA}$$

In the MM-(PB/GB)SA methods the solute entropy is approximated by classical statistical expression and normal-mode analysis. The computational time required to retrieve the results is less demanding than FEP or TI and, at the same time, conserves their accuracy[121,122]. In the last years several groups have evaluated the performance of these methods by calculating the correlation with experimental measure data[123–125].

As the FEP and TI these, the MM-(PB/GB)SA methods suffer the dependency of the set of coordinates provided, thus the selection of the starting complex is of utmost importance. The docking pose can directly provide the starting set of coordinates or the conformation can be refined by a minimization or MD approach. Thus these strategies allow to select a starting conformation that can be either a local minimum or the most explored one during a simulation.

Despite of MD scoring functions are more reliable than docking scores they are still not perfect and the time require is still too high to be intensively used in a vHTS. In addition the solvent models applied are inefficient to evaluate the water molecules interactions between solute and solvent interface.

1.3.3 ADMET

Once the lead compounds are optimized from a pharmacodynamic point of view is necessary to test their pharmacokinetic. The latter consists of Absorption, Distribution, Metabolism, Excretion and Toxicity (ADMET). Historically these properties have been evaluated by *in vitro* and *in vivo* studies that are undoubtedly high costly. In 2003 Gilbert *et al.* estimated that the preclinical phase determine the highest attrition rate in the R&D[126]. Hence the balance of potency, selectivity and ADMET properties need to be balanced to propose a drug candidate.

CADD methodologies were developed also to deal with this aspect of the drug discovery process.

In silico modeling of ADMET properties can be classified into three main categories: *i)* molecular modeling*; ii)* physiologically based pharmacokinetic modeling; *iii)* statistical modeling[127]*.*

The pharmacokinetic properties comprehend an high diversity types of parameters, which are mostly obtained from descriptors as it was seen previously for the Ligand-Based approaches, as seen previously. The descriptors are determined from experimental data available and thus, these values are fitted with mathematical models such as Partial Least Square (PLS), Multiple Linear Regression (MLR), Artificial Neural Networks (ANNs) and Decision Trees (DT). As it was mentioned above, several of them are linear techniques (PLS, MLR) and others can handle non-linear functions (ANNs). Moreover, the latter is known to overtraining the data.

Molecular modeling approaches can be applied in particular to investigate metabolic aspects based on metabolizing enzymes. For example the human cytochrome P450 (CYP) enzymes family are one of the most important phase I drug-metabolizing enzymes implicated into detoxification of xenobiotic compounds, bio-activation of non-toxic and toxic intermediates and pro-carcinogens. Moreover CYPs are also involved in drug-drug interactions (DDIs) mediated by drug inhibition and induction[128]. Due to the fact that the 75% of the drugs in the market are metabolized by this enzyme family, the researchers have made a big efforts to produce as many three-dimensional structures as possible, three-dimensional structures. By using the protein structure knowledge is possible to apply similar techniques as those seen in the Structure-Based approaches. It is clear that the enhancing of the prediction accuracy of these methodologies will have a strong impact in the R&D spending of a big pharmaceutical industry.

1.4 Bibliography

1. Tollenaere, J. P. The role of structure-based ligand design and molecular modelling in drug discovery. *Pharm. World Sci.* **18,** 56–62 (1996).
2. Ooms, F. Molecular modeling and computer aided drug design. Examples of their applications in medicinal chemistry. *Curr. Med. Chem.* **7,** 141–158 (2000).
3. Bolten, B. M. & DeGregorio, T. Trends in development cycles. *Nat. Rev. Drug Discov.* **1,** 335–336 (2002).
4. PR Tufts CSDD 2014 Cost Study | Tufts Center for the Study of Drug Development. at <http://csdd.tufts.edu/news/complete_story/pr_tufts_csdd_2014_cost_study>
5. Oprea, T. I. Current trends in lead discovery: are we looking for the appropriate properties? *J. Comput. Aided Mol. Des.* **16,** 325–334 (2002).
6. Pammolli, F., Magazzini, L. & Riccaboni, M. The productivity crisis in pharmaceutical R&amp;D. *Nat. Rev. Drug Discov.* **10,** 428–438 (2011).
7. Research, C. for D. E. and. New Drugs at FDA: CDER's New Molecular Entities and New Therapeutic Biological Products. at <http://www.fda.gov/Drugs/DevelopmentApprovalProcess/DrugInnovation/ucm20025676.htm>
8. Paul, S. M. *et al.* How to improve R&amp;D productivity: the pharmaceutical industry's grand challenge. *Nat. Rev. Drug Discov.* (2010). doi:10.1038/nrd3078
9. Ehrlich, P. Über den jetzigen Stand der Chemotherapie. *Berichte Dtsch. Chem. Ges.* **42,** 17–47 (1909).
10. Sharma, A. & Lal, S. P. Tanimoto Based Similarity Measure for Intrusion Detection System. *J. Inf. Secur.* **02,** 195–201 (2011).
11. Guner, O. F., Hughes, D. W. & Dumont, L. M. An integrated approach to three-dimensional information management with MACCS-3D. *J. Chem. Inf. Comput. Sci.* **31,** 408–414 (1991).
12. Pötter, T. & Matter, H. Random or rational design? Evaluation of diverse compound subsets from chemical structure databases. *J. Med. Chem.* **41,** 478–488 (1998).
13. Lipinski, C. A. Lead- and drug-like compounds: the rule-of-five revolution. *Drug Discov. Today Technol.* **1,** 337–341 (2004).
14. Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* **64,** 4–17 (2012).

15. Hann, M. M., Leach, A. R. & Harper, G. Molecular Complexity and Its Impact on the Probability of Finding Leads for Drug Discovery. *J. Chem. Inf. Model.* **41,** 856–864 (2001).

16. Oprea, T. I., Davis, A. M., Teague, S. J. & Leeson, P. D. Is There a Difference between Leads and Drugs? A Historical Perspective. *J. Chem. Inf. Model.* **41,** 1308–1315 (2001).

17. 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. *J. Med. Chem.* **53,** 2719–2740 (2010).

18. Baell, J. & Walters, M. A. Chemical con artists foil drug discovery. *Nat. Comment* **513,** (2014).

19. Vilar, S. *et al.* Similarity-based modeling in large-scale prediction of drug-drug interactions. *Nat. Protoc.* **9,** 2147–2163 (2014).

20. Kurogi, Y. & Guner, O. F. Pharmacophore modeling and three-dimensional database searching for drug design using catalyst. *Curr. Med. Chem.* **8,** 1035–1055 (2001).

21. van Drie, J. H. Pharmacophore discovery-lessons learned. *Curr. Pharm. Des.* **9,** 1649–1664 (2003).

22. Jorgensen, W. L. Rusting of the Lock and Key Model for Protein-Ligand Binding. *Science* **254,** 954–955 (1991).

23. Michal Vieth, J. D. H. Do active site conformations of small ligands correspond to low free-energy solution structures? J Comput Aided Mol Des. *J. Comput. Aided Mol. Des.* **12,** 563–72 (1998).

24. López-Rodríguez, M. L. *et al.* Optimization of the Pharmacophore Model for 5-HT $_7$ R Antagonism. Design and Synthesis of New Naphtholactam and Naphthosultam Derivatives. *J. Med. Chem.* **46,** 5638–5650 (2003).

25. Lee, C.-H., Huang, H.-C. & Juan, H.-F. Reviewing Ligand-Based Rational Drug Design: The Search for an ATP Synthase Inhibitor. *Int. J. Mol. Sci.* **12,** 5304–5318 (2011).

26. Yang, S.-Y. Pharmacophore modeling and applications in drug discovery: challenges and recent advances. *Drug Discov. Today* **15,** 444–450 (2010).

27. All Nobel Prizes in Chemistry. at <http://www.nobelprize.org/nobel_prizes/chemistry/laureates/>

28. Bambini, S. & Rappuoli, R. The use of genomics in microbial vaccine development. *Drug Discov. Today* **14,** 252–260 (2009).

29. Lundstrom, K. Micro-RNA in disease and gene therapy. *Curr. Drug Discov. Technol.* **8,** 76–86 (2011).

30. Wuethrich, K. The development of nuclear magnetic resonance spectroscopy as a technique for protein structure determination. *Acc. Chem. Res.* **22,** 36–44 (1989).

31. Allen, F. H. *et al.* The development of versions 3 and 4 of the Cambridge Structural Database System. *J. Chem. Inf. Comput. Sci.* **31,** 187–204 (1991).

32. Böhm, H.-J. & Klebe, G. What Can We Learn from Molecular Recognition in Protein–Ligand Complexes for the Design of New Drugs? *Angew. Chem. Int. Ed. Engl.* **35,** 2588–2614 (1996).

33. G Wagner, S G Hyberts & Havel, T. F. NMR Structure Determination in Solution: A Critique and Comparison with X-Ray Crystallography. *Annu. Rev. Biophys. Biomol. Struct.* **21,** 167–198 (1992).

34. Wilfred van Gunsteren & Berendsen, H. Computer simulation as a tool for tracing the conformational differences between proteins in solution and in crystalline state. *J. Mol. Biol.* **176,** 559–564 (2015).

35. Jacobson, M. P., Friesner, R. A., Xiang, Z. & Honig, B. On the Role of the Crystal Environment in Determining Protein Side-chain Conformations. *J. Mol. Biol.* **320,** 597–608 (2002).

36. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28,** 235–242 (2000).

37. RCSB PDB. at <http://www.rcsb.org/pdb/static.do?p=general_information/pdb_statistics/index.html>

38. UniProtKB/Swiss-Prot Release 2015_12 statistics. at <http://web.expasy.org/docs/relnotes/relstat.html>

39. Levitt, M. Growth of novel protein structural data. *Proc. Natl. Acad. Sci.* **104,** 3183–3188 (2007).

40. Lundstrom, K. Structural genomics and drug discovery. *J. Cell. Mol. Med.* **11,** 224–238 (2007).

41. Venselaar, H. *et al.* Homology modelling and spectroscopy, a never-ending love story. *Eur. Biophys. J.* **39,** 551–563 (2009).

42. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).

43. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402 (1997).

44. Cheng, J. A multi-template combination algorithm for protein comparative modeling. *BMC Struct. Biol.* **8,** 18 (2008).

45. Jones, D. T. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences1. *J. Mol. Biol.* **287,** 797–815 (1999).

46. Bränd´en, C.-I. & Alwyn Jones, T. Between objectivity and subjectivity. *Nature* **343,** 687–689 (1990).

47. Brünger, A. T. Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature* **355,** 472–475 (1992).

48. Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47,** 110–119 (1991).

49. Qiu, J. & Elber, R. SSALN: an alignment algorithm using structure-dependent substitution matrices and gap penalties learned from structurally aligned protein pairs. *Proteins* **62,** 881–891 (2006).

50. Benkert, P., Tosatto, S. C. E. & Schomburg, D. QMEAN: A comprehensive scoring function for model quality assessment. *Proteins* **71,** 261–277 (2008).

51. Greer, J. Model for haptoglobin heavy chain based upon structural homology. *Proc. Natl. Acad. Sci.* **77,** 3393–3397 (1980).

52. Moult, J., Pedersen, J. T., Judson, R. & Fidelis, K. A large-scale experiment to assess protein structure prediction methods. *Proteins Struct. Funct. Bioinforma.* **23,** ii–iv (1995).

53. Abrams, C. F. & Vanden-Eijnden, E. Large-scale conformational sampling of proteins using temperature-accelerated molecular dynamics. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 4961–4966 (2010).

54. Sugita, Y. & Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **314,** 141–151 (1999).

55. Laio, A. & Parrinello, M. Escaping free-energy minima. *Proc. Natl. Acad. Sci.* **99,** 12562–12566 (2002).

56. Fischer, E. Einfluss der Configuration auf die Wirkung der Enzyme. *Berichte Dtsch. Chem. Ges.* **27,** 2985–2993 (1894).

57. Foote, J. & Milstein, C. Conformational isomerism and the diversity of antibodies. *Proc. Natl. Acad. Sci.* **91,** 10370–10374 (1994).

58. Koshland, D. E. Application of a Theory of Enzyme Specificity to Protein Synthesis*. *Proc. Natl. Acad. Sci. U. S. A.* **44,** 98–104 (1958).

59. Kauzmann, W. Some factors in the interpretation of protein denaturation. *Adv. Protein Chem.* **14,** 1–63 (1959).

60. Brooijmans, N. & Kuntz, I. D. Molecular Recognition and Docking Algorithms. *Annu. Rev. Biophys. Biomol. Struct.* **32,** 335–373 (2003).

61. Kitchen, D. B., Decornez, H., Furr, J. R. & Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discov.* **3,** 935–949 (2004).

62. Jones, G., Willett, P., Glen, R. C., Leach, A. R. & Taylor, R. Development and validation of a genetic algorithm for flexible docking1. *J. Mol. Biol.* **267,** 727–748 (1997).

63. Huey, R., Morris, G. M., Olson, A. J. & Goodsell, D. S. A semiempirical free energy force field with charge-based desolvation. *J. Comput. Chem.* **28,** 1145–1152 (2007).

64. Muegge, I. A knowledge-based scoring function for protein-ligand interactions: Probing the reference state. *Perspect. Drug Discov. Des.* **20,** 99–114 (2000).

65. Goodford, P. J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* **28,** 849–857 (1985).

66. Cole, J. C., Murray, C. W., Nissink, J. W. M., Taylor, R. D. & Taylor, R. Comparing protein-ligand docking programs is difficult. *Proteins Struct. Funct. Bioinforma.* **60,** 325–332 (2005).

67. Wandzik, I. Current molecular docking tools and comparisons thereof. *MATCH* **55,** 271–278 (2006).

68. Kontoyianni, M., McClellan, L. M. & Sokol, G. S. Evaluation of Docking Performance: Comparative Data on Docking Algorithms. *J. Med. Chem.* **47,** 558–565 (2004).

69. Friesner, R. A. *et al.* Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **47,** 1739–1749 (2004).

70. Halgren, T. A. *et al.* Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening. *J. Med. Chem.* **47,** 1750–1759 (2004).

71. Schulz-Gasch, T. & Stahl, M. Binding site characteristics in structure-based virtual screening: evaluation of current docking tools. *J. Mol. Model.* **9,** 47–57 (2003).

72. Bursulaya, B. D., Totrov, M., Abagyan, R. & Brooks, C. L. Comparative study of several algorithms for flexible ligand docking. *J. Comput. Aided Mol. Des.* **17,** 755–763 (2003).

73. Perola, E., Walters, W. P. & Charifson, P. S. A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. *Proteins Struct. Funct. Bioinforma.* **56,** 235–249 (2004).

74. Kellenberger, E., Rodrigo, J., Muller, P. & Rognan, D. Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins Struct. Funct. Bioinforma.* **57,** 225–242 (2004).

75. Kroemer, R. T. *et al.* Assessment of Docking Poses: Interactions-Based Accuracy Classification (IBAC) versus Crystal Structure Deviations. *J. Chem. Inf. Comput. Sci.* **44,** 871–881 (2004).

76. Cummings, M. D., DesJarlais, R. L., Gibbs, A. C., Mohan, V. & Jaeger, E. P. Comparison of Automated Docking Programs as Virtual Screening Tools. *J. Med. Chem.* **48,** 962–976 (2005).

77. Erickson, J. A., Jalaie, M., Robertson, D. H., Lewis, R. A. & Vieth, M. Lessons in Molecular Recognition: The Effects of Ligand and Protein Flexibility on Molecular Docking Accuracy. *J. Med. Chem.* **47,** 45–55 (2004).

78. Cotesta, S. *et al.* Virtual screening to enrich a compound collection with CDK2 inhibitors using docking, scoring, and composite scoring models. *Proteins* **60,** 629–643 (2005).

79. Truchon, J.-F. & Bayly, C. I. Evaluating Virtual Screening Methods: Good and Bad Metrics for the 'Early Recognition' Problem. *J. Chem. Inf. Model.* **47,** 488–508 (2007).

80. Hevener, K. E. *et al.* Validation of Molecular Docking Programs for Virtual Screening against Dihydropteroate Synthase. *J. Chem. Inf. Model.* **49,** 444–460 (2009).

81. Warren, G. L. *et al.* A Critical Assessment of Docking Programs and Scoring Functions. *J. Med. Chem.* **49,** 5912–5931 (2006).

82. Jorgensen, W. L. The many roles of computation in drug discovery. *Science* **303,** 1813–1817 (2004).

83. Scior, T. *et al.* How to recognize and workaround pitfalls in QSAR studies: a critical review. *Curr. Med. Chem.* **16,** 4297–4313 (2009).

84. Wang, M.-Y. *et al.* Synthesis, biological evaluation and 3D-QSAR studies of imidazolidine-2,4-dione derivatives as novel protein tyrosine phosphatase 1B inhibitors. *Eur. J. Med. Chem.* **103,** 91–104 (2015).

85. Fu, Y., Li, H., Zhao, Q. & Ye, F. CoMFA study on novel Substitute oxazolidine derivatives. *Adv. Mater. Res.* **345,** 320–325 (2012).

86. Sharma, H. *et al.* Synthesis, biological evaluation and 3D-QSAR studies of 3-keto salicylic acid chalcones and related amides as novel HIV-1 integrase inhibitors. *Bioorg. Med. Chem.* **19,** 2030–2045 (2011).

87. Akamatsu, M. Current state and perspectives of 3D-QSAR. *Curr. Top. Med. Chem.* **2,** 1381–1394 (2002).

88. Verma, R. P. & Hansch, C. Camptothecins: a SAR/QSAR study. *Chem. Rev.* **109,** 213–235 (2009).

89. Chen, H., Zhou, J. & Xie, G. PARM: A Genetic Evolved Algorithm To Predict Bioactivity. *J. Chem. Inf. Comput. Sci.* **38,** 243–250 (1998).

90. Rogers, D. & Hopfinger, A. J. Application of genetic function approximation to quantitative structure-activity relationships and quantitative structure-property relationships. *J. Chem. Inf. Comput. Sci.* **34,** 854–866 (1994).

91. Xue, L., Godden, J. W. & Bajorath, J. Evaluation of descriptors and mini-fingerprints for the identification of molecules with similar activity. *J. Chem. Inf. Comput. Sci.* **40,** 1227–1234 (2000).

92. James H. Wikel, E. R. D. The Use of Neural Networks for Variable Selection in QSAR. *Bioorganic Amp Med. Chem. Lett.* **3,** 645–651 (1993).

93. Itskowitz, P. & Tropsha, A. kappa Nearest neighbors QSAR modeling as a variational problem: theory and applications. *J. Chem. Inf. Model.* **45,** 777–785 (2005).

94. Geladi, P. & Kowalski, B. R. Partial least-squares regression: a tutorial. *Anal. Chim. Acta* **185,** 1–17 (1986).

95. Jain, A. N., Koile, K. & Chapman, D. Compass: Predicting Biological Activities from Molecular Surface Properties. Performance Comparisons on a Steroid Benchmark. *J. Med. Chem.* **37,** 2315–2327 (1994).

96. Shahlaei, M., Fassihi, A. & Saghaie, L. Application of PC-ANN and PC-LS-SVM in QSAR of CCR1 antagonist compounds: A comparative study. *Eur. J. Med. Chem.* **45,** 1572–1582 (2010).

97. Kohavi, R. A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection. in *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2* 1137–1143 (Morgan Kaufmann Publishers Inc., 1995). at <http://dl.acm.org/citation.cfm?id=1643031.1643047>

98. Weiss, S. M. & Kulikowski, C. A. *Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems*. (Morgan Kaufmann Publishers Inc., 1991).

99. Cramer, R. D., Patterson, D. E. & Bunce, J. D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **110,** 5959–5967 (1988).

100. Klebe, G., Abraham, U. & Mietzner, T. Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. *J. Med. Chem.* **37,** 4130–4146 (1994).

101. Silverman, B. D. & Platt, D. E. Comparative molecular moment analysis (CoMMA): 3D-QSAR without molecular superposition. *J. Med. Chem.* **39,** 2129–2140 (1996).

102. Masso, M., Lu, Z. & Vaisman, I. I. Computational mutagenesis studies of protein structure-function correlations. *Proteins Struct. Funct. Bioinforma.* **64,** 234–245 (2006).

103. Schwans, J. P. *et al.* Experimental and Computational Mutagenesis To Investigate the Positioning of a General Base within an Enzyme Active Site. *Biochemistry (Mosc.)* **53,** 2541–2555 (2014).

104. Cristiani, A. *et al.* Conformational Changes of Congenital FVII Variants with Defective Binding to Tissue Factor ARG304GLN (FVII Padua), ARG 304TRP (FVII Nagoya) and ARG79GLN (FVII Shinjo or Tondabayashi). *Int. J. Biomed. Sci. IJBS* **9,** 185–193 (2013).

105. Pearlman, D. A. Evaluating the Molecular Mechanics Poisson−Boltzmann Surface Area Free Energy Method Using a Congeneric Series of Ligands to p38 MAP Kinase. *J. Med. Chem.* **48,** 7796–7807 (2005).

106. Karplus, M. & McCammon, J. A. Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.* **9,** 646–652 (2002).

107. Abel, R., Young, T., Farid, R., Berne, B. J. & Friesner, R. A. Role of the active-site solvent in the thermodynamics of factor Xa ligand binding. *J. Am. Chem. Soc.* **130,** 2817–2831 (2008).

108. Wang, L., Berne, B. J. & Friesner, R. A. Ligand binding to protein-binding pockets with wet and dry regions. *Proc. Natl. Acad. Sci.* **108,** 1326–1330 (2011).

109. Sabbadin, D., Ciancetta, A. & Moro, S. Bridging Molecular Docking to Membrane Molecular Dynamics To Investigate GPCR–Ligand Recognition: The Human A2A Adenosine Receptor as a Key Study. *J. Chem. Inf. Model.* **54,** 169–183 (2014).

110. Kollman, P. Free energy calculations: applications to chemical and biochemical phenomena. *Chem. Rev.* **93,** 2395–2417 (1993).

111. Mezei, M. & Beveridge, D. L. Free Energy Simulationsa. *Ann. N. Y. Acad. Sci.* **482,** 1–23 (1986).

112. Aqvist, J., Medina, C. & Samuelsson, J. E. A new method for predicting binding affinity in computer-aided drug design. *Protein Eng.* **7,** 385–391 (1994).

113. Aqvist, J. & Marelius, J. The linear interaction energy method for predicting ligand binding free energies. *Comb. Chem. High Throughput Screen.* **4,** 613–626 (2001).

114. Knight, J. L. & Brooks, C. L. λ-Dynamics free energy simulation methods. *J. Comput. Chem.* **30,** 1692–1700 (2009).

115. Reddy, M. R. & Erion, M. D. *Free Energy Calculations in Rational Drug Design*. (Springer Science & Business Media, 2001).

116. Kollman, P. A. *et al.* Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc. Chem. Res.* **33,** 889–897 (2000).

117. McCammon, J. A. in *Computational Approaches in Supramolecular Chemistry* (ed. Wipff, G.) 515–517 (Springer Netherlands, 1994). at <http://link.springer.com/chapter/10.1007/978-94-011-1058-7_33>

118. Nicholls, A. & Honig, B. A Rapid Finite Difference Algorithm, Utilizing Successive Over-relaxation to Solve the Poisson-Boltzmann Equation. *J Comput Chem* **12,** 435–445 (1991).

119. Bashford, D. & Case, D. A. Generalized born models of macromolecular solvation effects. *Annu. Rev. Phys. Chem.* **51,** 129–152 (2000).

120. Still, W. C., Tempczyk, A., Hawley, R. C. & Hendrickson, T. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.* **112,** 6127–6129 (1990).

121. Huo, S., Massova, I. & Kollman, P. A. Computational alanine scanning of the 1:1 human growth hormone-receptor complex. *J. Comput. Chem.* **23,** 15–27 (2002).

122. Huo, S., Wang, J., Cieplak, P., Kollman, P. A. & Kuntz, I. D. Molecular dynamics and free energy analyses of cathepsin D-inhibitor interactions: insight into structure-based ligand design. *J. Med. Chem.* **45,** 1412–1419 (2002).

123. Sgobba, M., Caporuscio, F., Anighoro, A., Portioli, C. & Rastelli, G. Application of a post-docking procedure based on MM-PBSA and MM-GBSA on single and multiple protein conformations. *Eur. J. Med. Chem.* **58,** 431–440 (2012).

124. Brown, S. P. & Muchmore, S. W. Large-scale application of high-throughput molecular mechanics with Poisson-Boltzmann surface area for routine physics-based scoring of protein-ligand complexes. *J. Med. Chem.* **52,** 3159–3165 (2009).

125. Kuhn, B., Gerber, P., Schulz-Gasch, T. & Stahl, M. Validation and Use of the MM-PBSA Approach for Drug Discovery. *J. Med. Chem.* **48,** 4040–4048 (2005).

126. Gilbert, J., Henske, P. & Singh, A. Rebuilding big pharma's business model. *VIVO-N. Y. THEN NORWALK-* **21,** 73–80 (2003).

127. Segall, M. & beresford, A. P. *Virtual ADME-Tox: The Prom- ise of Technology in Preclinical Development, in: C. Sansom (Ed.), Enabling Technologies: Delivering the Future for Pharmaceutical R&D*. (PJP Publications Ltd., 2002).

128. Bode, C. The nasty surprise of a complex drug-drug interaction. *Drug Discov. Today* **15,** 391–395 (2010).

2

AIM OF THE PROJECT

The main aim of in this PhD project is based on the application of *in silico* approaches in order to provide answers to different pharmaceutical problems. Most of the techniques to be applied come from the Structure-Based field, thus the knowledge of the three-dimensional structure is the core of the investigation.

In several cases the answers to certain questions cannot be obtained by a straightforward workflow. Therefore, the computer scientists need to know a large variety of softwares, methodologies and even how to manage different type of files. These knowledge and skills are an important aspect of the daily research in CADD field, thus any implementation or strategy to overcome these difficulties are of utmost importance.

For this reasons we tried to developed different softwares to perform complex and exhaustive analysis of diverse aspects of the computational chemistry.

To retrieve several information is essential the application of an ample variety of arduous methodologies, and inevitably the users need to have a good proficiency in handling multiple types of files. Moreover, such analysis need to be performed for more than one single molecule, thus these simulations are iterated multiple times. Hence, a part from the necessary computational time, also the time used by user manual procedures has a relevant implication. Consequently, the balance between the results accuracy and the time required to achieve the outcomes are the main focus in our software development. This task will be addressed by creating new as pipelines that guide the user from the input submission to the generation of the results likes plots, graphical representations and raw data.

Indeed such pipelines are thinking to reduce the demanded time to perform the analysis and to lighten the proficiency required by the user to achieve the outcomes. Finally these new tools would be integrated in our common routine analyses, in order to prove their usefulness.

3

Ph.D. Thesis
Alberto Cuzzolin

2016

SCIENTIFIC PUBLICATIONS

During this Ph.D. they were carried out several projects to face different pharmaceutical aspects. Particularly, by taking advantage of structure-based techniques we provide answers in the field of protein-ligand recognition events. Although multiple tools are available nowadays, in several cases these methodologies were not completely suitable for the current study. Therefore, many efforts were made to develop novel methodologies to overcome the deficiency of existing computational software.

The current chapter contains seven scientific publications, which reflect the work done and the main results obtained during these years. The articles are organized into two sections: methodology development and computational techniques application.

The former contains the following four articles:

1. “Implementing the “Best Template Searching” tool into Adenosiland platform“
2. “Alternative Quality Assessment Strategy to Compare Performances of GPCR-Ligand Docking Protocols: The Human Adenosine A_{2A} Receptor as a Case Study”
3. “DockBench: An Integrated Informatic Platform Bridging the Gap between the Robust Validation of Docking Protocols and Virtual Screening Simulations”
4. “Deciphering the Complexity of Ligand-protein Recognition Pathways using Supervised Molecular Dynamics (SuMD) Simulations.”

The computational techniques section includes three publications, in two of which our in-home software called SuMD was applied:

5. “Exploring the recognition pathway at the human A_{2A} adenosine receptor of the endogenous agonist adenosine using supervised molecular dynamics simulations“
6. “Understanding allosteric interactions in G protein-coupled receptors using Supervised Molecular Dynamics: a prototype study analysing the human A_3 adenosine receptor positive allosteric modulator LUF6000”
7. “ALK Kinase Domain Mutations in Primary Anaplastic Large Cell Lymphoma: Consequences on NPM-ALK Activity and Sensitivity to Tyrosine Kinase Inhibitors”

3.1 Implementing the “Best Template Searching” tool into Adenosiland platform

Matteo Floris, Davide Sabbadin, Antonella Ciancetta, Ricardo Medda, Alberto Cuzzolin and Stefano Moro

Abstract

Background: Adenosine receptors (ARs) belong to the G protein-coupled receptors (GCPRs) family. The recent release of X-ray structures of the human A_{2A} AR (hA_{2A} AR) in complex with agonists and antagonists has increased the application of structure-based drug design approaches to this class of receptors. Among them, homology modeling represents the method of choice to gather structural information on the other receptor subtypes, namely A_1, A_{2B}, and A_3 ARs. With the aim of helping users in the selection of either a template to build its own models or ARs homology models publicly available on our platform, we implemented our web-resource dedicated to ARs, Adenosiland, with the “Best Template Searching” facility. This tool is freely accessible at the following web address: http://mms.dsfarm.unipd.it/Adenosiland/ligand.php.

Findings: The template suggestions and homology models provided by the “Best Template Searching” tool are guided by the similarity of a query structure (putative or known ARs ligand) with all ligands co-crystallized with hA_{2A} AR subtype. The tool computes several similarity indexes and sort the outcoming results according to the index selected by the user.

Conclusions: We have implemented our web-resource dedicated to ARs Adenosiland with the “Best Template Searching” facility, a tool to guide template and models selection for hARs modelling. The underlying idea of our new facility, that is the selection of a template (or models built upon a template) whose co-crystallized ligand shares the highest similarity with the query structure, can be easily extended to other GPCRs.

Keywords: G protein-coupled receptors; Adenosine receptors; Receptor modelling; Bioinformatics platform; Adenosiland

Findings

The template suggestions and homology models pro- vided by the “Best Template Searching”tool are guided by the similarity of a query structure (putative or known ARs ligand) with all ligands co-crystallized with hA_{2A} AR subtype. The tool computes several similarity indexes and sort the outcoming results according to the index

Published - Floris, M.; Sabbadin, D.; Ciancetta, A.; Medda, R.; Cuzzolin, A.; Moro, S. *Silico Pharmacol.* **2013**, *1* (1), 25.

selected by the user.

Background

Adenosine receptors (ARs) belong to the G protein- coupled receptors (GCPRs) family. The known four subtypes, termed adenosine A_1, A_{2A}, A_{2B} and A_3 receptors, are widely distributed in human body and involved in several physio-pathological processes[1]. The release of X-ray structures of the human A_{2A} AR in complex with agonists[2,3] and antagonists[4–8] has enabled to extend structure-based drug design approaches to this class of receptors. With the use of homology model- ling techniques, indeed, structural information on the other subtypes can also be derived. As a key step when building homology models is the selection of a proper template, we have developed a tool to guide the user in this crucial choice by implementing the "Best Template Searching" facility in our web-resource dedicated to ARs, Adenosiland[9]. This tool is freely accessible at the following web address*:* *http://mms.dsfarm.unipd.it/Adenosiland/ligand.php*. The underlying idea behind this facility is to help the user in selecting the best template or ARs model to get the highest quality receptor for further molecular docking studies. A possible strategy herein presented is to compute the similarity between a known or putative agonist/antagonist and all co-crystallized ARs ligands.

Table 1 - Values of the in-house validation of the combined similarity index

Input ligand	Suggested template	Combined similarity value
Adenosine	2YDO	0.83
NECA	2YDO	0.72
UK-432,097	3QAK	0.37
ZMA 241385	4EIY	0.69
T4G	3UZA	0.84
T4E	3UZC	0.92
XAC	3REY	0.67
Caffeine	3RFM	0.98

Tool description

The "Best Template Searching" tool works as follows: the user is asked to input a query molecule either by uploading a SMILES string or by directly drawing the 2D structure by using the JME interface; the similarity of the input molecule is then computed against all the ligands co-crystallized with the hA_{2A} AR. The following similarity indexes are calculated: *i)* shape similarity (based on the Manhattan distance between USR descrip- tors), *ii)* 2D similarity (based on the Tanimoto and Tversky Similarities of Pubchem Fingerprints), *iii)* phar- macophoric similarity (based on the Tanimoto similarity of Pharmacophoric triplets), and *iv)* a combined similarity (derived by the following function: 0.6 * pharmacophoric similarity + 0.4 * shape

Published - Floris, M.; Sabbadin, D.; Ciancetta, A.; Medda, R.; Cuzzolin, A.; Moro, S. *Silico Pharmacol.* **2013**, *1* (1), 25.

similarity).

The values of the two coefficients composing the latter similarity index have been derived by running a preliminary in-house validation based on all available crystallographic structures: In particular, the two values have been chosen so that by providing as input the structures of the co-crystallized ligand the corresponding receptor structure results the best ranked one according to the combined similarity index. The values obtained for the structures considered for the internal validation are reported in Table 1. For all the structure except one, the suggested template results the corresponding crystal structure. The only exception is represented by NECA for which the structure co-crystallized with adenosine is suggested as best template. Considering the high structural similarity between the two agonist structures, the results is in line with the others. Simultaneously to the best template searching process, a similarity search screening is also performed against all adenosine agonists and antagonists deposited in ChEMBL, release 14[10]. In more details, the query is compared to 760 A_1, 469 A_{2A}, 559 A_{2B} and 290 A_3 AR ligands and the comparison is based on the calculation of the similarity measures previously described. The identified compounds are reported in a table along with the associated binding data available in literature.

Tool validation

Ligand similarity biased template selection criteria at the basis of the “Best Template Searching” tool has been successfully applied to rationalize the Structure Activity Relationships (SAR) of a series of [5-substituted-4- phenyl-1,3-thiazol-2-yl] furamides as antagonist of the hARs[11]. The most potent derivative of the furamides series, the furan-2-carboxylic acid (4- phenyl-5-pyridin-4-yl-thiazol-2-yl)-amide, has been selected as query molecule: As reported in Table 2, a similarity sorting of the templates based on the combined similarity criteria has been taken into account to select the most suit- able models for receptor-based ligand design. The selected workflow is summarized in Figure 1: Starting from the suggested best template, namely the structure with the 3UZA PDB ID, co-crystallized with the 6-(2,6-dimethylpyr- idin-4-yl)-5-phenyl-1,2,4-triazin-3-amine (T4G), we have constructed A_1, A_{2B} and A_3 AR models through homology modeling and used the so derived structural information to provide hypotheses of ligand-receptor interaction and ligand-receptor selectivity profile[11].

Published - Floris, M.; Sabbadin, D.; Ciancetta, A.; Medda, R.; Cuzzolin, A.; Moro, S. *Silico Pharmacol.* **2013**, *1* (1), 25.

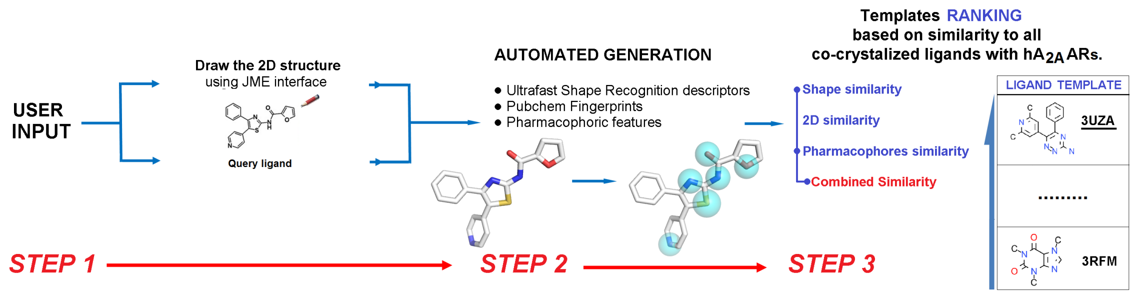

USER INPUT
Draw the 2D structure using JME interface
Query ligand
AUTOMATED GENERATION
● Ultrafast Shape Recognition descriptors
● Pubchem Fingerprints
● Pharmacophoric features
Templates RANKING based on similarity to all co-crystalized ligands with hA$_{2A}$ ARs.
Shape similarity
2D similarity
Pharmacophores similarity
Combined Similarity
LIGAND TEMPLATE
3UZA
....
3RFM
STEP 1
STEP 2
STEP 3

Figure 1 - Workflow of the homology modeling template selection based on the structure of furan-2-carboxylic acid (4-phenyl-5- pyridin-4-yl-thiazol-2-yl)-amide.

Table 2 -Similarity sorting of human A2A AR templates based on furan-2carboxylic acid (4-phenyl-5-pyridin-4-yl-thiazol-2-yl)-amide query ligand

Ligand	PDB ID Templ.	Shape simil.	[a]2D simil.	[b]2D simil.	[a]Pharmac. simil.	[b]Pharmac. simil.	Combined simil. (Shape & FP)
T4G	**3UZA**	**0.33**	**0.86**	**0.89**	**0.46**	**0.65**	**0.52**
ZM 241385	3PWH	0.58	0.90	0.93	0.27	0.42	0.48
T4E	3UZC	0.37	0.84	0.89	0.44	0.54	0.47
ZM 241385	4EIY	0.34	0.90	0.93	0.27	0.43	0.39
ZM 241385	3EML	0.35	0.90	0.93	0.27	0.42	0.39
NECA	2YDV	0.51	0.82	0.87	0.17	0.31	0.39
ZM 241385	3VG9	0.32	0.90	0.93	0.27	0.43	0.38
XAC	3REY	0.21	0.89	0.94	0.25	0.48	0.37
ZM 241385	3VGA	0.28	0.90	0.93	0.27	0.42	0.36
Adenosine	2YDO	0.33	0.82	0.86	0.18	0.31	0.31
Caffeine	3RFM	0.26	0.81	0.85	0.21	0.34	0.30
UK-432,097	3QAK	0.16	0.87	0.93	0.14	0.35	0.27

[a] **Tanimoto** [b] **Tversky**

Published - Floris, M.; Sabbadin, D.; Ciancetta, A.; Medda, R.; Cuzzolin, A.; Moro, S. *Silico Pharmacol.* **2013**, *1* (1), 25.

Methods

The “Best Template Searching” tool is part of the Adenosiland infrastructure, based on Ubuntu 9.10 Linux operating system, which is a patchwork of several informatics tools (for more details see Floris et al. 2013). The similarity indexes are calculated by using different approaches: 2D similarity based on Tanimoto and Tversky indexes[12,13] are calculated from Pubchem Fingerprints (CDK implementation), the shape similarity is calculated by using an in-house implementation of the Ultrafast Shape Recognition method[14,15], and the pharmacophoric features of the pharmacophore-based similarity index are described by Gaussian 3D volumes[16].

Conclusions

We have implemented a novel tool, called “Best Template Searching” to provide template suggestions and homology models of all four hARs based on the similarity between a query structure provided by the user and all co-crystallized ARs ligands. It is well known that ligand-driven induced fit of the receptor is a key feature to facilitate the identification or the optimization of novel potent and selective agonists and antagonists, in particular through molecular docking studies. We therefore believe that choosing as template the structure co-crystallized with the ligand that shares the highest structural similarity with the scaffold of interest may represent an effective strategy. This is in facts the under- lying idea of our platform implementation: By using the “Best Template Searching” option, users can upload a SMILES string or directly draw the 2D structure by using the JME interface of the scaffold of interest and search the most similar ligand co-crystallized so far with the hA_{2A} AR. Several similarity indexes are calculated by using different approaches such as a 2D similarity, shape similarity, pharmacophore-based similarity, and simple consensus shape- and pharmacophore-based similarity index. We are also confident that the proposed strategy can be easily and effectively extended to other GPCRs.

Abbreviations

ARs:	Adenosine receptors
GPCRs:	G protein-coupled receptors
NECA:	N-ethyl-5′-carboxamido adenosine;
T4E:	4-(3-amino-5-phenyl-1,2,4-triazin-6-yl)-2- chlorophenol;
T4G:	6-(2,6-dimethylpyridin-4-yl)-5-phenyl-1,2,4-triazin-3-amine;
ZM 241385:	4-(2-(7-amino-2-(2-furyl)(1,2,4)triazolo(2,3-a)(1,3,5)triazin-5-yl-amino) ethyl)phenol;
XAC:	N-(2-aminoethyl)-2-[4-(2,6-dioxo-1,3-dipropyl- 2,3,6,7- tetrahydro-1H-purin-8-yl)ph

Published - Floris, M.; Sabbadin, D.; Ciancetta, A.; Medda, R.; Cuzzolin, A.; Moro, S. *Silico Pharmacol.* **2013**, *1* (1), 25.

Bibliography

1. Fredholm, B. B., IJzerman, A. P., Jacobson, K. A., Klotz, K. N. & Linden, J. International Union of Pharmacology. XXV. Nomenclature and classification of adenosine receptors. *Pharmacol. Rev.* **53,** 527–552 (2001).

2. Lebon, G. *et al.* Agonist-bound adenosine A2A receptor structures reveal common features of GPCR activation. *Nature* **474,** 521–525 (2011).

3. Xu, F. *et al.* Structure of an agonist-bound human A2A adenosine receptor. *Science* **332,** 322–327 (2011).

4. Jaakola, V.-P. *et al.* The 2.6 angstrom crystal structure of a human A2A adenosine receptor bound to an antagonist. *Science* **322,** 1211–1217 (2008).

5. Doré, A. S. *et al.* Structure of the adenosine A(2A) receptor in complex with ZM241385 and the xanthines XAC and caffeine. *Struct. Lond. Engl. 1993* **19,** 1283–1293 (2011).

6. Hino, T. *et al.* G-protein-coupled receptor inactivation by an allosteric inverse-agonist antibody. *Nature* **482,** 237–240 (2012).

7. Congreve, M. *et al.* Discovery of 1,2,4-triazine derivatives as adenosine A(2A) antagonists using structure based drug design. *J. Med. Chem.* **55,** 1898–1903 (2012).

8. Liu, W. *et al.* Structural basis for allosteric regulation of GPCRs by sodium ions. *Science* **337,** 232–236 (2012).

9. Floris, M. *et al.* Implementing the ‘Best Template Searching’ tool into Adenosiland platform. *Silico Pharmacol.* **1,** 25 (2013).

10. Gaulton, A. *et al.* ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **40,** D1100–1107 (2012).

11. Inamdar, G. S. *et al.* New insight into adenosine receptors selectivity derived from a novel series of [5-substituted-4-phenyl-1,3-thiazol-2-yl] benzamides and furamides. *Eur. J. Med. Chem.* **63,** 924–934 (2013).

12. Steinbeck, C. *et al.* The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics. *J. Chem. Inf. Model.* **43,** 493–500 (2003).

13. Steinbeck, C. *et al.* Recent developments of the chemistry development kit (CDK) - an open-source java library for chemo- and bioinformatics. *Curr. Pharm. Des.* **12,** 2111–2120 (2006).

Published - Floris, M.; Sabbadin, D.; Ciancetta, A.; Medda, R.; Cuzzolin, A.; Moro, S. *Silico Pharmacol.* **2013**, *1* (1), 25.

14. Floris, M., Masciocchi, J., Fanton, M. & Moro, S. Swimming into peptidomimetic chemical space using pepMMsMIMIC. *Nucleic Acids Res.* **39,** W261–W269 (2011).

15. Pedro J Ballester, W. G. R. Ultrafast Shape Recognition to Search Compound Databases for Similar Molecular Shapes. *J. Comput. Chem.* **28,** 1711–23 (2007).

16. Taminau, J., Thijs, G. & De Winter, H. Pharao: pharmacophore alignment and optimization. *J. Mol. Graph. Model.* **27,** 161–169 (2008).

Published - Floris, M.; Sabbadin, D.; Ciancetta, A.; Medda, R.; Cuzzolin, A.; Moro, S. *Silico Pharmacol.* **2013**, *1* (1), 25.

3.2 Alternative Quality Assessment Strategy to Compare Performances of GPCR-Ligand Docking Protocols: The Human Adenosine A_{2A} Receptor as a Case Study

Antonella Ciancetta, Alberto Cuzzolin, and Stefano Moro*

Abstract

The progress made in the field of G protein-coupled receptors (GPCRs) structural determination has increased the adoption of docking-driven approaches for the identification or optimization of novel potent and selective ligands. In this work, we compared the performances of the 16 different docking/scoring combinations using the recently released crystal structures of the human A_{2A} AR (hA_{2A} AR) in complex with both agonists and antagonists. The proposed evaluation strategy encompasses the use of three complementary “quality descriptors”: *a)* the number of conformations generated by a docking algorithm having a RMSD value lower than the crystal structure resolution (R); *b)* a novel consensus-based function defined as “protocol score”; and *c)* the interaction energy maps (IEMs) analysis, based on the identification of key ligand−receptor interactions observed in the crystal structures.

Introduction

The progress made in the field of G protein-coupled receptors (GPCR) structural determination has increased the adoption of docking-driven approaches for the identification or the optimization of novel potent and selective ligands[1-8]. As routinely demonstrated, docking programs are usually successful in generating multiple poses that include binding modes similar to the crystallographically determined bound structure, whereas scoring functions are much less successful at correctly identify the corresponding “bioactive” binding mode[9]. This intrinsic limitation generally implies the need for the calibration of the docking protocol through benchmark studies prior to applying it. Traditionally, these benchmarks have focused on redocking the cognate ligand of a crystallographic receptor−ligand complex to measure geometric pose prediction accuracy[10-12].

In this work, we propose an alternative quality assessment strategy to compare the performances of the 16 different docking/scoring combinations using the recently released crystal structures of the human A_{2A} AR (hA_{2A} AR) in complex with both agonists and antagonists, as summarized in Table 1[13-19]. Among them, one is cocrystallized with the endogenous agonist adenosine (PDB ID: 2YDO[15]), one with its synthetic analogue NECA (N-ethyl-5′-carboxamido adenosine, PDB ID, 2YDV[15]), and the remaining eight with five antagonists, namely, ZM 241385 (4- (2- (7-amino-2- (2-furyl) (1,2,4) triazolo-(2,3-a) (1,3,5) triazin-5-yl-amino)ethyl) phenol, PDB IDs

Published - Ciancetta, A.; Cuzzolin, A.; Moro, S. *J. Chem. Inf. Model.* **2014**, *54* (8), 2243–2254.

3EML[13], 3PWH[16], 3VGA[17], 4EIY[19]); T4G (6- (2,6-dimethyl- pyridin-4-yl) -5-phenyl-1,2,4-triazin-3-amine, PDB ID 3UZA[18]); T4E (4-(3-amino-5-phenyl-1,2,4-triazin-6-yl)-2-chlorophenol, PDB ID 3UZC[18]); caffeine (PDB ID 3RFM[27]); and XAC (N-(2-aminoethyl)-2-[4-(2,6-dioxo-1,3-dipropyl-2,3,6,7-tetrahy-dro-1H-purin-8-yl)phenoxy] acetamide, PDB ID 3REY[16]). The structures of the cocrystallized ligands are shown in Figure 1. The pharmacology of adenosine receptors and the potential applications of their agonists and antagonists have already been extensively described and recently reviewed[20]. The performances of all docking/scoring combinations were evaluated using an alternative assessment strategy defined by three complementary “quality descriptors”: *a)* the number of conformations generated by the docking algorithm having a root mean square deviation (RMSD) value lower than the crystal structure resolution (R); *b)* a novel consensus-based function defined as “protocol score”; and *c)* the interaction energy maps (IEMs) analysis, based on the identification of key ligand−receptor interactions observed in the crystal structures.

Table 1 - hA_{2A} AR Crystall Structures Available to Date

PDB ID	**Release date**	**R (Å)**	**Ligand name**	**Ligand name abbr.**[b]	**Crystal. Strategy**	**Ligand type**
3EML	08/10/14	2.60	ZM 241385	ZMA	T4 lysozime fusion[c]	Antagonist
2QAK[A]	11/03/09	2-71	UK-432,097	UKA	T4 lysozime fusion[c]	Agonist
2YDO	11/05/18	3.00	Adenosine	ADO	StaR[d]	Agonist
2YDY	11/05/18	2.60	NECA	NEC	StaR[d]	Agonist
3PWH	11/09/07	3.30	ZM 241385	ZMA	StaR[e]	Antagonist
3REY	11/09/07	3.31	XAC	XAC	StaR[e]	Antagonist
3RFM	11/09/07	3.60	Caffeine	CFF	StaR[e]	Antagonist
3VG9[A]	12/02/01	2.70	ZM 241385	ZMA	Fab2838 complex[f]	Antagonist
3VGA	12/02/01	3.10	ZM 243815	ZMA	Fab2838 complex[f]	Antagonist
3UZA	12/03/21	3.27	T4G	T4G	StaR[e]	Antagonist
3UZC	12/03/21	3.34	T4E	T4E	StaR[e]	Antagonist
4IEY	12/07/25	1.80	ZM 241385	ZMA	ApoCytochrom e b562RIL chimera[g]	Antagonist

[a] Structures not considered in this study. [b] Three letter code assigned in the PDB file. [c] A_{2A}AR-T4L-ΔC: IL3 replaced by T4 and C-Term deleted. [d] A_{2A} AR-GL31- ΔC: thermostabilizing mutations (L48A, A54L, T65A and Q89A), N154A mutation and C-Term deleted. [e] A_{2A} AR-StaR2- ΔC: thermostabilizing mutations (A54L, T88A, K122A, V239A, R107A, L202A, L235A and S277A), N154 mutation and C-Term deleted. [f] A_{2A} AR-Fab2838- ΔC complex with mouse monoclonal-antibody Fab fragment (Fab2838), N154 mutation and C-Term deleted. [g] A_{2A}AR-BRIL- ΔC: IL3 replaced by apocytochrome b562RIL and C-Term, deleted.

The comparison of IEMs enables a fast and graphical selection of the conformations based on the quality of the interactions (in terms of the number of

Published - Ciancetta, A.; Cuzzolin, A.; Moro, S. *J. Chem. Inf. Model.* **2014**, *54* (8), 2243–2254.

established interactions and their relative strength) occurring between each docking pose and selected key residues. This type of analysis was proposed in the past by several authors and in different forms but rarely applied due to the lack of automation and the subjectivity of interactions selection[21-24]. Both limitations have been over- come by developing an in house python script.

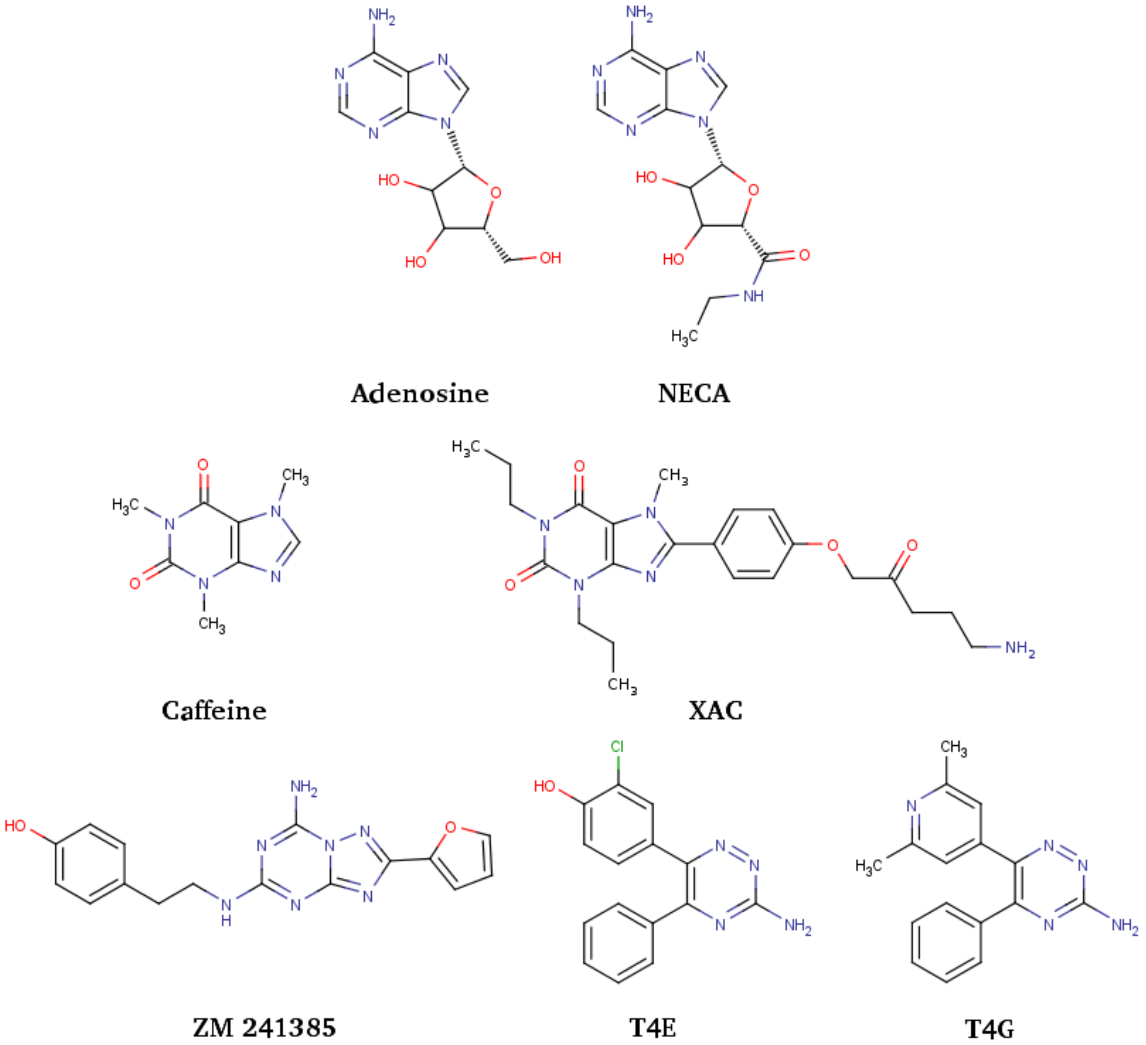


Figure 1 – Structures of hA_{2A} AR cocrystalized ligands used to perform the docking benchmark study.

Materials and Methods

Numbering and Name Conventions. The residues of the hA_{2A} AR are indicated according to the following scheme: the three letter residue name is followed by the residue number and the Ballesteros and Weinstein notation reported in brackets[25]. The latter nomenclature scheme, generally indicated by "TM ± 50", identifies the residues by the helix number (TM) followed by the position relative (±) to a reference residue among the most conserved amino acids in that helix, to which the number 50 is arbitrarily assigned. Compound names correspond to the ligands three letter codes assigned in the PDB file, whereas protein identifiers are the corresponding PDB entries. Docking protocols are named according to the following scheme: "program name abbreviation-scoring function/search algorithm".

Computational Facilities. Energy calculations and analyses of docking poses were performed with the Molecular Operating Environment (MOE, version 2012.10) suite[26]. Extraction and analysis of docking results were performed by using in-house bash and python scripts. Maps and graphs were created with Gnuplot, version 4.4[27].

Protein Structures Selection and Preparation. Of the 12 hA_{2A} AR crystal structures available, listed in Table 1, the following structures were used to perform the benchmark study (PDB IDs): 2YDO[15], 2YDV[15], 3EML[13], 3PWH[16], 3REY[16], 3RFM[16], 3UZA[18], 3UZC[18], 3VGA[17], and 4EIY[19]. The structures identified by the 3QAK[14] and 3VGA[17] PDB IDs were not considered: We excluded from our analysis the ligand cocrystallized in the 3QAK structure because it has a high number of heavy atoms and a number of rotatable bonds exceeding the cutoff allowed for some of the in-house available docking programs; the inverse agonist cocrystallized in the 3VGA structure was not considered because it has low occupancy in the PDB structure. The selected structures were retrieved from the RCSB PDB database (http://www.rcsb.org)[28]. Before the preparation procedure, all the proteins were aligned and superimposed to a selected reference structure (3EML[13]). Antibody portions, ions, and crystallization solvents were removed, whereas water molecules and cocrystallized ligands were retained for the hydrogen atoms assignment step and then removed. Fused proteins (lysozime and apocytochrome in the specific cases) as well as point mutations were retained, and the structures were not subjected to any conformational changes. Missing loop domains, N-terminal and C-terminal, were not modeled. Ionization states and hydrogen positions were assigned with the "Protonate-3D" tool[29], as implemented in the MOE suite. Then, to minimize contacts among hydrogen atoms, the structures were subjected to energy minimization with Amber99 force field[30] until the root mean square (RMS) of the conjugate gradient was <0.05 $kcal\cdot mol^{-1}$ $Å^{-1}$, by keeping the heavy atoms fixed at their positions. After the protonation step, for each structure, the coordinates of the binding site center (barycenter of cocrystallized ligand) were determined and saved, then ligand and water molecules were removed and protein atoms partial charges computed with the Amber99 force field[30].

Ligand Structures. Co-crystallized ligands were extracted from the corresponding crystallographic complex and checked for errors. Hydrogen atoms were added, and the protonation state (pH 7.4) was assigned. The structures were not subjected to energy minimization, so that for each ligand the starting conformation is the same one observed in the crystal structure. Partial charges on ligands atoms were computed on the basis of the PM3/ESP semiempirical Hamiltonian[31,32].

Docking Settings. The performances of the following six docking programs were assessed: Autodock[33], GOLD[34], Glide[35], PLANTS[36], Molegro Virtual Docker[37], and MOE- dock[36]. The versions of the programs were the most up to date available at the time we performed the calculations. Among all the scoring functions and search algorithms available in the considered programs, we discarded those that did not allow us to return a user-defined number of output conformations without duplicates and postdocking refinement. In the end, a total amount of 16 different docking algorithm/scoring function combinations were assessed, as detailed in Table 2. To make the results obtained with different protocols as homogeneous as possible, we set the common settings reported in Table 3.

Docking Stages. Each ligand structure was first docked into the corresponding crystal structure with the different docking protocols (cognate ligand docking). Then, for the protocols giving the best performances in the cognate ligand docking stage, we performed ensemble docking runs to assess whether the protocol is able to assign to each ligand (or ligand conformation in case of ZMA) the corresponding cocrystallized structure by selecting it among all the tested proteins. For the ensemble docking step, three different strategies for the definition of the binding site center have been evaluated. In particular, the binding site center was set as *i)* the centroid of the barycenters of the ligands; *ii)* the ZMA barycenter in 3EML structure; and *iii)* the barycenter of each ligand in its corresponding crystal structure. Finally, we evaluated the effect of two additional parameters on the docking outcomes: the reconstruction of the second extracellular loop (EL2) and the starting conformation.

Analysis of Docking Results. To judge the performances of the different tested protocols, the RMSD values between predicted and crystallographic poses were calculated. In the case of the XAC ligand, which has a highly flexible solvent-exposed tail, the RMSD values were computed only for the heavy atoms of the aromatic cores. The performances of the docking protocols were evaluated on the basis of the lowest, highest, and average RMSD values ($\mathrm{RMSD_{min}}$, $\mathrm{RMSD_{max}}$, and $\mathrm{RMSD_{ave}}$, respectively) as well as the highest number of conformations with a RMSD value lower than the corresponding X-ray resolution (R), $\mathrm{N^{(RMSD<R)}}$.

Table 2 - List of docking programs along with Search Algorithm (or Placing Method) and Scoring Function used to perform the docking benchmark study

Programs	Search Algorithm (+ placing method)	Scoring Function	Abbr.
Autodock 4.2	Genetic algorithm	AutoDock SF	AD-GA
	Lamarkian GA	AutoDock SF	AD-LGA
	Local search	AutoDock SF	AD-LS[a]
Glide 5.8	Glide algorithm	Standard Precision	Glide-SP
GOLD 5.1	Generic algorithm	Goldscore	Gold-Gold
	Generic algorithm	Chemscore	Gold-Chem
	Generic algorithm	ASP	Gold-ASP
	Generic algorithm	PLP	Gold-PLP
PLANTS 1.2	ACO algorithm	ChemPLP	Plants-PLP
	ACO algorithm	PLP95	Plants-PLP95
	ACO algorithm	PLP	Plants-PLP
MOE 2012.10	Systematic search + alpha triangle	London dG	MOE-AT
	Systematic search + alpha PMI	London dG	MOE-APMI
	Systematic search + triangle matcher	London dG	MOE-TM
Molegro Virtual Docker 5.5	Iterated simplex	MolDock SF	MVD-IS
	MolDock optimizer	MolDock SF	MVD-MDO
	MolDock simplex	MolDock SF	MVD-MDSE

Protocol Score. To compare at a glance the performances of the different protocols tested, we merged the above- discussed parameters and in particular the $\mathrm{RMSD}_{\mathrm{ave}}$ and the $N^{(\mathrm{RMSD}<R)}$ in a unique statistical value, that we called protocol score, defined

Published - Ciancetta, A.; Cuzzolin, A.; Moro, S. *J. Chem. Inf. Model.* **2014**, *54* (8), 2243–2254.

as follows: *a)* one point is assigned to each protocol that has the $RMSD_{ave}$ value lower than R; *b)* one point is assigned to each protocol that generates at least 10 conformations having RMSD values with respect to the X-ray binding mode lower than R; and *c)* two points are assigned to the protocols that satisfy both of the above-mentioned requirements. Moreover, to discern the best protocols among the good ones, three points are assigned to the protocols that give the lowest $RMSD_{ave}$ value and, at the same time, returns the highest number of conformers with a RMSD value lower than R. The protocol score assignment criteria are summarized in Table 4.

Table 3 - Common docking settings for the evaluated protocols

Parameter	Value/Setting
Ligand input conformation	X-ray binding mode
Ligand initial partial charges	PM3/ESP
Water molecules	Excluded
Output	20 conformations
RMSD threshold	1.0 Å
Binding cavity center	Ligand barycenter in X-ray structure
Binding cavity radius	20 Å
Grid spacing (for grid-based calculations)	0.3 Å
Refinement and rescoring	Turned off

Table 4 - Protocol Score Assignment Criteria

Condition	Score
$RMSD_{ave}<R$	1
$N^{(RMSD<R)}>10$	1
$RMSD_{ave}<R$ and $N^{(RMSD>R)}>10$	2
Protocol returns: Min($RMSD_{ave}$) and Max($N^{(RMSD>R)}$)	3

Analysis of Ligand−Receptor Interactions. To analyze the ligand−receptor interactions, we calculated the individual electrostatic and hydrophobic contributions to the interaction energy (hereby denoted as IEele and IEhyd, respectively) of key residues involved in the binding with the ligands, as emerged from detailed analyses and comparisons among the different crystallographic binding modes. In particular, the electrostatic contribution is computed on the basis of the nonbonded electrostatic interaction energy term of the force field[38], whereas the hydrophobic contributions is calculated by using the directional hydrophobic interaction term based on contact surfaces as implemented in the MOE scoring function[26]. As a consequence, energy (expressed in kcal/mol) is associated with the electrostatic contribution, whereas a score (the higher the better) is related to the hydrophobic contribution. The analysis of these contributions have been reported as "interaction energy maps" (hereby indicated as IEMs), graphically displayed as heat-like maps reporting the key residues involved in the binding with the considered ligands along with a quantitative estimate of the occurring interactions.

Results and Discussion

Overview of X-ray Binding Modes. Prior to discussing the results of our docking benchmark, we briefly report an overview of the binding modes observed in the ligand-hA_{2A} AR complexes under study. It has to be pointed out that the analysis herein reported lacks water mediated interactions, as we intentionally did not include water molecules in our docking simulations. We instead briefly addressed this topic in a recent study[39], and a deeper investigation of the role of water molecules in hA_{2A} AR ligand binding is the focus of a study being currently conducted in our research group[40].

Figure 2 depicts the IEM (for more details, see the Materials and Methods section) of the ligands under study. The map has been derived stepwise (see Figure S1, Supporting Information). We first computed for all the ligands the individual contribution of each residue to the interaction energy (per residue analysis). From a comparison of the results of the per residue analyses, we then identified the key residues involved in the binding with all ligands and reported the occurring interaction in the IEM in Figure 2. The common interaction pattern for all ligands involves an aromatic π−π stacking with the conserved Phe168, located in the second extracellular loop (EL2), and additional hydrophobic contacts with Leu249 (6.51) and Ile274 (7.39) side chains. Strong polar interactions are established with the side chain of the conserved Asn253 (6.55)[41]. The IEM along with the three-dimensional representation of the corresponding binding modes (Figure S2, Supporting Information) helps in appreciating the different extent (and type) of interaction networks and the different sizes of the ligands by allowing a direct comparison of all ligands to each other, between pairs of structurally related compounds and different conformations of the same molecule. With respect to the antagonists structures, agonists (NECA and adenosine) interact through fewer hydrophobic interactions while establishing hydrogen bonds with two additional residues, namely, Thr88 (3.36) and Ser277 (7.42), mediated by the ribose moiety. With respect to the latter interaction, it has to be pointed out that Thr88 (3.36) and Ser277 (7.42) are mutated to alanine in the so-called “StaR2” structures (3PWH, 3REY, 3RFM, 3UZC, and 3UZA), and therefore for those constructs, the interaction with these residues cannot be detected. Among the antagonist structures, the differences in the binding patterns between ZMA and caffeine reflect their binding affinities to the hA_{2A} AR: ZMA shows an extended pattern with strong polar and hydrophobic interactions, whereas caffeine establishes fewer and less intense interactions.

All of the above-described interactions are consistent with the available mutagenesis data: in particular, the far back known role of Asn253 (6.55) and the more recent mutagenesis data highlighting the roles of Phe168 (EL2) and Leu249 (6.51) for both agonist and antagonist binding and that of Thr88 (3.36) for agonist binding[42].

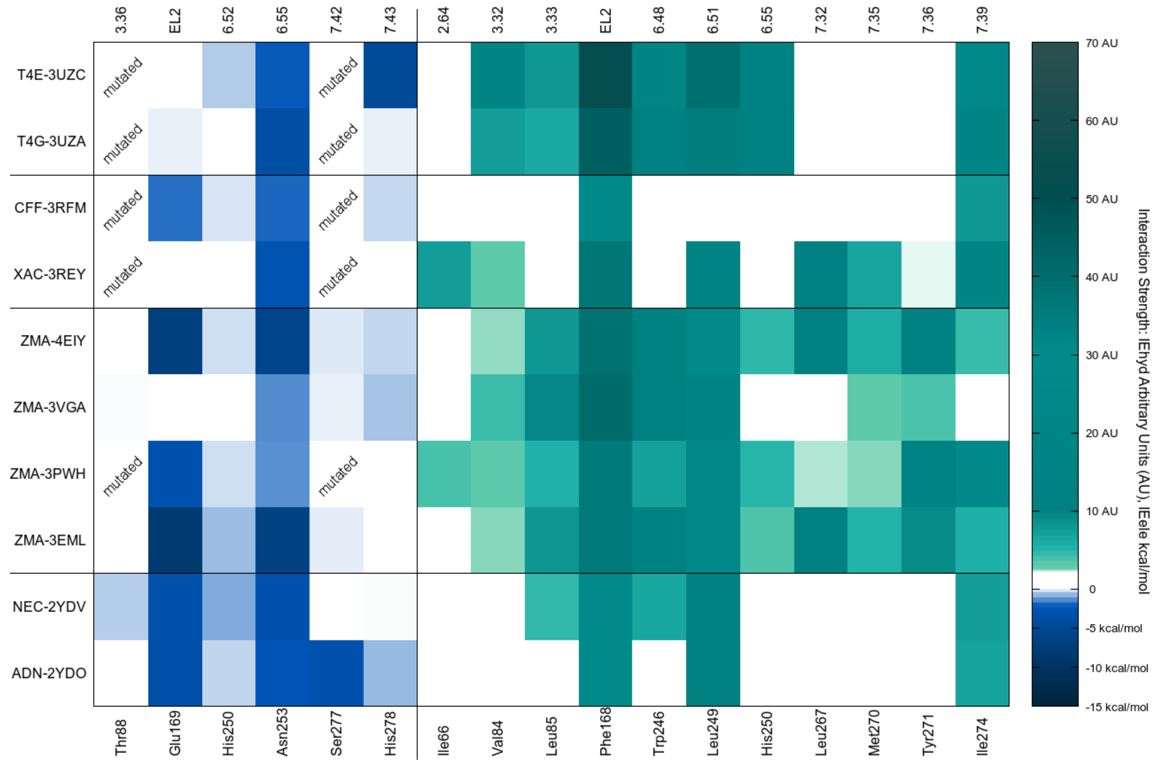


Figure 2 – Interaction energy map (IEM) for the hA_{2A} AR cocrystallized ligands under study. Ligands are identified by the three letter codes assigned in the PDB file followed by the PDB IDs.

Workflow. The workflow of the computational protocol is shown in Figure 3. Starting from the protein−ligand complexes (PDB files, see Table 1), the protein and ligand structures were prepared and docking simulations run with the selected protocols, listed in Table 2. Then, for each protein structure, the RMSD values of the conformations generated by the different docking protocols with respect to the ligand cocrystallized binding modes were calculated and statistical analyses performed. As detailed in the Materials and Methods section, during the protein preparation step, water molecules and cocrystallized ligands were retained. Once the hydrogen atoms were added and minimized, water molecules were removed and no longer considered. Complexed antibody portions, ions, and crystallization solvents were removed, whereas fused proteins as well as point mutations were retained. Unsolved protein regions (loops, N-terminal, and C- terminal) were not remodeled, and the structures were not energy minimized. Ligand structures were extracted from the original PDB files and checked for errors, and hydrogen atoms were added. The structures were not energy minimized in order to retain for each ligand the X-ray observed conformation. As it is generally accepted that the performances of docking protocols, especially those relying upon genetic algorithms, can be affected by the starting conformation, we also run test calculations on selected cases by supplying different random generated conformations as input and evaluated their effects on the performances of the docking protocols. The results are discussed in the following text.

After the protein and ligand preparation steps, we run the different docking protocols listed in Table 2: the assessed scoring functions include knowledge-based (Gold-ASP), force- field-based (AutodockSF and GoldScore), and empirical scoring functions, whereas tested algorithms comprise both deterministic (MOE and Glide) and stochastic search method approaches. Moreover, the variety of tested protocols also encompasses different types of protein representation, such as grid (Autodock and Glide) and all atom. We defined common settings (summarized in Table 3) for the different programs in order to ascribe the differences in the performances to the selected combination of search algorithms (or placing method) and scoring functions. In particular, we chose the same settings for the binding cavity (center, radius, and grid spacing for grid- based calculations), the ligand input (conformation and partial charges), and the program output (number of saved conformations, RMSD threshold value, refinement, and rescoring).

The results of the docking calculation were collected, and RMSD values with respect the cocrystallized ligand were computed for all conformations generated by the protocols.

The performances were evaluated on the basis of lowest, highest, and mean RMSD values as well as the highest number of conformations with a RMSD value lower than the corresponding X-ray resolution, $N^{(RMSD<R)}$. As the considered crystal structures range from high (4EIY, 1.8 Å) to low (3RFM, 3.6 Å) resolutions, to evaluate the latter statistical parameter ($N^{(RMSD<R)}$) we decided to compare each structure with its own R value rather than setting a fixed threshold.

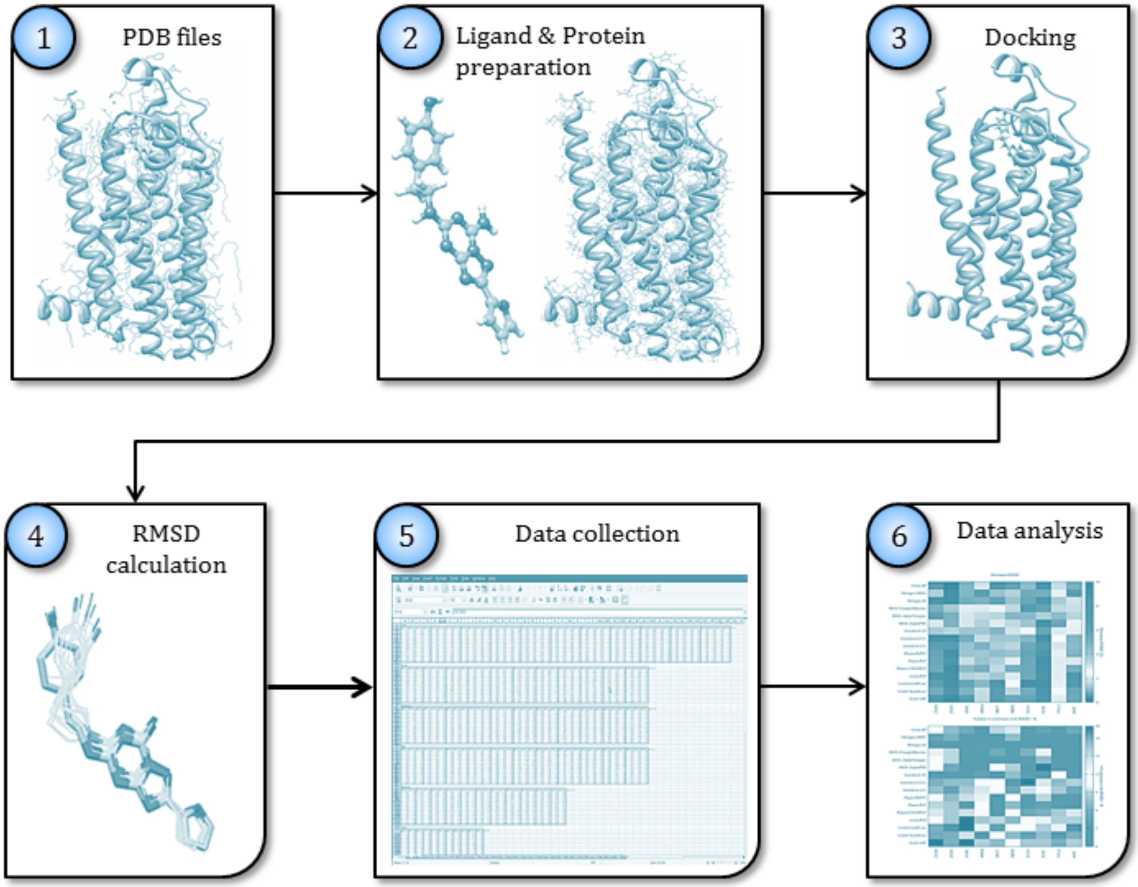


Figure 3 – Workflow of the benchmark study.

With this procedure, we intrinsically take into account the quality of the experimental data to be reproduced: the less confidence we have in the atomic coordinates experimentally determined, the less strictly we judge the performances of a protocol in reproducing them.

Cognate Ligand Docking. The results of the cognate ligand docking step are reported in Table S1 (Supporting Information), and the most relevant statistical parameters are graphically summarized in Figure 4. In the map in Figure 4A, the minimum RMSD value ($RMSD_{min}$) of all the tested docking protocols are reported for each considered crystal structure: At first glance, it can be noted that there are protocols, such as Gold-ASP; Gold-Gold; Gold-PLP; and Plants-PLP, able to generate at least one pose that reproduces the X-ray binding mode with satisfying accuracy regardless of the specific structure under consideration. From the map, it is also straightforward that agonist binding poses are predicted better than antagonist ones and in particular that the caffeine binding mode is the most challenging to be reproduced. As we also highlighted in our recent membrane molecular dynamics simulations of caffeine docking poses[39], the dynamical evolution of different starting binding modes involves a consistent number of water molecules in rapid exchange around the ligand structure as a consequence of the weak interactions established with the receptor. This feature makes it challenging to reproduce the crystallographic binding mode without taking into account the dynamical behavior of surrounding water molecules. The difficulty is even increased if water molecules are not considered at all, as in this specific docking exercise.

The ability to reproduce at least once the X-ray observed binding mode is not a sufficient criterion to judge the quality of the docking protocols. Indeed, when extending the analysis by computing the average RMSD value ($\mathrm{RMSD}_{\mathrm{ave}}$, Figure 4B), some of the above-mentioned protocols worsen their performances. Moreover, other protocols that accurately reproduced at least in one conformer the X-ray binding mode (Gold-Chem, MVD-IS) show $\mathrm{RMSD}_{\mathrm{ave}}$ values well over the structure resolutions. These data are consistent with the well-known limit of the docking procedure. Although search algorithms are usually successful at generating poses that include binding modes crystallographically observed, the scoring functions developed to date are much less successful at identifying and ranking the correct binding pose[9]. The best protocol would be the one that is able to accurately reproduce the X-ray binding mode and rank the conformations with the lowest RMSD values at the top of the list.

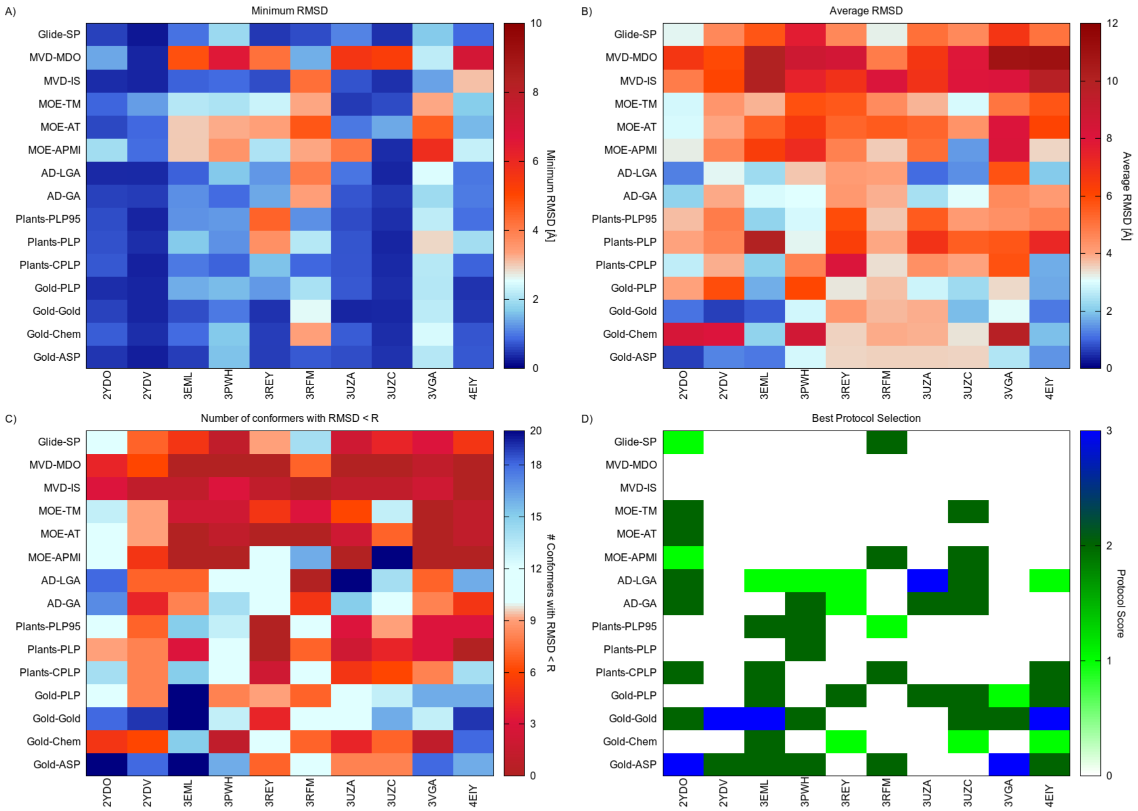


Figure 4 – Results of the cognate ligand docking procedure.

Published - Ciancetta, A.; Cuzzolin, A.; Moro, S. *J. Chem. Inf. Model.* **2014**, *54* (8), 2243–2254.

This is also in view of the application of the docking protocol in a virtual screening framework, where only a selected percentage of the top generated conformations is considered and further processed. Whit this in mind, we defined an additional parameter by computing for each protocol the number of conformers having a RMSD value lower than the X-ray structure resolution ($N^{(RMSD<R)}$), which represents somehow the best accuracy one might expect from a computational procedure based on experimental data. From the map in Figure 4C, it can be evinced that there are only a few protocols able to generate a high number of conformations close to the crystallographic binding mode and that only in a few cases (Gold-ASP/2YDO; Gold-ASP/3EML; Gold-Gold/3EML: Gold-PLP/3EML; AD- LGA/3UZA; MOE-APMI/3UZC) all of the conformations generated by the protocol have RMSD values below the structure resolution.

We therefore tried to merge all of the above-discussed parameters in a unique statistical value that we called protocol score, defined as detailed in the Materials and Methods section and summarized in Table 4. The map in Figure 4D graphically displays the scores assigned to all of the tested protocols. The obtained results suggest that it is not possible to identify the best protocol for low resolution structures (R > 3.00) and that the protocols that perform better for most structures are represented by the docking program Gold combined with the GoldScore and ASP scoring functions.

Consensus Scoring. A well-known technique to improve the performances of the scoring functions is to combine the results of different functions into a consensus score[43]. Moreover, it has been demonstrated that combining different types of scoring functions increases the accuracy, as each scoring function compensates for the weaknesses of the other one. We therefore combined the results of the two scoring functions performing at best for the majority of the considered structures (ASP and GoldScore) and evaluate the performances of the thus obtained consensus scoring function (Table 5). As can be noted, the consensus score improves the overall prediction accuracy and gives good performances for all the structures (protocol score = 2). Moreover, for several protein structures, such as 2YDV, 3EML, 3PWH, 3UZC, and 4EIY, the consensus score represents the best performing protocol (protocol score = 3).

Table 5 - Consensus Score Results[a]

PDB ID	Consensus protocol			Best single prtocol		
	RMSD$_{ave}$ [Å]	N$_{(RMSD<R)}$	Score	RMSD$_{ave}$ [Å]	N$_{(RMSD<R)}$	Score
2YDO	0.694	20	2	**0.593**	**20**	**3**
2YDY	**0.415**	**20**	**3**	0.600	19	3
3EML	**0.783**	**20**	**3**	0.853	20	3
3PWH	**2.661**	**17**	**3**	2.861	16	2
3REY	2.719	11	2	3.266	9	2
3RFM	3.661	12	2	3.163	14	2
3UZA	2.969	14	2	**1.033**	**20**	**3**
3UZC	1.081	19	2	1.489	20	2
3VGA	2.766	16	2	**2.470**	**18**	**3**
4EIY	**1.127**	**19**	**3**	1.156	19	3

[a] For each protein structure, the best protocols among all the tested ones in this study are reported in bold face text.

Ensemble Docking. For the above-described consensus protocol, we run ensemble docking calculations to assess whether the protocol is able to assign to each ligand (or ligand conformation in the case of ZMA) the corresponding cocrystallized structure by selecting it among all the considered proteins. For ensemble docking, we evaluated three different strategies for the definition of the binding site center by setting it as *i)* the centroid of the barycenters of the ligands; *ii)* the ZMA barycenter in the 3EML structure (upon which all of the other structures were aligned at the beginning of the docking procedure); and *iii)* the barycenter of each ligand in its corresponding crystal structure. The results collected in Table 6 report the percentage of protein selection for the generation of 20 conformers for each ligand. As can be noted, the definition of the binding site center does not significantly affect the outcomes. In all of the considered cases, the protocol preferentially selects three proteins, namely, the 4EIY, 2YDO, and 2YDV structures. These structures are characterized by high to low resolution (1.8, 2.6, and 3.0 Å, respectively) and the complete resolution of the second extracellular loop (EL2). We therefore ascribe the preference of the docking protocol for these structures to their completeness rather than to their resolution.

Published - Ciancetta, A.; Cuzzolin, A.; Moro, S. *J. Chem. Inf. Model.* **2014**, *54* (8), 2243–2254.

Table 6 - Percentage of protein selection of the ensemble docking runs

PDB ID	Strategy 1	Strategy 2	Strategy 3
2YDV	28.00	26.00	27.50
2YDO	24.50	25.50	22.50
3EML	1.00	2.00	1.50
3PWH	2.00	1.50	1.00
3REY	0.00	0.00	0.00
3RFM	1.50	1.50	1.00
3UZA	1.00	0.00	0.50
3UZC	1.00	0.00	0.00
3VGA	0.50	0.00	0.00
4EIY	40.50	43.50	46.00

Evaluation of Additional Parameters. IEMs Inspection. As mentioned in the Introduction section, we employed a complementary metrics to evaluate the protocol performance, that we called IEMs. In our implementation, the IEMs are based on the analysis of key ligand-binding interactions observed in the crystal structures (Figure 2) and enable a fast and graphical selection of the conformations generated by the docking algorithms. The selection is guided by the quality, in terms of the number of established interactions and their relative strength, of the interactions occurring between each docking pose and selected key residues. The identification of key residues can be either knowledge-based, such as a comparison with available protein−ligand crystal structures (in our case) or mutagenesis data, or blind. In the latter case, the IEMs are computed for all residues surrounding the binding site within a user-defined radius and are devoid of any subjectivity. Moreover, to solve the automation issue that has been previously pointed out by several authors[21-24], we developed a python script that automatically generates the IEMs from the computed interactions.

Figure 6A depicts the IEMs of the output conformations generated by the consensus protocol for the 3UZA structure. A first glance at the map highlights the presence of two clusters of conformations.

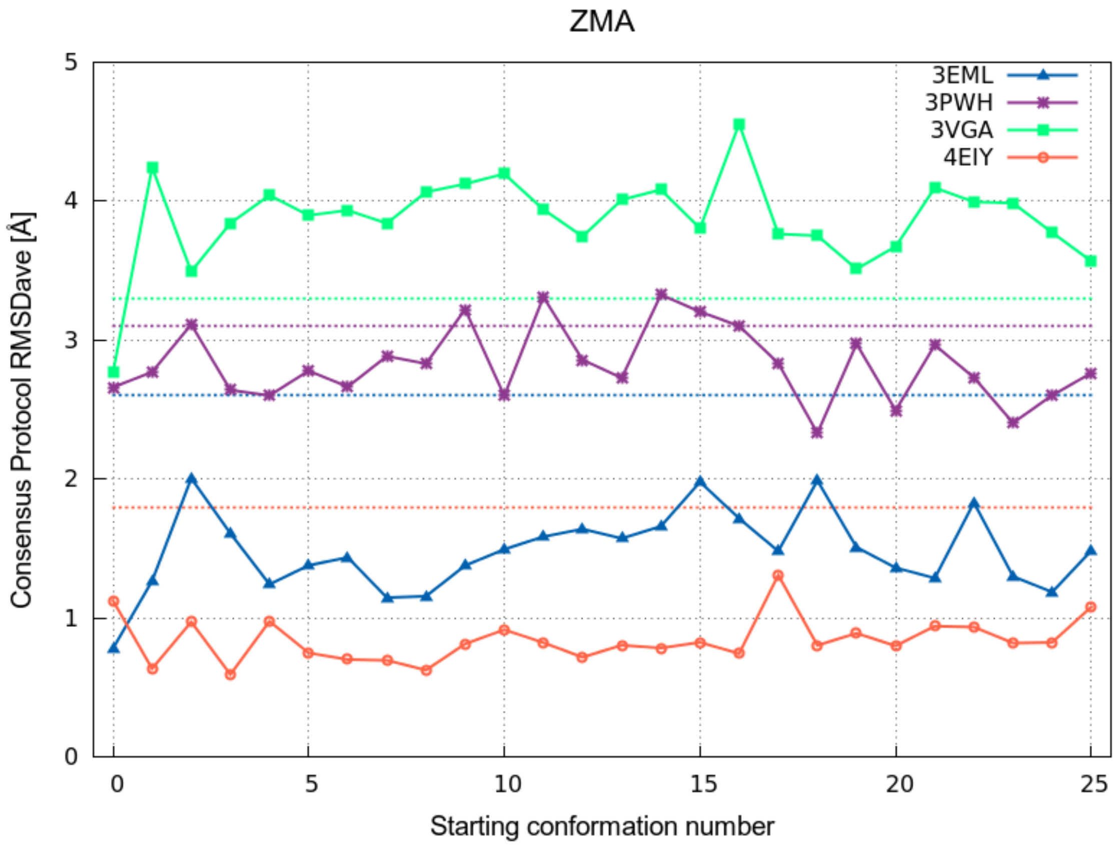


Figure 5 – Effect of starting conformation on consensus $RMSD_{ave.}$ Structure resolution is depicted as a dotted line.

The comparison with the interaction energy computed for the cocrystallized ligand (representing the lowest row of the map) suggests that poses 2, 6, 8, 9, 12, and 13 considerably differ from the binding mode observed in the crystal structure. A superimposition of the poses (Figure 6B, Video S7, Supporting Information) reveals that those structures are placed far away from the binding site by the docking protocol and therefore have a higher RMSD with respect to the cocrystallized ligand. In this specific case, the IEM inspection helped in identifying pose clusters according to the types of interaction they establish with the receptor. The comparison of IEMs relative to the conformation generated by the same protocols for different structures (Figures S4−S8 and Videos S1−S10, Supporting Information) helps in evaluating protocol performances as well as the reproducibility of a crystal binding mode. By comparing panels A and B of Figure S5 (Supporting Information), it is straightforward that the binding mode of ZMA in the 3PWH structure (Video S4, Supporting Information) is more challenging to be reproduced as compared to the one observed in the 3EML structure (Video S3, Supporting Information). However, the irregularity of the interactions in the IEMs of the 3PWH, 3REY, and 3VGA structures (Figures S5B; S6A; and S8A; and Videos S4; S5; and S9, respectively, Supporting

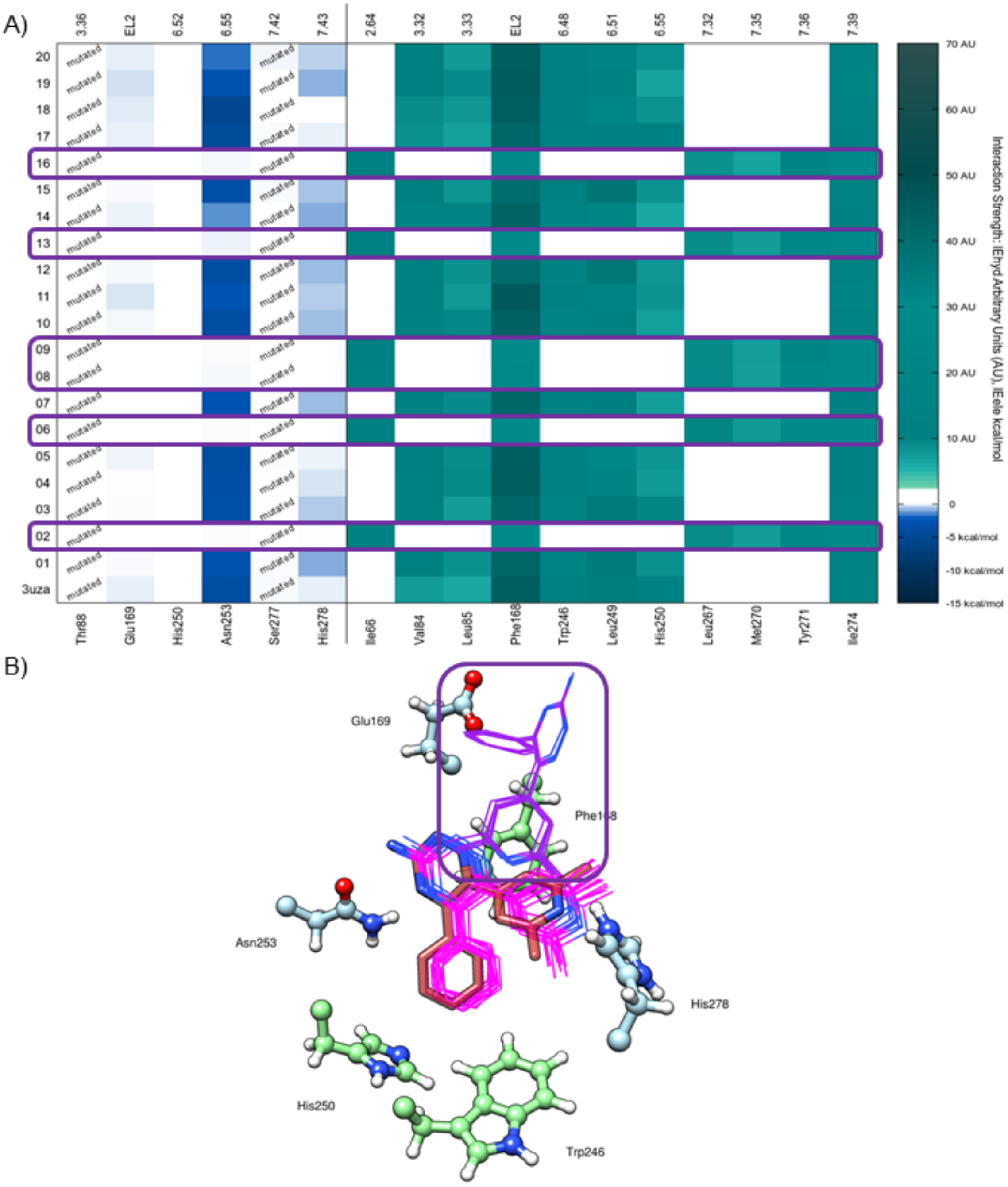


Figure 6 – (A) IEM o the conformations generated by the consensus protocol for the 3UZA structure. **(B)** The two clusters of conformations are superimposed to the crystal structure (red sticks): the poses with the lower RMSD values are represented in magenta wires, whereas poses with higher RMSD are depicted as purple wires.

Information) mirrors the higher average RMSD obtained for those structures. The same conclusions can be drawn by comparing the performances of different protocols on the same structure (data not shown).

We therefore believe that the IEMs can represent a complementary metrics that can help in evaluating the results of docking calculations in a fast graphic way.

Published - Ciancetta, A.; Cuzzolin, A.; Moro, S. *J. Chem. Inf. Model.* **2014**, *54* (8), 2243–2254.

Reconstructed EL2. The results obtained in the ensemble docking procedure show that the protocol preferentially selects complete structures. We therefore decided to evaluate the effect of the reconstruction of the second extracellular loop (EL2) for the structures that have this portion not completely solved. The loop has been reconstructed by using the "Paste Fragment" tool implemented in MOE. The missing portion of the EL2 (see Table S2, Supporting Information) has been copied from the 4EIY structure and pasted into the others, by using the "graft" options, that fits the fragment by superposing the flanking residues of the missing sequence and performs a short minimization. This method represents the least computationally expensive choice and resulted in the most meaningful option in our specific case, as three of the available crystal structures have been solved with complete EL2 and structure superposition highlights very high structural similarity among them (see Figure S3, Supporting Information).

We repeated the cognate ligand docking calculations for the thus-obtained structures by comparing the performances of the consensus protocol prior to and after loop reconstruction (Table 7).

Table 7 - Results of consensus protocol for structures with reconstructed EL2

PDB ID	**Original structure**			**Reconstructed EL2**		
	RMSD$_{ave}$ [Å]	**N$_{(RMSD<R)}$**	**score**	**RMSD$_{ave}$ [Å]**	**N$_{(RMSD<R)}$**	**score**
3EML	0.783	20	3	1.680	20	2
3PWH	2.661	17	3	2.998	15	2
3REY	2.719	11	2	3.490	8	0
3RFM	3.661	12	2	3.684	0	0
3UZA	2.969	14	2	2.376	15	2
3UZC	1.081	19	3	**0.811**	**20**	**3**
3VGA	2.766	16	2	2.939	14	2

As can be noted, the reconstruction of the EL2 worsens the performance of the consensus protocol for the 3REY and 3RFM structures. The 3UZC structure, on the contrary, benefits from loop reconstruction, whereas the performances of other structures are only slightly affected. These results suggest that it is not possible to draw a general conclusion about the benefits of loop reconstruction. We therefore recommend when selecting a crystal structure for docking studies to pay attention to the completeness of the structure in the surroundings of the binding cavity. In case none of the available structures is completely solved, different loop reconstruction methods should be tested.

Starting Ligand Conformation. To evaluate the effects of the starting structure on the docking outcomes, we carried out test calculations by supplying different conformations as input. We performed the analysis by considering 25 different conformations of ZMA as initial structures: 24 conformations were obtained through a stochastic conformational search by leaving the default setting in the MOE conformational search (apart from the RMSD threshold that was increased to 1.0 Å), and another conformation was generated by simply performing an energy minimization. In both cases, the starting point was a structure designed with the MOE builder tool. We evaluated the performances of the consensus protocol for the 3EML, 3PWH, 3VGA, and 4EIY structures (Figure 5). As can be noted, the performances of the consensus protocol are less affected by the starting conformation, and the $RMSD_{ave}$ values remain well below the crystal structure resolution for the 4EIY and 3EML structures, whereas for the 3PWH and to a greater extent for the 3VGA structures the performances are considerably worse. In the case of the 3VGA structure, all of the $RMSD_{ave}$ values are all above the structure resolution. In all cases, no consistent trend is evidenced.

Conclusion

We have presented here an alternative assessment strategy to compare the performances of GPCR-ligand docking protocols based on complementary “quality descriptors”: *a)* the number of conformations generated by a docking algorithm having a RMSD value lower than the crystal structure resolution ($N^{(RMSD<R)}$); *b)* a novel consensus-based function defined as protocol score; and *c)* the IEM analysis, based on the identification of key ligand−receptor interactions observed in the crystal structures. We have selected as test case the hA_{2A} AR in complex with different ligands and evaluated the perform- ances of 16 different docking/scoring combinations in generating poses close to the conformations observed in the X-ray structures. Common settings among the different selected docking programs have been defined, and two issues potentially affecting the docking outcomes, such as the input conformation and the reconstruction of protein missing portions around the binding site, have been tested and discussed.

The conclusions of our analysis can be summarized into a few points: *i)* as expected, no universal docking protocol exists that can reproduce with satisfying accuracy all of the observed X-ray binding modes even when relative to the same receptor subtype cocrystallized with structurally related ligands; *ii)* in the analyzed test case, the overall performances of the docking protocols benefit from the use of a consensus scoring function and are not considerably affected by the input conformation as well as by EL2 reconstruction. Exceptions to those general considerations have been observed for low resolution structures and for structures cocrystallized with low affinity ligands.

In view of the obtained results, we suggest using complementary metrics and additional statistical parameters to the traditional RMSD value to judge and compare the performances of docking protocols.

Abbreviations

ARs	adenosine receptors
EL2	second extracellular loop
GPCRs	G protein-coupled receptors
NECA	N-ethyl-5′- carboxamido adenosine
T4E	4-(3-amino-5-phenyl-1,2,4-tria- zin-6-yl)-2-chlorophenol
T4G	6-(2,6-dimethylpyridin-4-yl)-5- phenyl-1,2,4-triazin-3-amine
TM	transmembrane; ZM 241385, 4-(2-(7-amino-2-(2-furyl)(1,2,4)triazolo(2,3-a)(1,3,5)- triazin-5-yl-amino)ethyl)phenol
XAC	N-(2-aminoethyl)-2-[4- (2,6-dioxo-1,3-dipropyl- 2,3,6,7-tetrahydro-1H-purin-8-yl)-phenoxy]acetamide

Published - Ciancetta, A.; Cuzzolin, A.; Moro, S. *J. Chem. Inf. Model.* **2014**, *54* (8), 2243–2254.

Bibliography

(1) Carlsson, J.; Yoo, L.; Gao, G. Z.; Irwin, J. I.; Shoichet, B. K.; Jacobson, K. A. Structure-based discovery of A2A adenosine receptor ligands. J. Med. Chem. 2010, 53, 3748−3755.

(2) Kolb, P.; Rosenbaum, D. M.; Irwin, J. J.; Fung, J. J.; Kobilka, B. K.; Shoichet, B. K. Structure-based discovery of beta2-adrenergic receptor ligands. Proc. Natl. Acad. Sci. U.S.A. 2009, 106, 6843−6848.

(3) Kufareva, I.; Rueda, M.; Katritch, V.; Stevens, R. C.; Abagyan, R. Status of GPCR modeling and docking as reflected by community- wide GPCR Dock 2010 assessment. Structure 2011, 19, 1108−1126.

(4) Carlsson, J.; Coleman, R. G.; Setola, V.; Irwin, J. J.; Fan, H.; Schlessinger, A.; Sali, A.; Roth, B. L.; Shoichet, B. K. Ligand discovery from a dopamine D3 receptor homology model and crystal structure. Nat. Chem. Biol. 2011, 7, 769−778.

(5) Katritch, V.; Jaakola, V. K.; Lane, J. R.; Lin, J.; Ijzerman, A. P.; Yeager, M.; Kufareva, I.; Stevens, R. C.; Abagyan, R. Structure-based discovery of novel chemotypes for adenosine A(2A) receptor antagonists. J. Med. Chem. 2010, 53, 1799−1809.

(6) Reynolds, K. A.; Katritch, V.; Abagyan, R. Identifying conformational changes of the beta(2) adrenoceptor that enable accurate prediction of ligand/receptor interactions and screening for GPCR modulators. J. Comput.-Aided Mol. Des. 2009, 23, 273−288.

(7) de Graaf, C.; Rognan, D. Selective structure-based virtual screening for full and partial agonists of the beta2 adrenergic receptor. J. Med. Chem. 2008, 51, 4978−4985.

(8) de Graaf, C.; Kooistra, A. J.; Vischer, H. F.; Katritch, V.; Kuijer, M.; Shiroishi, M.; Iwata, S.; Shimamura, T.; Stevens, R. C.; de Esch, I. J.; Leurs, R. Crystal structure-based virtual screening for fragment-like ligands of the human histamine H(1) receptor. J. Med. Chem. 2011, 54, 8195−8206.

(9) Warren, G. L.; Andrews, C. W.; Capelli, A.-M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A critical assessment of docking programs and scoring functions. J. Med. Chem. 2006, 49, 5912−5931.

(10) Katritch, V.; Abagyan, R. GPCR agonist binding revealed by modeling and crystallography. Trends Pharmacol. Sci. 2011, 32, 637− 643.

(11) Vilar, S.; Karpiak, J.; Berk, B.; Costanzi, S. In silico analysis of the binding of agonists and blockers to the beta2-adrenergic receptor. J. Mol. Graphics Modell.

2011, 29, 809−817.

(12) Beuming, T.; Sherman, W. Current assessment of docking into GPCR crystal structures and homology models: successes, challenges, and guidelines. J. Chem. Inf. Model. 2012, 52, 3263−3277.

(13) Jaakola, V. P.; Griffith, M. T.; Hanson, M. A.; Cherezov, V.; Chien, E. Y. T.; Lane, J. R.; Ijzerman, A. P.; Stevens, R. C. The 2.6 angstrom crystal structure of a human A2A adenosine receptor bound to an antagonist. Science 2008, 322, 1211−1217.

(14) Xu, F.; Wu, H.; Katritch, V.; Han, G. W.; Cherezov, V.; Stevens, R. Structure of an agonist-bound human A2A adenosine receptor. Science 2011, 332, 322−327.

(15) Lebon, G.; Warne, T.; Edwards, P. C.; Bennett, K.; Langmead, C. J.; Leslie, A. G. W.; Tate, C. G. Agonist-bound adenosine A2A receptor structures reveal common features of GPCR activation. Nature 2011, 474, 521−525.

(16) Dore,́ A. S.; Robertson, N.; Errey, J. C.; Ng, I.; Hollenstein, K.; Tehan, B.; Hurrell, E.; Bennett, K.; Congreve, M.; Magnani, F.; Tate, C. G.; Weir, M.; Marshall, F. H. Structure of the adenosine A2A receptor in complex with ZM241385 and the xanthines XAC and caffeine. Structure 2011, 19, 1283−1293.

(17) Hino, T.; Arakawa, T.; Iwanari, H.; Yurugi-Kobayashi, T.; Ikeda- Suno, C.; Nakada-Nakura, Y.; Kusano-Arai, O.; Weyand, S.; Shimamura, T.; Nomura, N.; Cameron, A. D.; Kobayashi, T.;

Hamakubo, T.; Iwata, S.; Murata, T. G-protein-coupled receptor inactivation by an allosteric inverse-agonist antibody. Nature 2012, 482, 237−240.

(18) Congreve, M.; Andrews, S. P.; Dore, A. S.; Hollenstein, K.; Hurrell, E.; Langmead, C. J.; Mason, J. S.; Ng, I. W.; Tehan, B.; Zhukov, A.; Weir, M.; Marshall, F. H. Discovery of 1,2,4-triazine derivatives as adenosine A2A antagonists using structure based drug design. J. Med. Chem. 2012, 55, 1898−1903.

(19) Liu, W.; Chun, E.; Thompson, A. A.; Chubukov, P.; Xu, F.; Katritch, V.; Han, G. W.; Heitman, L. H.; Ijzerman, A. P.; Cherezov, V.; Stevens, R. C. Structural basis for allosteric regulation of GPCRs by sodium ions. Science 2012, 337, 232−236.

(20) Müller, C. E.; Jakobson, K. A. Recent developments in adenosine receptor ligands and their potential as novel drugs. Biochim. Biophys. Acta 2011, 1808, 1290−1308.

(21) Perez-Nueno, V. I.; Rabal, O.; Borrell, J. I.; Teixido, J. APIF: a new interaction fingerprint based on atom pairs and its application to virtual screening. J. Chem. Inf. Model. 2009, 49, 1245−1260.

(22) Novikov, F. N.; Stroylov, V. S.; Stroganov, O. V.; Kulkov, V.; Chilov, G. G.

Developing novel approaches to improve binding energy estimation and virtual screening: a PARP case study. J. Mol. Model. 2009, 15, 1337−1347.

(23) Agostino, M.; Sandrin, M. S.; Thompson, P. E.; Yuriev, E.; Ramsland, P. In silico analysis of antibody-carbohydrate interactions and its appli-cation to xenoreactive antibodies. Mol. Immunol. 2009, 47, 233−246.

(24) Agostino, M.; Sandrin, M. S.; Thompson, P. E.; Yuriev, E.; Ramsland, P. A. Identification of preferred carbohydrate binding modes in xenor-eactive antibodies by combining conformational filters and binding site maps. Glycobiology 2010, 20, 724−735.

(25) Ballesteros, J. A.; Weinstein, H. Integrated methods for the construction of three dimensional models and computational probing of structure-function relationships in G-protein coupled receptors. Methods Neurosci. 1995, 25, 366−428.

(26) MOE (Molecular Operating Environment), version 2012.10; Chemical Computing Group Inc.: Montreal, Canada, 2012.

(27) Williams, T. Kelley, C. Gnuplot: An Interactive Plotting Program; available at http://www.gnuplot.info (accessed Apr 18, 2014).

(28) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. Protein Data Bank. Nucl. Acids. Res. 2000, 28, 235−242.

(29) Labute, P. Protonate3D: assignment of ionization states and hydrogen coordinates to macromolecular structures. Proteins 2009, 75, 187−205.

(30) Wang, J.; Cieplak, P.; Kollman, P. A. How well does a restrained electrostatic potential (RESP) model perform in calculating conforma- tional energies of organic and biological molecules? J. Comput. Chem. 2000, 21, 1049−1074.

(31) Stewart, J. J. P. Optimization of parameters for semiempirical methods I. Method. J. Comput. Chem. 1989, 10, 209−220.

(32) Stewart, J. J. P. Optimization of parameters for semiempirical methods II. Applications. J. Comput. Chem. 1989, 10, 221−264.

(33) Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J. Autodock4 and AutoDockTools4: automated docking with selective receptor flexiblity. J. Comput. Chem. 2009, 16, 2785−2791.

(34) Glide, version 5.8; Schrödinger, LLC: New York, NY, 2012.

(35) GOLD Suite, version 5.0; Cambridge Crystallographic Data Centre: Cambridge, UK. http://www.ccdc.cam.ac.uk.

(36) Korb, O.; Stützle, T.; Exner, T. E. PLANTS: application of ant colony

optimization to structure-based drug design. Lec. Notes Comput. Sci. 2006, 4150, 247−258.

(37) Thomsen, R.; Christensen, M. H. MolDock: A new technique for high-accuracy molecular docking. J. Med. Chem. 2006, 49, 3315− 3321.

(38) Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. J. Comput. Chem. 1996, 17, 490−519.

(39) Sabbadin, D.; Ciancetta, A.; Moro, S. Bridging molecular docking to membrane molecular dynamics to investigate GPCR− ligand recognition: The human A2A adenosine receptor as a key study. J. Chem. Inf. Model. 2014, 54, 169−183.

(40) Sabbadin, D.; Ciancetta, A.; Moro, S. Perturbation of fluid dynamics properties of water molecules during GPCR-ligand recognition: the human A_{2A} adenosine receptor as a key study. J. Chem. Inf. Model. 2014, Submitted.

(41) Jaakola, V.-P.; Lane, J. R.; Lin, J. Y.; Katritch, V.; Ijzerman, A. P.; Stevens, R. C. Ligand binding and subtype selectivity of the human A(2A) adenosine receptor: Identification and characterization of essential amino acid residues. J. Biol. Chem. 2010, 285, 13032−13044.

(42) Cristalli, G.; Lambertucci, C.; Marucci, G.; Volpini, R.; Dal Ben, D. A2A adenosine receptor and its modulators: Overview on a druggable GPCR and on structure-activity relationship analysis and binding requirements of agonists and antagonists. Curr. Pharm. Des. 2008, 14, 1525−1552.

(43) Cheng, T.; Li, X.; Li, Y.; Liu, Z.; Wang, R. Comparative assessment of scoring functions on a diverse test set. J. Chem. Inf. Model. 2009, 49, 1079−1093.

3.3 DockBench: An Integrated Informatic Platform Bridging the Gap between the Robust Validation of Docking Protocols and Virtual Screening Simulations

Alberto Cuzzolin †, Mattia Sturlese †, Ivana Malvacio †, Antonella Ciancetta and Stefano Moro*

Abstract

Virtual screening (VS) is a computational methodology that streamlines the drug discovery process by reducing costs and required resources through the in silico identification of potential drug candidates. Structure-based VS (SBVS) exploits knowledge about the three-dimensional (3D) structure of protein targets and uses the docking methodology as search engine for novel hits. The success of a SBVS campaign strongly depends upon the accuracy of the docking protocol used to select the candidates from large chemical libraries. The identification of suitable protocols is therefore a crucial step in the setup of SBVS experiments. Carrying out extensive benchmark studies, however, is usually a tangled task that requires users' proficiency in handling different file formats and philosophies at the basis of the plethora of existing software packages. We present here DockBench 1.0, a platform available free of charge that eases the pipeline by automating the entire procedure, from docking benchmark to VS setups. In its current implementation, DockBench 1.0 handles seven docking software packages and offers the possibility to test up to seventeen different protocols. The main features of our platform are presented here and the results of the benchmark study of human Checkpoint kinase 1 (hChk1) are discussed as validation test.

Introduction

Virtual screening is a computational methodology aimed at streamlining the drug discovery process through the in silico identification of novel hits from large chemical libraries[1]. After emerging in the late 1990s[2] as a strategy to reduce the time and cost of chemical synthesis and in vitro testing, VS nowadays represents an integral part of the drug discovery pipeline both in industry and in academic environments[3]. The main purpose of a VS campaign is to select appropriate compounds while removing unsuitable structures, thus significantly reducing costs and required resources. Depending on the amount of information available about the system of interest, VS is historically classified into two main categories[4]: ligand-based VS (LBVS) and structure-based VS (SBVS). SBVS exploits knowledge about the three-dimensional (3D) structure of the target gathered either experimentally by X-ray crystallography or NMR spectroscopy, or computationally through homology modeling and performs docking calculations to rank candidates

Published - Cuzzolin, A.; Sturlese, M.; Malvacio, I.; Ciancetta, A.; Moro, S. *Molecules* **2015**, *20* (6), 9977–9993.

on the basis of estimated binding affinity or complementarity to the binding site[5].

Consequently, the success of a SBVS campaign strongly depends upon the accuracy of the engine used to generate, place, and rank the conformation of candidates into a target binding site[6]. A crucial step in the setup of a SBVS experiment is therefore the selection of a proper docking protocol, *i.e.*, the combination of search algorithm and scoring function that yields the best accuracy achievable.

The comparison of different docking protocols is not a trivial task[7], as it requires expertise in handling the different philosophies behind the large variety of available software packages. As a result, non-expert users are usually discouraged in enriching the pool of docking programs to test due to difficulties in input and output formats syntax comprehension, conversion and management. Moreover, the time required in merging and comparing the results arising from different protocols usually is incompatible with the requests of the experimental counterpart.

Within this framework, we have recently proposed[8] a strategy to compare the performances of docking protocols based on two quality metrics: The “Protocol Score”, and the number of conformations generated by the docking protocol with a RMSD value below the resolution (R) of the crystal structure “$N^{(RMSD<R)}$”. With the aim to broaden their exploitation also by non-expert users, we have proposed the presentation of the results as coloured maps of immediate interpretation in a benchmark study focused on the human adenosine 2A receptor.

In the present work, we move a step forward and present DockBench 1.0, a platform available free of charge upon request that fully automates the entire procedure from the setup of docking benchmarks to VS campaigns. In its current implementation, DockBench 1.0 handles seven different docking software packages and provides the user with the possibility to test up to seventeen protocols. A GUI guides the user step-by-step throughout all the stages required to perform the entire pipeline, from the choice of the docking protocol to assess to the VS of large chemical libraries. The results are expressed in terms of the above mentioned quality metrics and returned as easy to interpret coloured maps. The outputs of the different software packages are returned in a unique format and are analysed with a standardized procedure to avoid software related biases.

We describe as validation case a docking benchmark study focused on human checkpoint kinase 1 (hChk1). hChk1 is a serine/threonine kinase responsible for the arrest of the cell cycle that allow DNA repair in tumour cells in response to a damage[9]. Therefore, hChk1 inhibition represents a strategy to increase the therapeutic efficacy of anticancer drugs, thus enhancing the apoptosis induced by alkylating agents[10,11].

Published - Cuzzolin, A.; Sturlese, M.; Malvacio, I.; Ciancetta, A.; Moro, S. *Molecules* **2015**, *20* (6), 9977–9993.

Results and Discussion

DockBench General Features

The Flowchart of the DockBench 1.0 platform is reported in Figure 1. All the functionalities are embedded in a graphical user interface (GUI, Figure 2) and are organized into five main tabs, corresponding to the tasks required to carry out a complete pipeline, from docking benchmark studies to VS experiments: (1) Input Settings; (2) Docking Protocols Settings; (3) Results Visualization; (4) Plots Visualization; (5) Virtual Screening Settings. The main features of each tab are discussed in the following. DockBench 1.0 is available free of charge and can be requested at the project page[12].

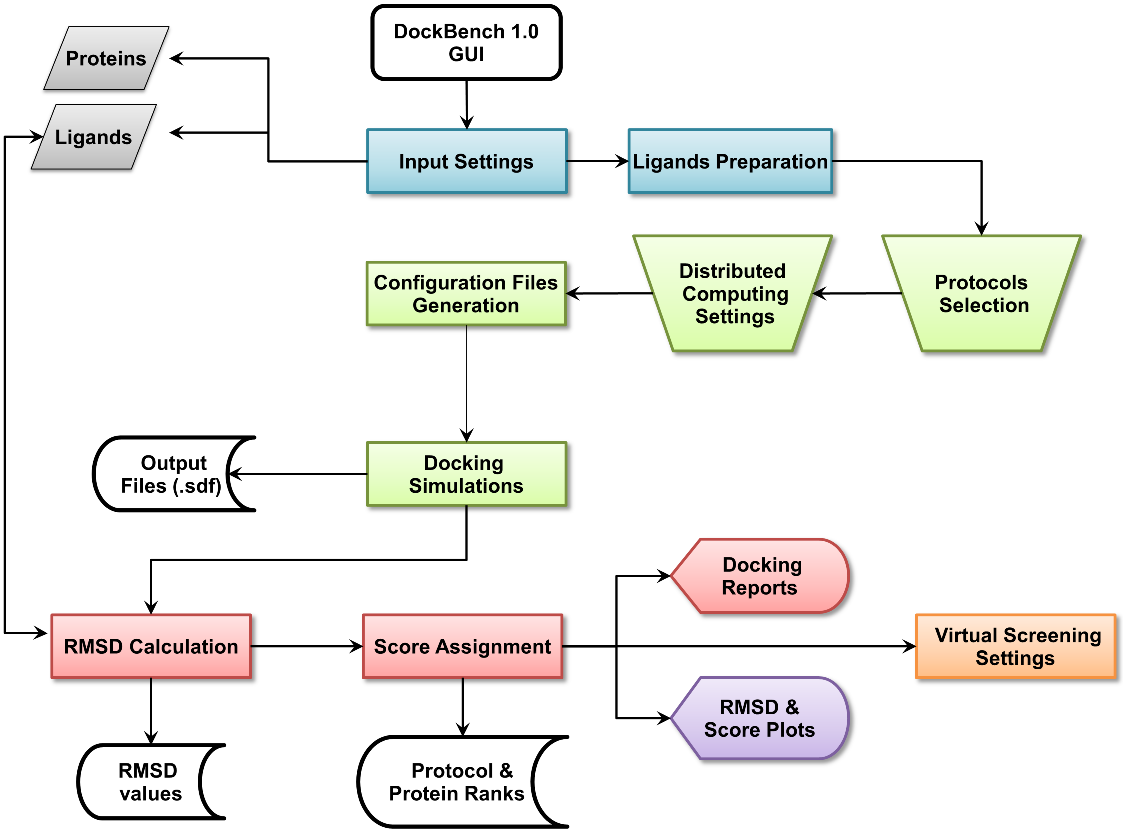


Figure 1 - DockBech 1.0 workflow. The platform is accessed through a GUI, the different stages of the pipeline are highlighted with different colours.

Published - Cuzzolin, A.; Sturlese, M.; Malvacio, I.; Ciancetta, A.; Moro, S. *Molecules* **2015**, *20* (6), 9977–9993.

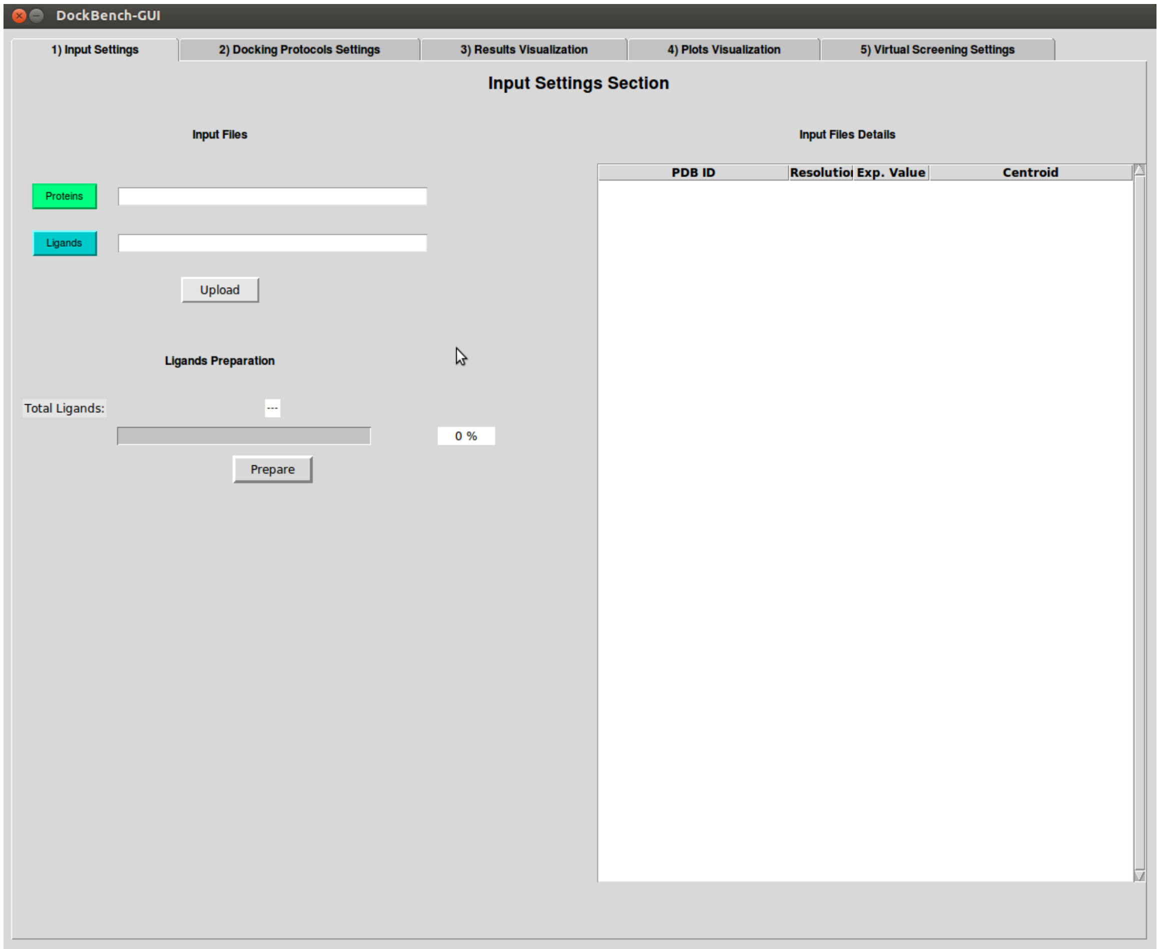


Figure 2 - DockBench GUI tabs: (1) Input Settings; (2) Docking Protocols Settings; (3) Results Visualization; (4) Plots Visualization; (5) Virtual Screening Settings.

Input Settings

DockBench 1.0 significantly eases benchmark and VS procedures and allows the user to submit jobs with the different implemented software packages at once. The user is asked to provide a few files including ligands and receptors structures in Tripos .mol2 format and receptor structures .pdb format as retrieved from the Protein Data Bank (PDB)[13] (Figure 1, grey boxes). The supplied structures, apart from original pdb files, must be prepared in advance. In particular, hydrogen atoms need to be added by setting the correct protonation states for both the ligand and the protein structures. Moreover, the user must take care of generating proper ligands tautomeric and stereoisomeric states. Once the structures have been uploaded, the R values, ligand names and pdb codes are automatically extracted from PDB Remark section, displayed in a table on the GUI and saved for subsequent use for files nomenclature and results visualization. Available binding data information for the co-crystallized ligands are directly retrieved from PDB source page and displayed. In case several data are available, the following selection hierarchy is applied: Ki, Kd, and IC50.

After the input structures have been uploaded, the coordinates of ligands centroids are computed according to Equation 1) and set as the binding cavity centre for the subsequent docking simulations. To avoid biases due to input conformations, a ligands preparation step (Figure 1, blue boxes) is performed with the obminimize[14] tool, as detailed in the Methods section.

Docking Protocols Settings

In the protocol selection tab all the implemented docking software packages are listed. In case a docking program is not available to the user, it will be automatically set as inactive. DockBench 1.0 offers the possibility to select up to 17 different protocols, sorted alphabetically as reported in Table 1. Briefly, AutoDock[15] is embedded with three different global optimizer approaches coupled with the AutoDock Scoring Function: Genetic algorithm (GA), Lamarkian genetic algorithm (LGA), and local search (LS). AutoDock Vina[16] is included with its standard optimization algorithm and standard scoring function. Glide[17,18], is implemented with the Standard Precision mode. Four scoring functions are available for the GOLD suite (ASP, Chemscore, Goldscore and PLP)[19]. Plants[20] is available with three different scoring functions[21] (ChemPLP, PLP, PLP95) that are coupled to the Ant Colony Optimization (ACO) algorithm. The Triangle Matcher placing method of the MOE docking tool is implemented along with three different scoring functions (Affinity dG, London dG, GBVI/WSA)[22]. rDock[23] can be run with or without desolvation potential term with its standard scoring function. Each protocol is managed independently by providing the user with the possibility to select all of them or individual ones.

Several advanced options (Figure 1, green trapezoids) can be customized by the user prior to running the docking simulations: The number of output poses (default 20), the threshold RMSD value to define unique poses (default 1.0 Å, not available for AutoDock Vina and rDOCK), and the radius (default 20 Å) of the binding site. As several software packages describe the binding site using inclusion spheres (GOLD, PLANTS, rDOCK), the sphere radius *r* is set as common parameter to define the cavity. Nevertheless, to maintain comparable volumes for the protocols adopting parallelepiped-shaped cavities, the cube side *l* is scaled according to Equation (2). Along with the options pertaining to the configuration files, DockBench 1.0 allows the user to optimize calculations performances by setting distributed computing and licenses management features. This functionality is designed to take advantage of multicore CPUs and makes a sophisticated use of semaphores, as implemented in GNU Parallel[24]. In details, all the jobs (docking runs) are classified and redirect to hardware resources according to two parameters: The total number of cores to be used—that is automatically detected by DockBench 1.0 but that can be edited by the user (*i.e.*, in case the calculations will run on a remote machine with a different cores number)—and the number of licenses available for commercial software packages. According to a classification based on the presence of licenses, the jobs are launched in different “traffic lines”: Protocols without license limits are redirected to the same traffic line, whereas to each licensed program a unique traffic line will be reserved.

Table 1 - List of docking protocols available in DockBench 1.0

Program	**Search Algorithm/Placing Method**	**Scoring Function**	**Protocol Abbreviation**
Autodock 4.2	Local Search Lamarkian GA Genetic Algorithm	AutoDock SF AutoDock SF AutoDock SF	AUTODOCK-ls AUTODOCK-lga AUTODOCK-ga
AutoDock Vina 1.1.2	Monte Carlo + BFGS local search	Standard Vina SF	VINA-std
Glide 6.5	Glide Algorithm	Standard Precision	GLIDE-sp
GOLD 5.2	Generic Algorithm Genetic Algorithm Generic Algorithm Genetic Algorithm	Goldscore Chemscore ASP PLP	GOLD-goldscore GOLD-chemscore GOLD-asp GOLD-plp
MOE 2014.09	Triangle Matcher Triangle Matcher Triangle Matcher	London-dG Affinity-dG GBIVIWSA	MOE-londondg MOE-affinitydg MOE-gbiviwsa
PLANTS 1.2	ACO Algorithm ACO Algorithm ACO Algorithm	PLP PLP95 ChemPLP	PLANTS-plp PLANTS-plp95 PLANTS-chemplp
rDock 2013.1	Genetic Algorithm + Monte Carlo + Simplex minimization Genetic Algorithm + Monte Carlo + Simplex minimization	Standard rDock master SF Standard rDock master SF + desolvation potential	RDOCK-std RDOCK-solv

Published - Cuzzolin, A.; Sturlese, M.; Malvacio, I.; Ciancetta, A.; Moro, S. *Molecules* **2015**, *20* (6), 9977–9993.

The number of licenses defines the width of the unique traffic lines, *i.e.*, how many jobs will simultaneously run for a given program. Therefore, the traffic lines reserved for licensed software packages will be subtracted from total number of cores and saturated by non-licensed jobs. For instance, on a workstation equipped with an eight core/threads CPU, DockBench1.0 (with default settings) will run simultaneously one GLIDE job, one GOLD job, and one MOE job. The remaining five cores will be saturated by the protocols not limited by licenses (AutoDock, AutoDock Vina, PLANTS, and rDOCK).

Results Visualization

At the end of the docking simulations (Figure 1, green boxes), DockBench1.0 converts all the output files from formats specific to each docking software package to structure-data files (.sdf). A check is performed to detect any missing output, thus automatically identifying job failures. A summary of the chosen options as well as details on considered ligand-protein systems and tested protocols is reported in a format table on the GUI. For each structure-docking protocol pair, minimum ($RMSD_{min}$), maximum ($RMSD_{max}$) and average RMSD ($RMSD_{ave}$) values with respect to the X-ray binding mode are calculated (Equation (3), Experimental Section) and a text file summarizing all these results is produced (Figure 1, red boxes). These value are then used to compute the quality metrics[8] $N^{(RMSD<R)}$ and the Protocol Score. At this stage, protocol and protein ranks are drafted and displayed in tabular format on the GUI according to the computed Protocol Score values. Protocol based ranks are derived by summing up the scores obtained by each protocol for all the considered protein structures. Protein based ranks are compiled by listing in descending orders the protein structures with higher sums of protocol scores. The GUI allow the user to shift from one rank to another according to which piece of information is considered more relevant.

Plots Visualization

DockBench 1.0 provide the users with the possibility to graphically display the results as easy to interpret coloured maps. In the plot visualization tab, four plots depicting the $RMSD_{min}$, $RMSD_{ave}$, $N^{(RMSD<R)}$ and Protocol Score trends are displayed. These graphs along with the above mentioned ranks are intended to guide the user in the selection of the best performing protocols as well as in the protein structure yielding more robust results for the subsequent VS jobs. In particular, each plot returns the list of tested protocols against the considered systems and display the analysed value ($RMSD_{min}$, $RMSD_{ave}$, $N^{(RMSD<R)}$ or Protocol Score) with a colour code. To ease the interpretation of the results, colour codes have been devised so that blue spots identify the best results obtained for each value.

Virtual Screening Settings

As anticipated, DockBench 1.0 offers the possibility to perform VS campaigns by selecting one or more of the previously evaluated docking protocols (Figure 1, orange box). The user is asked to upload a molecular database in .sdf format and has the possibility to automatically include the ligands used for the benchmark study (useful for enrichment analyses), and to define the number of posed to be returned for each ligand. Depending on the size of the loaded library and on the performance of the selected protocol detected during the benchmark procedure, an estimate of the time required to screen the whole library is provided. Similarly to the benchmark calculations, the VS scheme takes advantage of GNU parallel[24]. Calculation can be performed on a single workstation as well as on a cluster, by indicating the hostname and the number of cores to be used for each node. The jobs are monitored and in case of interruption, a restart input file is provided. To further speed up the calculations, the loaded library is splitted according to its size into more sdf files with an in-house python script implemented in the code. At the end of the VS procedure, the resulting conformers are merged and a global ranking is performed.

Case Study

The results of our validation test are reported in Figure 3. The $RMSD_{min}$ analysis highlights the protocols (VINA-std; GOLD-plp; GOLD-goldscore; GOLD-asp and AUTODOCK-ga) able to generate at least one pose that reproduces the X-ray observed binding mode with significant accuracy (Figure 3A). Some of these protocols, however, worsen their performances when $RMSD_{ave}$ values are inspected (Figure 3B). Conversely, other protocols that accurately reproduced at least once the crystallographic pose for a given structure (GOLD-asp/c73-3PA5) show $RMSD_{ave}$ values over the structures resolutions.

By analysing the data in terms of $N^{(RMSD<R)}$, it emerges that there are few protocols able to generate a high $N^{(RMSD<R)}$ and that only in the 10% of the examined cases (32/340) all the conformations generated by the protocol have RMSD value below the structure resolution ($N^{(RMSD<R)}$= 20, blue spots in Figure 3C).

The inspection of the Protocol Score results (Figure 3D) reveals that some protocols (RDOCK-solv, GOLD-plp, GOLD-goldscore, GOLD-asp and AUTODOCK-lga) generate the highest score for at least one protein structure.

At a first glance, these results suggest that it is not possible to identify the best docking protocol for all the considered structures. Therefore, the selection of a proper protocol for subsequent docking simulations depends upon the selected protein structure.

Published - Cuzzolin, A.; Sturlese, M.; Malvacio, I.; Ciancetta, A.; Moro, S. *Molecules* **2015**, *20* (6), 9977–9993.

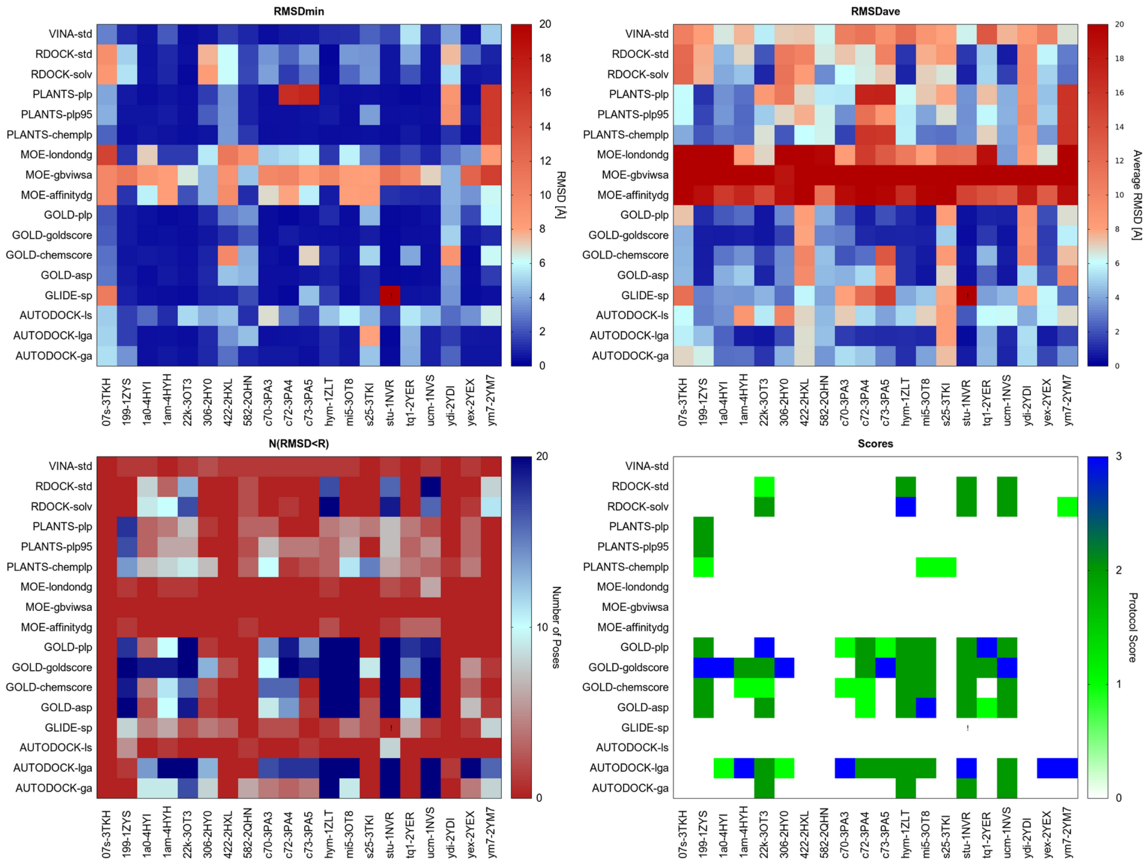


Figure 3 – Results of the docking benchmark study on human checkpoint kinase 1. **(A)** Minimum RMSD values ($RMSD_{min}$) returned by the tested docking protocol (y-values) for the considered X-ray structures (x-values); **(B)** Average RMSD values ($RMSD_{ave}$); **(C)** Numbers of conformations returned by each docking protocol having a RMSD value lower than the X.ray structure resolution ($N^{(RMSD<R)}$); **(D)** Protocol Score. RMSD is expressed in Å, whereas the Protocol Score ona 0-3 points scale. Values are rendered with a colour code, blue spots identify the best obtained results. Exclamation marks warn that an error occurred during docking calculations.

For instance, GOLD-goldscore could be used coupled to structures corresponding to PDB codes 1ZYS and 1NVS, whereas AUTODOCK-lga could be used in conjunction with the 1NVR structure. Overall, AUTODOCK-lga and GOLD-goldscore represent the protocols yielding the highest scores for a greater number of different proteins.

DockBench 1.0 Performances

To evaluate the performances of the distributed computing system we integrated in DockBench 1.0, we have tested the efficiency in the jobs management by DockBench 1.0 as compared with a traditional one by one job routine. In Table 2, the average execution time and the total calculation time for each protocol are reported. The docking calculation of the whole hChk1 case study (20 proteins; 17 Protocols) was achieved by the traditional routine in 16 h and 54 min. To complete the same task, DockBench 1.0 spent in 2 h 24 min, by using two licenses for GOLD, two licenses for GLIDE, two licenses for MOE and no license limit for the other software packages. It has to be pointed out that the DockBench 1.0 performances in this comparison were mainly affected by the low number of licenses used. A more reliable comparison has been drawn by running the same case study by using only non-licensed protocols (AutoDock, PLANTS, rDock, Vina). In this case, the traditional routine spent 11 h 13 min whereas DockBench 1.0 carried out the calculations in 27 min.

Table 2 - ***Cont.***

Abbreviation	**Average Execution Time (s)**	**Total Time (s)**
AUTODOCK-ga	973.5	20,445.3
AUTODOCK-lga	633.3	13,299.1
AUTODOCK-ls	7.45	156.58
GLIDE-sp	46.8	984.2
GOLD-asp	133.4	2,801.8
GOLD-chemscore	136.2	2,860.4
GOLD-goldscore	401.7	8,436.5
GOLD-plp	98.6	2,071.9
PLANTS-chemplp	61.4	1,290
PLANTS-plp	23.3	958.1
PLANTS-plp95	16.6	348.3
MOE-affinitydg	17.6	352.5
MOE-londondg	18.4	368.4
MOE-gbviwsa	131.9	2,638.8
RDOCK-std	20.0	426.2
RDOCK-solv	31.9	671.0
VINA-std	132.7	2,786.7

Published - Cuzzolin, A.; Sturlese, M.; Malvacio, I.; Ciancetta, A.; Moro, S. *Molecules* **2015**, *20* (6), 9977–9993.

Experimental Section

Computational Facilities

All computations were performed on a 200 cores cluster based on Ubuntu OS (14.04, 64 bit) and under the network file system (NFS) service. Performance timing of DockBench 1.0 was performed on a single HP ProLiant server DL585G7, equipped with four AMD Opteron Processor 6282 servers, for a total of 64 CPU cores.

DockBench 1.0 Platform

Programming Languages and Software Dependencies

DockBench 1.0 is written in Python and patches several Bash scripts to launch and analyse molecular docking simulations. To integrate the MOE docking tool[22], in-house built Scientific Vector Language (SVL) scripts have been embedded in the code. DockBench 1.0 also integrates third party applications and the following packages are required to fully utilize the platform features: OpenBabel chemical toolbox 2.3.2[14], GNU parallel 20130922[24] and Gnuplot 4.6.

Names Conventions

All the files generated by DockBench 1.0 are named according to the following scheme: “Ligand abbreviation—protein identifier—protocol abbreviation”. Ligands abbreviations correspond to the three letter codes assigned in the PDB files, whereas proteins identifiers are the corresponding PDB entries. Docking protocols abbreviations (Table 1) are named according to the following scheme: “Program name abbreviation-scoring function/search algorithm”.

Implemented Docking Protocols and Standard Settings

In its current implementation, DockBench 1.0 handles the following docking software packages for a total of 17 different protocols (see Table 1 for more details): AutoDock 4.2.5.1 [15], AutoDock Vina1.1.2 [16], Glide 6.5 [17,18], GOLD 5.2 [25], MOE 2014.09 [22], PLANTS 1.2 [20], rDock [21]. Several common options among the different protocols have been set (Table 3).
The coordinates of the binding cavity centre (centroid) are computed as the weighted centre of mass of all ligand atoms (Equation (1)):

$$Centroid = \left(\frac{\sum_i^n (x_i * m_i)}{\sum_i^n m_i}\right), \left(\frac{\sum_i^n (y_i * m_i)}{\sum_i^n m_i}\right), \left(\frac{\sum_i^n (z_i * m_i)}{\sum_i^n m_i}\right) \quad (1)$$

Published - Cuzzolin, A.; Sturlese, M.; Malvacio, I.; Ciancetta, A.; Moro, S. *Molecules* **2015**, *20* (6), 9977–9993.

Table 3 - Common docking settings for the evaluated protocols

Parameter	**Value/Setting**
Ligand input conformation	Structures generated by minimization
Ligand initial partial charges	Provided by the user
Water molecules	Excluded
Output	20 conformations (*customizable*)
RMSD threshold	1.0 Å (*customizable*)
Binding cavity centre (*Centroid*)	Ligand barycenter in X-ray structure
Binding cavity radius (r)	20 Å (*customizable*)
Grid spacing (for grid-based calculations)	0.475 Å
Refinement and re-scoring	Turned off

To maintain similar cavity volumes for the protocols defining the binding cavity with a parallelepiped, we set cubes having similar volume to the sphere by scaling the side, *l*, according to Equation (2):

$$l = \sqrt[3]{\frac{4\pi}{3}r} \qquad (2)$$

Moreover, at variance with the previously published procedure[8], a pre-processing step of the input conformations has been implemented to avoid biases arising from ligand input conformation. The input structures are therefore minimized with the minimize tool[14], using the conjugate gradient algorithm and a maximum of 2500 steps to reach convergence criteria of 1e-16 based on the MMFF94 force field[26]. Finally, RMSD values with respect to the co-crystallized ligands are calculated as reported in Equation (3) with an in-house built Python script. Given two sets of n heavy atoms a and b:

$$RMSD(a,b) = \sqrt{\frac{1}{n}\sum_{i=i}^{n}((a_{ix}-b_{ix})^2+(a_{iy}-b_{iy})^2+(a_{iz}-b_{iz})^2)} \qquad (3)$$

Case Study Input Files Preparation

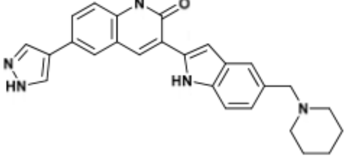
Protein Structures

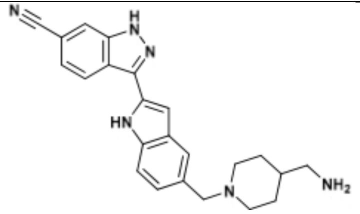
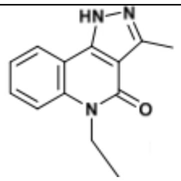
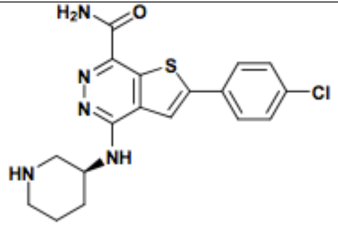
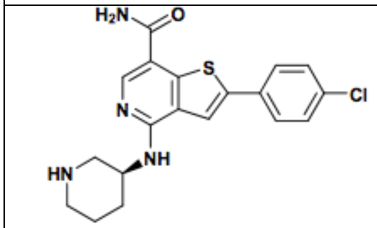
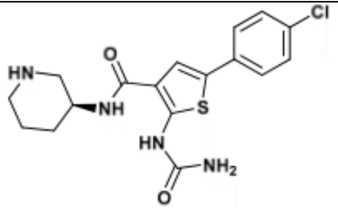
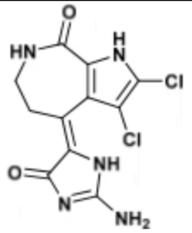
Among the 108 available X-ray structures for the hChk1, the following 20 ligand-protein complexes were selected for the docking benchmark (PDB IDs): 3TKH[27], 3TKI[27], 1ZYS[28], 4HYI[29], 4HYH[29], 3OT3[30], 2HY0[31], 2HXL[31], 2QHN[32], 3PA3[33], 3PA4[33], 3PA5[33], 1ZLT[34], 3OT8[35], 1NVR[36], 1NVS[36], 2YEX[37], 2YER[37], 2YDI[38], 2YM7[39]. The structures were retrieved from the RCSB PDB database[13] and selected on the basis of their X-ray resolution (R, selection criterion = R < 1.8 Å). Before the preparation procedure, all the proteins were aligned and superimposed to a selected reference structure. Crystallization solvent and ions were removed, whereas water molecules and co-crystallized ligands were retained for the hydrogen atoms assignment step and then removed. Ionization states and hydrogen positions were assigned with the 'Protonate-3D' tool[40], as implemented in the Molecular Operating Environment (MOE, version 2014.09) suite[22]. Then, the structures were subjected to energy minimization with Amber99 force field[41], by keeping the heavy atoms fixed at their positions. Finally, ligand and water molecules were removed and protein atoms partial charges computed with the Amber99 force field[41].

Ligand Structures

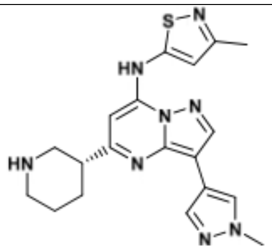
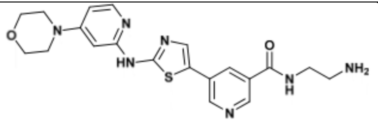
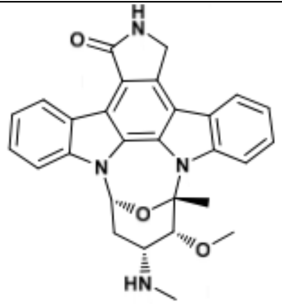
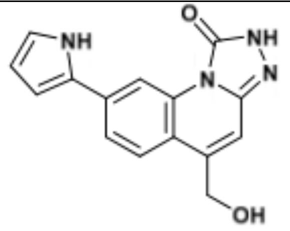
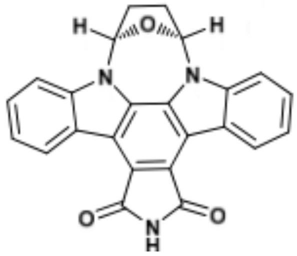
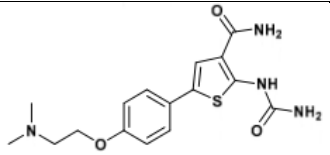
Co-crystallized ligands were extracted from the corresponding crystallographic complex and checked for errors. Hydrogen atoms were added and the protonation state (pH: 7.4) was assigned. Partial charges on ligands atoms were computed on the basis of the PM3/ESP semiempirical Hamiltonian[42,43]. The structures have been then subjected to the ligand preparation procedure of the DockBench 1.0 platform. A full list of ligands considered in this study along with their structures and names is reported in Table 4.

Table 4 - Cont.

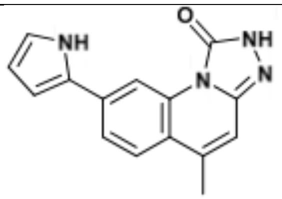
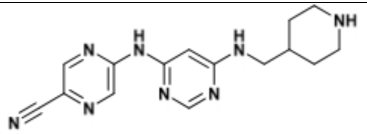
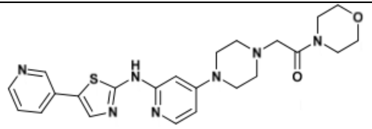
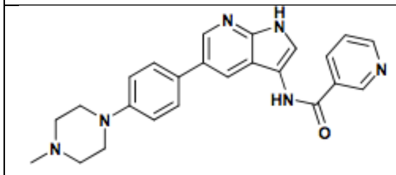
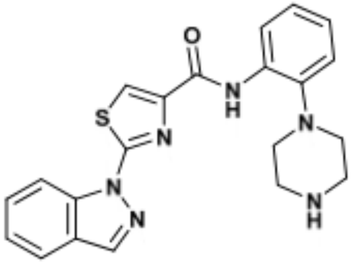
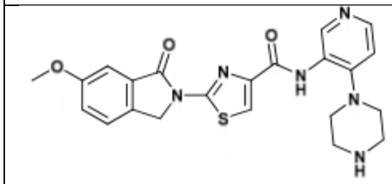
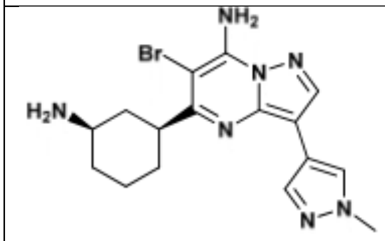
Structure	IUPAC Name	Ligand Abbreviation
	3-[5-(piperidin-1-ylmethyl)-1H-indol-2-yl]-6-(1H-pyrazol-4-yl)-1H-quinolin-2-one	306-2HY0

	3-[5-[[4-(aminomethyl)piperidin-1-yl]methyl]-1H-indol-2-yl]-1H-indazole-6-carbonitrile	422-2HXL
	5-ethyl-3-methyl-1H-pyrazolo[4,5-c]quinolin-4-one	582-2QHN
	2-(4-chlorophenyl)-4-[[(3S)-piperidin-3-yl]amino]thieno[2,3-d]pyridazine-7-carboxamide	C70-3PA3
	2-(4-chlorophenyl)-4-[[(3S)-piperidin-3-yl]amino]thieno[3,2-c]pyridine-7-carboxamide	C72-3PA4
	2-(aminocarbonylamino)-5-(4-chlorophenyl)-N-[(3S)-piperidin-3-yl]thiophene-3-carboxamide	C73-3PA5
	(4Z)-4-(2-amino-5-oxo-3H-imidazol-4-ylidene)-2,3-dichloro-1,5,6,7-tetrahydropyrrolo[2,3-c]azepin-8-one	Hym-1ZLT

Published - Cuzzolin, A.; Sturlese, M.; Malvacio, I.; Ciancetta, A.; Moro, S. *Molecules* **2015**, *20* (6), 9977–9993.

	3-(1-methyl-1H-pyrazol-4-yl)-N-(3-methyl-1,2-thiazol-5-yl)-5-[(3R)-piperidin-3-yl]pyrazolo[1,5-a]pyrimidin-7-amine	Mi5-3OT8
	N-(2-azanylethyl)-5-[2-[(4-morpholin-4-ylpyridin-2-yl)amino]-1,3-thiazol-5-yl]pyridine-3-carboxamide	S25-3TKI
	(5S,6R,7R,9R)-6-methoxy-5-methyl-7-(methylamino)-6,7,8,9,15,16-hexahydro-17- oxa-4b,9a,15-triaza-5,9 methanodibenzo[b,h]cyclonona[jkl]cyclopenta[e-as-indacen-14(5H)-one	Stu-1NVR
	5-(hydroxymethyl)-8-(1H-pyrrol-2-yl)[1,2,4]triazolo[4,3-a]quinolin-1(2H)-one	Tq1-2YER
	(5R,8S)-5,6,7,8-tetrahydro-13H-5,8-epoxy-4b,8a,14-triazadibenzo[b,h]cycloocta[1,2,3,4-jkl]cyclopenta[e]-as-indacene-13,15(14H)-dione	Ucm-1NVS
	2-(carbamoylamino)-5-{4-[2-(dimethylamino)ethoxy]phenyl}thiophene-3-carboxamide	Ydi-2YDI

Published - Cuzzolin, A.; Sturlese, M.; Malvacio, I.; Ciancetta, A.; Moro, S. *Molecules* **2015**, *20* (6), 9977–9993.

	5-methyl-8-(1H-pyrrol-2-yl)-2H-[1,2,4]triazolo[4,3-a]quinolin-1-one	Yex-2YEX
	5-((6-((piperidin-4-ylmethyl)amino)pyrimidin-4-yl)amino)pyrazine-2-carbonitrile	Ym7-2YM7
	1-morpholin-4-yl-2-[4-[2-[(5-pyridin-3-yl-1,3-thiazol-2-yl)amino]pyridin-4-yl]piperazin-1-yl]ethanone	07s-3TKH
	N-{5-[4-(4-methylpiperazin-1-yl)phenyl]-1H-pyrrolo[2,3-b]pyridin-3-yl}pyridine-3-carboxamide	199-1ZYS
	2-indazol-1-yl-N-(2-piperazin-1-ylphenyl)-1,3-thiazole-4-carboxamide	1a0-4HYI
	2-(6-methoxy-1-oxoisoindolin-2-yl)-N-(4-(piperazin-1-yl)pyridin-3-yl)thiazole-4-carboxamide	1am-4HYH
	5-[(1R,3S)-3-azanylcyclohexyl]-6-bromo-3-(1-methylpyrazol-4-yl)pyrazolo[1,5-a]pyrimidin-7-amine	22k-3OT3

Published - Cuzzolin, A.; Sturlese, M.; Malvacio, I.; Ciancetta, A.; Moro, S. *Molecules* **2015**, *20* (6), 9977–9993.

Conclusions

We have introduced here DockBench 1.0, a platform available free of charge that fully automates the pipeline from docking benchmarks to VS campaigns setups. Currently, DockBench 1.0 implements seven different docking software packages (including commercial and freely available ones) and provides the user with the possibility to test up to seventeen protocols. The platform has been devised with the aim to minimize the user's required expertise by overcoming the main issues related to docking benchmark procedures: The management of input/output formats and the time required in running,
merging and comparing the results arising from different software packages. To this aim, a GUI guides the user step-by-step throughout all the stages from docking protocol assessment to VS of large chemical libraries. The outputs of the different software packages are returned in a unique format and are analysed with a standardized procedure to avoid biases. The distributed computing philosophy based on GNU parallel semaphores has been integrated in the platform, thus allowing the users to speeds up the calculations while cleverly using the available resources. As validation case, we have reported on the benchmark study of 20 hChk1 structures by testing all the protocols available in the platform. DockBench 1.0 is available free of charge and can be requested at the project web page[12].

Published - Cuzzolin, A.; Sturlese, M.; Malvacio, I.; Ciancetta, A.; Moro, S. *Molecules* **2015**, *20* (6), 9977–9993.

Bibliography

1. Sotriffer, C. Methods and principles in medicinal chemistry. In Virtual Screening: Principles, Challenges, and Practical Guidelines; Wiley-VCH: Weinheim, Germany, 2011.

2. Horvath, D. A virtual screening approach applied to the search for trypanothione reductase inhibitors. J. Med. Chem. 1997, 40, 2412–2423.

3. Lill, M. Virtual screening in drug design. In In Silico Models for Drug Discovery; Kortagere, S., Ed.; Humana Press: Totowa, NJ, USA, 2013; Volume 993, pp. 1–12.

4. Wilton, D.; Willett, P.; Lawson, K.; Mullier, G. Comparison of Ranking Methods for Virtual Screening in Lead-Discovery Programs. J. Chem. Inf. Model. 2003, 43, 469–474.

5. Kitchen, D.B.; Decornez, H.; Furr, J.R.; Bajorath, J. Docking and scoring in virtual screening for drug discovery: Methods and applications. Nat. Rev. Drug Discov. 2004, 3, 935–949.

6. Houston, D.R.; Walkinshaw, M.D. Consensus Docking: Improving the Reliability of Docking in a Virtual Screening Context. J. Chem. Inf. Model. 2013, 53, 384–390.

7. Cole, J.C.; Murray, C.W.; Nissink, J.W.M.; Taylor, R.D.; Taylor, R. Comparing protein-ligand docking programs is difficult. Proteins Struct. Funct. Bioinform. 2005, 60, 325–332.

8. Ciancetta, A.; Cuzzolin, A.; Moro, S. Alternative Quality Assessment Strategy to Compare Performances of GPCR-Ligand Docking Protocols: The Human Adenosine A2A Receptor as a Case Study. J. Chem. Inf. Model. 2014, 54, 2243–2254.

9. Sanchez, Y.; Wong, C.; Thoma, R.S.; Richman, R.; Wu, Z.; Piwnica-Worms, H.; Elledge, S.J. Conservation of the Chk1 checkpoint pathway in mammals: Linkage of DNA damage to Cdk regulation through Cdc25. Science 1997, 277, 1497–1501.

10. Bartek, J.; Lukas, J. Chk1 and Chk2 kinases in checkpoint control and cancer. Cancer Cell 2003, 3, 421–429.

11. Converso, A.; Hartingh, T.; Garbaccio, R.M.; Tasber, E.; Rickert, K.; Fraley, M.E.; Yan, Y.; Kreatsoulas, C.; Stirdivant, S.; Drakas, B.; et al. Development of thioquinazolinones, allosteric Chk1 kinase inhibitors. Bioorg. Med. Chem. Lett. 2009, 19, 1240–1244.

12. MMs DockBench. Available online:

http://mms.dsfarm.unipd.it/mmsdockbench.html (accessed on 25 May 2015).

13. Berman, H.M. The Protein Data Bank. Nucleic Acids Res. 2000, 28, 235–242.

14. O'Boyle, N.M.; Banck, M.; James, C.A.; Morley, C.; Vandermeersch, T.; Hutchison, G.R. Open Babel: An open chemical toolbox. J. Cheminform. 2011, 3, doi:10.1186/1758-2946-3-33.

15. Morris, G.M.; Huey, R.; Lindstrom, W.; Sanner, M.F.; Belew, R.K.; Goodsell, D.S.; Olson, A.J. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. J. Comput. Chem. 2009, 30, 2785–2791.

16. Trott, O.; Olson, A.J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J. Comput. Chem. 2010, 31, 455–461.

17. 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.; et al. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. J. Med. Chem. 2004, 47, 1739–1749.

18. 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. J. Med. Chem. 2004, 47, 1750–1759.

19. Verdonk, M.L.; Cole, J.C.; Hartshorn, M.J.; Murray, C.W.; Taylor, R.D. Improved protein-ligand docking using GOLD. Proteins Struct. Funct. Bioinform. 2003, 52, 609–623.

20. Korb, O.; Stützle, T.; Exner, T.E. Plants: Application of ant colony optimization to structure-based drug design. In Ant Colony Optimization and Swarm Intelligence; Dorigo, M., Gambardella, L.M., Birattari, M., Martinoli, A., Poli, R., Stützle, T., Eds.; Springer Berlin Heidelberg: Berlin, Heidelberg, Germany, 2006; Volume 4150, pp. 247–258.

21. Korb, O.; Stützle, T.; Exner, T.E. Empirical Scoring Functions for Advanced Protein-Ligand Docking with PLANTS. J. Chem. Inf. Model. 2009, 49, 84–96.

22. Molecular Operating Environment (MOE), 2014.09. Chemical Computing Group Inc.: 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2015.

23. Ruiz-Carmona, S.; Alvarez-Garcia, D.; Foloppe, N.; Garmendia-Doval, A.B.;

Juhos, S.; Schmidtke, P.; Barril, X.; Hubbard, R.E.; Morley, S.D. rDock: A Fast, Versatile and Open Source Program for Docking Ligands to Proteins and Nucleic Acids. PLoS Comput. Biol. 2014, 10, e1003571.

24. Tange, O. GNU Parallel—The Command-Line Power Tool. Login USENIX Mag. 2015, 36, 42–47.

25. GOLD suite, version 5.2; Cambridge Crystallographic Data Centre: 12 Union Road, Cambridge CB2 1EZ, UK.

26. Halgren, T.A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. J. Comput. Chem. 1996, 17, 490–519.

27. Dudkin, V.Y.; Rickert, K.; Kreatsoulas, C.; Wang, C.; Arrington, K.L.; Fraley, M.E.; Hartman, G.D.; Yan, Y.; Ikuta, M.; Stirdivant, S.M.; et al. Pyridyl aminothiazoles as potent inhibitors of Chk1 with slow dissociation rates. Bioorg. Med. Chem. Lett. 2012, 22, 2609–2612.

28. Stavenger, R.A.; Zhao, B.; Zhou, B.-B.S.; Brown, M.J.; Lee, D.; Holt, D.A. Pyrrolo[2,3-b]pyridines Inhibit the Checkpoint Kinase Chk1. Available online: http://www.rcsb.org/pdb/static.do?p=general_information/about_pdb/policies_references.html (accessed on 28 May 2015).

29. Huang, X.; Cheng, C.C.; Fischmann, T.O.; Duca, J.S.; Richards, M.; Tadikonda, P.K.; Reddy, P.A.; Zhao, L.; Arshad Siddiqui, M.; Parry, D.; et al. Structure-based design and optimization of 2-aminothiazole-4-carboxamide as a new class of CHK1 inhibitors. Bioorg. Med. Chem. Lett. 2013, 23, 2590–2594.

30. Labroli, M.; Paruch, K.; Dwyer, M.P.; Alvarez, C.; Keertikar, K.; Poker, C.; Rossman, R.; Duca, J.S.; Fischmann, T.O.; Madison, V.; et al. Discovery of pyrazolo[1,5-a]pyrimidine-based CHK1 inhibitors: A template-based approach—Part 2. Bioorg. Med. Chem. Lett. 2011, 21, 471–474.

31. Huang, S.; Garbaccio, R.M.; Fraley, M.E.; Steen, J.; Kreatsoulas, C.; Hartman, G.; Stirdivant, S.; Drakas, B.; Rickert, K.; Walsh, E.; et al. Development of 6-substituted indolylquinolinones as potent Chek1 kinase inhibitors. Bioorg. Med. Chem. Lett. 2006, 16, 5907–5912.

32. Brnardic, E.J.; Garbaccio, R.M.; Fraley, M.E.; Tasber, E.S.; Steen, J.T.; Arrington, K.L.; Dudkin, V.Y.; Hartman, G.D.; Stirdivant, S.M.; Drakas, B.A.; et al. Optimization of a pyrazoloquinolinone class of Chk1 kinase inhibitors. Bioorg. Med. Chem. Lett. 2007, 17, 5989–5994.

33. Zhao, L.; Zhang, Y.; Dai, C.; Guzi, T.; Wiswell, D.; Seghezzi, W.; Parry, D.;

Fischmann, T.; Siddiqui, M.A. Design, synthesis and SAR of thienopyridines as potent CHK1 inhibitors. Bioorg. Med. Chem. Lett. 2010, 20, 7216–7221.

34. Lee, C.C.; Ng, K.; Wan, Y.; Gray, N.; Spraggon, G. Crystal Structure of Chk1 Complexed with a Hymenaldisine Analog. Available online: http://www.rcsb.org/pdb/static.do?p=general_information/about_pdb/policies_references.html (accessed on 28 May 2015).

35. Dwyer, M.P.; Paruch, K.; Labroli, M.; Alvarez, C.; Keertikar, K.M.; Poker, C.; Rossman, R.; Fischmann, T.O.; Duca, J.S.; Madison, V.; et al. Discovery of pyrazolo[1,5-a]pyrimidine-based CHK1 inhibitors: A template-based approach—Part 1. Bioorg. Med. Chem. Lett. 2011, 21, 467–470.

36. Zhao, B. Structural Basis for Chk1 Inhibition by UCN-01. J. Biol. Chem. 2002, 277, 46609–46615.

37. Oza, V.; Ashwell, S.; Brassil, P.; Breed, J.; Ezhuthachan, J.; Deng, C.; Grondine, M.; Horn, C.; Liu, D.; Lyne, P.; et al. Synthesis and evaluation of triazolones as checkpoint kinase 1 inhibitors. Bioorg. Med. Chem. Lett. 2012, 22, 2330–2337.

38. Oza, V.; Ashwell, S.; Almeida, L.; Brassil, P.; Breed, J.; Deng, C.; Gero, T.; Grondine, M.; Horn, C.; Ioannidis, S.; et al. Discovery of Checkpoint Kinase Inhibitor (S)-5-(3-Fluorophenyl)-N- (piperidin-3-yl)-3-ureidothiophene-2-carboxamide (AZD7762) by Structure-Based Design and Optimization of Thiophenecarboxamide Ureas. J. Med. Chem. 2012, 55, 5130–5142.

39. Reader, J.C.; Matthews, T.P.; Klair, S.; Cheung, K.M.J.; Scanlon, J.; Proisy, N.; Addison, G.; Ellard, J.; Piton, N.; Taylor, S.; et al. Structure-Guided Evolution of Potent and Selective CHK1 Inhibitors through Scaffold Morphing. J. Med. Chem. 2011, 54, 8328–8342.

40. Labute, P. Protonate3D: Assignment of ionization states and hydrogen coordinates to macromolecular structures. Proteins 2009, 75, 187–205.

41. Wang, J.; Cieplak, P.; Kollman, P.A. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? J. Comput. Chem. 2000, 21, 1049–1074.

42. Stewart, J.J.P. Optimization of parameters for semiempirical methods I. Method. J. Comput. Chem. 1989, 10, 209–220.

43. Stewart, J.J.P. Optimization of parameters for semiempirical methods II. Applications. J. Comput. Chem. 1989, 10, 221–264.

Published - Cuzzolin, A.; Sturlese, M.; Malvacio, I.; Ciancetta, A.; Moro, S. *Molecules* **2015**, *20* (6), 9977–9993.

3.4 Deciphering the Complexity of Ligand-protein Recognition Pathways using Supervised Molecular Dynamics (SuMD) Simulations.

Alberto Cuzzolin, Mattia Sturlese, Giuseppe Deganutti, Veronica Salmaso, Davide Sabbadin, Antonella Ciancetta and Stefano Moro*

Abstract

Molecular recognition is a crucial issue in interpreting the mechanism of known active substances as well as in the development of novel active candidates, since both thermodynamic and kinetic aspects greatly affect the understanding of ligand-mediated signal transmission in living organisms or whether a chemical compound can be transformed in a drug candidate. The physicochemical bases governing the optimization of thermodynamic aspects of ligand binding are relatively well understood, but they remain still poorly comprehend for binding kinetics. Unfortunately, simulating this binding process is still a challenging task because it requires classical MD experiments in a long microsecond time scale that is affordable only with a high-level computational capacity. In order to overcome this limiting factor, we have recently implemented an alternative MD approach, named supervised molecular dynamics (SuMD) specifically in the field of G protein-coupled receptors (GPCRs). SuMD enables the investigation of ligand-receptor binding events independently from the starting position, chemical structure of the ligand, and also from its receptor binding affinity.

In this Article, we would like to present an extension of SuMD application domain including different types of proteins compared to GPCRs. In particular, we decided to deeply analyze the ligand-protein recognition pathways of six different case studies that we grouped into two different classes: globular and membrane proteins. Moreover, we would like to introduce the SuMD-Analyzer tool that we have specifically implemented to help the user in the analysis of the SuMD trajectories. Finally, we will emphasize the limit of SuMD applicability domain as well as its strengths in analyzing the complexity of ligand-protein recognition pathways.

Introduction

The essential features of ligand-protein interaction are very often summarized under the expression "molecular recognition" incorporating both thermodynamic aspects (quantified by the K_d , the equilibrium dissociation constant) and kinetic aspects of ligand binding (reflected by the rate constants k_{on} and k_{off}). Consequently, molecular recognition is thus a crucial issue in interpreting the mechanism of known active substances as well as in the development of novel active candidates, since

Submitted - Cuzzolin, A.; Sturlese, M.; Deganutti, G.; Salmaso, V.; Sabbadin, D.; Ciancetta, A.; Moro, S. *J. Med. Chem.*

both thermodynamic and kinetic aspects greatly affect the understanding of ligand-mediated signal transmission in living organisms or whether a chemical compound can be transformed in a drug candidate[1].

The physico-chemical bases governing the optimization of thermodynamic aspects of ligand binding are relatively well understood but, unluckily, they remain still poorly comprehend for binding kinetics. In fact, the equilibrium dissociation constant value depends on the free energy difference between the ligand-protein bound and unbound states, both of which are chemically stable and generally experimentally observable. On the contrary, k_{on} and k_{off} rate constants depend on the height of the free energy barrier separating those states and, in particular, the highest free energy barrier defined as transition state characterized only by a fleeting existence[2]. Consequently, the major challenge in the optimization of the kinetics parameters is the complexity in characterizing all plausible approaching pathways of the ligand to its protein. In fact, different approaching pathways can be characterized by different metastable intermediate states (referred also as meta-binding sites)[3] connected to each other, and to the final bound state, by different transition states. Understanding the molecular interactions between ligand and protein during the approaching pathways is thus central to the deep understanding and to the rational control of ligand binding kinetics. Even though experimental techniques for measuring the kinetic parameters of ligand binding have existed for decades, all of them only provide indirect evidence about transient structures visited along a ligand-binding pathway[2]. Alternatively, computational methods, and in particular molecular dynamics (MD) simulations, can provide detailed structural information on metastable intermediate states (meta-binding sites) and transition states at the atomistic level of detail[4]. Due to increases in computational power, it has recently become possible to simulate the full process of spontaneous ligand-protein association which typically occurs on the microsecond timescale, providing direct access to detailed information on binding mechanisms that have been difficult to access experimentally[4,5]. Unfortunately, simulating this binding process is still a challenging task because it requires classical MD experiments in a long microsecond time scale that is affordable only with a high-level computational capacity. However, the probability of reproduce ligand- protein binding or unbinding event on an accessible timescale can be enhanced through the introduction of biased potentials that facilitate the crossing of energy barriers or the application of external forces on the ligand, respectively[6]. An alternative strategy that does not require the introduction of biases or external forces and enables to explore the ligand-protein approaching path in nanosecond simulation time scale has been recently proposed by us specifically in the field of G protein-coupled receptors (GPCRs)[7,8] The "supervised molecular dynamics" (SuMD) approach exploit a tabu-like algorithm to monitor the distance between the center of masses of the ligand atoms and the

Submitted - Cuzzolin, A.; Sturlese, M.; Deganutti, G.; Salmaso, V.; Sabbadin, D.; Ciancetta, A.; Moro, S. *J. Med. Chem.*

protein binding site in short (600 ps) standard MD simulations (Figure 1, left panel). According to this strategy, an arbitrary number of distance points is collected “on the flight” at regular intervals and fitted into a linear function *f(x)=mx*. If the slope (*m*) is negative, the ligand-receptor distance is likely to be shortened and the simulation is restarted from the last set of coordinates. Otherwise, the simulation is restored from the original set of coordinates and started over. The supervision is repeated until the ligand-receptor distance is less than 5 Å. The results of a SuMD simulation are displayed in a graph reporting the interaction energy toward the distance between the ligand and the binding site (Figure 1, right panel). We have recently applied the SuMD approach to interpret at the molecular level: *i)* the binding of different antagonists at the human A_{2A} adenosine receptor (hA_{2A} AR) by detecting and characterizing a possible energetically stable meta-binding site[7] , *ii)* the binding of the natural agonist adenosine at the hA_{2A} AR by detecting and characterizing a possible energetically stable meta-binding site[9], *iii)* the positive allosteric modulation mediated by LUF6000 toward the human A_3 adenosine receptor (hA_3 AR) by suggesting at least two possible mechanisms to explain the available experimental data[10], and *iv)* the binding of different ligands at the human P2Y12 receptor by detecting and characterizing again possible energetically stable meta-binding site[11].

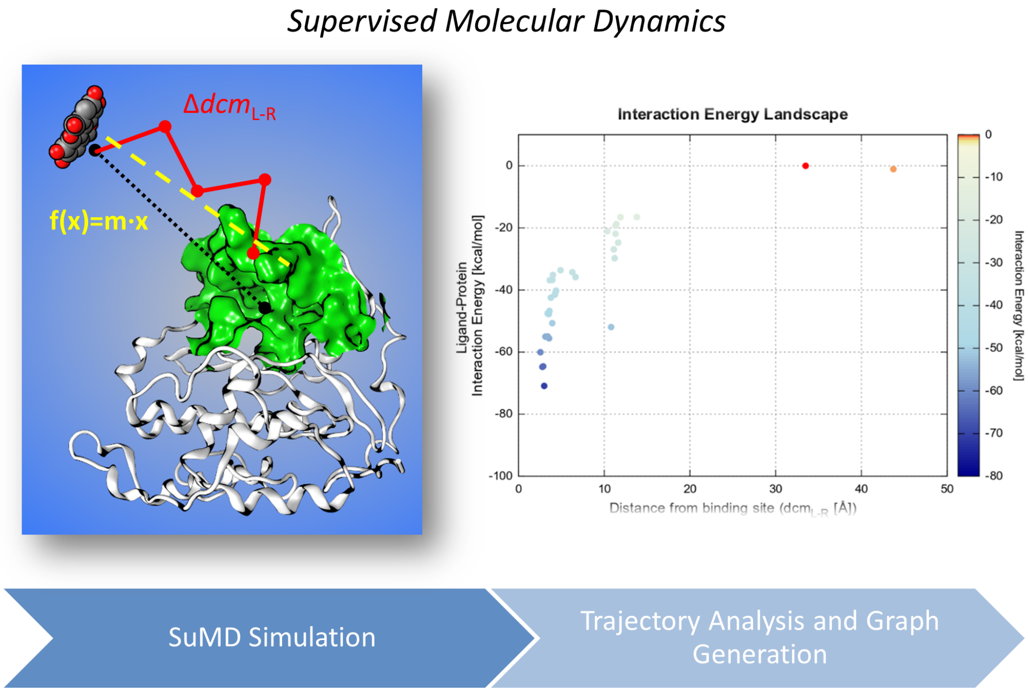

Supervised Molecular Dynamics
Δdcm L-R
f(x)=m·x
Interaction Energy Landscape
Ligand-Protein Interaction Energy [kcal/mol]
0
-20
-40
-60
-80
-100
0
10
20
30
40
50
Distance from binding site (dcm L-R [Å])
Interaction Energy [kcal/mol]
0
-10
-20
-30
-40
-50
-60
-70
-80
SuMD Simulation
Trajectory Analysis and Graph Generation

Figure 1 - a) Schematic representation of Supervised Molecular Dynamics (SuMD) algorithm (left) and the outcoming ligand−protein interaction energy landscape. Interaction Energy values: kcal mol^{-1}

In the present work, we would like to present an extension of SuMD application domain including different types of proteins compared to GPCRs. In particular, we decided to deeply analyzed the ligand-protein recognition pathways of six different

Submitted - Cuzzolin, A.; Sturlese, M.; Deganutti, G.; Salmaso, V.; Sabbadin, D.; Ciancetta, A.; Moro, S. *J. Med. Chem.*

case studies that we grouped into two different classes of proteins : globular and membrane proteins, as summarized in Table 1. Moreover, we would like to introduce the SuMD-Analyzer tool that we have specifically implemented to help, also a non expert user, in the analysis of the SuMD trajectories.

Table 1 - Structural summary of the selected ligand-protein PDB ID are reported

Globular System

PDB	**Protein**	**Ligand**	**Resolution [Å]**	**Affinity**	**Ligand MW**	**Ref.**
2ZJW	CK2	Ellagic Acid	2.40	K_i=0.04 µM	302.197	41
13GS	GSTP1-1	SASP	1.90	K_i=24 µM	398.39	44
4K7I	PRDX5	Benzen-1,2-diol	2.25	K_i=1500µM	110.11	45
2VDB	HSA	(S)-naproxen	2.25	K_i=1.2-1.8µM[1]	230.25	49,58

Transmembrane Systems

PDB	**Protein**	**Ligand**	**Resolution [Å]**	**Affinity**	**Ligand MW**	**Ref.**
3GWW	LeuT	(S)-dluoxetine	2.46	IC_{50}=355mM	345.79	51
2YDV	hA$_{2A}$AR	NECA	2.60	K_i=13.8nM	308.29	55

Materials and Methods

General.

All computations were performed on a hybrid CPU/GPU cluster. MD simulations were carried out with the ACEMD engine[12] on a GPU cluster equipped with four NVIDIA GTX 580, two NVIDIA GTX 680, three NVIDIA GTX 780, and four NVIDIA GTX 980. Before running SuMD simulations, the following preliminary phases were carried out: *i)* protein-ligand system preparation; *ii)* ligand parameterization; *iii)* solvated system setup and equilibration. Two different protocols based on AMBER12[13]/General Amber Force Filed (GAFF)[14] and the CHARM27[15]/CHARMM General Force Field (CGenFF), force fields combinations were adopted for globular and transmembrane systems, respectively[16,17].

Systems Preparation.

Protein-ligand complexes were retrieved from the RCSB PDB database[18]. Proteins structures were prepared with the protein preparation tool as implemented in MOE[19]: hydrogen atoms were added to the complex and appropriate ionization states were assigned by means of the Protonate-3D tool[20]. Missing atoms in protein side chains were built according to either the AMBER12[13] or the CHARM27[15] force field topology. Missing loops were modeled by the default homology modelling protocol implemented in the MOE protein preparation tool. Non- natural N-terminal and C-terminal were capped to mimic the previous residue. For each considered system, the conformer with highest occupancy was selected whenever available. To avoid protein-ligand long range interactions in the starting geometry, the ligand was then moved at least 15 Å from any protein atom.

Ligand Parametrization.

Globular systems. For the MD simulations based on the AMBER12 force field[13], the ligands were subjected to two energy minimization steps with MOPAC2012[21] using PM6 method[22] and Gaussian 09[23] (basis set: HF/6-31G*). After geometry minimization, ligand parameters were derived with GAFF[14] as implemented in ambertools2014[13] by using antechamber and parmchk tools. RESP partial charges where calculated with Gaussian 09[23] following the procedure suggested by antechamber.

Transmembrane systems. For the MD simulation based on the CHARMM27 force field[24], initial parameters for the ligands were retrieved from the paramchem service and subsequently optimized consistently to CGenFF[16,25] at the MP2/6-31G* level of theory[26] by using Gaussian 09[23] and the Force Field Toolkit[27] implemented in the VMD engine[28].

Solvated System Setup and Equilibration

Globular Systems. Protein-ligand complexes were assembled with tleap tool using AMBER14SB[29] as force field for the protein[29]. The systems were explicitly solvated by a cubic water box with cell borders placed at least 12 Å away from any protein or ligand atom using TIP3P as water model[30]. To neutralize the total charge Na^+/Cl^- counter-ions were added to a final salt concentration of 0.150 M. The systems were energy minimized by 2000 step with conjugate-gradient method, then 50000 step of NVE (100 ps) followed by 1 ns of NPT simulation were carried out, both using 2 fs as time step and applying an harmonic positional constrain on protein and ligand atoms gradually reduced with a scaling factor of 0.1. Pressure was maintained at 1 atm using a Berendsen barostat[31].

The Langevin thermostat was set with a low damping constant of 1 ps^{-1} [32]. Bond lengths involving hydrogen atoms were constrained using the M-SHAKE algorithm[33]. The MD productive runs were conducted in a NVT ensemble. Long-range Coulomb interactions were handled using the particle mesh Ewald summation method (PME) setting the mesh spacing to 1.0 Å[34]. A non-bonded cut-off distance of 9 Å with a switching distance of 7.5 Å was used.

Transmembrane Systems. Transmembrane proteins were embedded in a 1-palmitoyl-2- oleoyl-snglycero-3-phosphocholine (POPC) lipid bilayer according to the suggested orientation reported in the Orientations of Proteins in Membranes (OPM) database[35]. The systems were solvated with TIP3P[30] water using the program Solvate 1.0[36] and neutralized by Na^+/Cl^- counterions to a final concentration of 0.154 M. The systems were then equilibrated through a two steps procedure: in the first stage, after 2000 cycles of conjugate-gradient minimization algorithm (in order to reduce steric clashes produced by the system manual setting), 10 ns of MD simulation were performed in the NPT ensemble, restraining ligand and protein atoms by a force constant of 1 Kcal mol^{-1} Å^{-2}. The temperature was maintained at 298 K using a Langevin thermostat with a low damping constant of 1 ps^{-1} [32] pressure was maintained at 1 atm using a Berendsen barostat[31], bond lengths involving hydrogen atoms were constrained using the M-SHAKE algorithm[33] with an integration timestep of 2 fs. In the second stage, once water molecules diffused inside the protein cavity and the lipid bilayer reached equilibrium, the force constant was gradually reduced to 0.1 Kcal mol^{-1} Å^{-2} for the next 10 ns of MD simulation.

Supervised Molecular Dynamics (SuMD)

SuMD is a command line tool written in python, tcl, and bash that operates the supervision of MD trajectories according to the algorithm that has been previously described[7]. The program exploits Visual Molecular Dynamics (VMD) and Gnuplot functionalities[28,37]. In its current implementation, SuMD is interfaced with the ACEMD[12] engine and supports AMBER and CHARMM force fields.

SuMD Input files. SuMD requires a configuration file (selection.dat, Figure S1) organized in three major sections containing information about: *i)* the system; *ii)* the supervision procedure; and *iii)* the simulation settings. In the system settings section, the following details about the molecular system need to be provided: *i)* the pdb file name containing the starting coordinates; *ii)* the 3-letter code name of the ligand; and *iii)* the residues describing the target binding site. In the supervision settings section, the following values are declared: *i)* the slope threshold (default value: 0); and *ii)* the number of maximum consecutive failed steps (default value: 33) to stop the simulation. In the simulation settings section, the following details must be specified: *i)* the force field to use; *ii)* the parameter file; *iii)* the GPU device ID to which the calculation will be addressed. In this section, a Boolean operator manages the introduction of a randomization step that varies the position of the ligand through a 600 ps of non-supervised MD simulation. In the same directory where SuMD is launched, a file containing the cell dimension as well as a parameter file (prmtop/psf with the same name of the pdb) must also be provided.

SuMD Main Code. The workflow of the SuMD main code is reported in Figure 2A. As depicted, at the beginning of the simulation SuMD detects the atoms that identify the ligand and the target binding site, to define the distance between their mass centers $dcm_{(L-R)}$ that will be monitored. Then, a series of 600 ps classical MD simulations are performed. After each simulation, five $dcm_{(L-R)}$ distance points are collected at regular intervals of 75ps. Using these points, the slope value (*m*) is derived by a linear fitting. As previously described, if the resulting slope *m* is negative or below the user selected threshold (*i.e.* the distance $dcm_{(L-R)}$ is decreasing), the next simulation step starts from the last set of coordinates produced, otherwise the simulation is restarted by randomly assigning the atomic velocities. To avoid problematic starting geometries (*i.e.* geometries prone to lead to dead-end pathway), in the first simulation step, SuMD supervises the distance $dcm_{(L-R)}$ with a maximum threshold of 31 failed attempts (Preliminary Run). In the case this threshold is reached, SuMD callbacks a randomization process on the set of coordinates supplied by the user by a classical 600 ps MD simulation.

During the following steps, the simulations are perpetuated under the supervision rules. In particular, the first time a slope value below the threshold is recorded, the program enters the so- called "SuMD Run". When the distance $dcm_{(L-R)}$ drops below 5 Å the supervision is disabled and the simulation proceeds though a classical MD simulation. At the end of the simulation, only the productive steps are saved, chronologically numbered and stored in a separate directory.

SuMD log file. At each SuMD simulation step, a log file (Figure 1S) is updated collecting information about: *i)* the step number; *ii)* the $dcm_{(L-R)}$ distance; *iii)* the slope value (*m*); *iv)* the electrostatic and van der Waals potential energy contributions of the ligand-receptor interaction energy (IE). A counter keeps trace on how many times each SuMD step has been attempted. Furthermore, three counters corresponding to the $dcm_{(L-R)}$ distance ranges 0-2 Å, 2-5 Å, and 5-9 Å are reported. These distances monitors how many times the binding site is approached, *i.e*. how often the $dcm_{(L-R)}$ distance lies below the long-range interaction cutoff. These counters determine the program termination criteria (see following section) and, according to the binding site definition supplied by the user, they might represent: the target binding site, its neighbors, and putative allosteric/meta-binding sites, respectively.

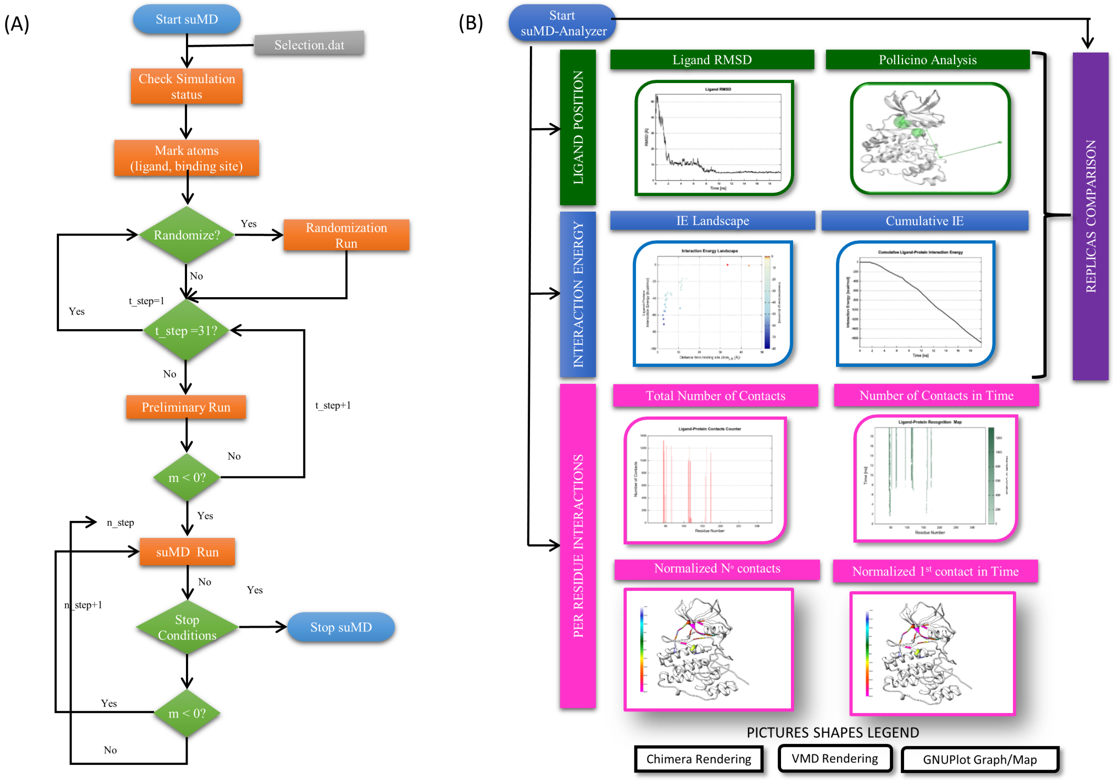


Figure 2 – (A) Workflow of the SuMD main code **(B)** Workflow of the SuMD-Analyzer tool.

SuMD Termination Criteria.

A SuMD simulation is terminated when one of the following criteria is satisfied: *i)* no negative slope (*m*) values are recorded for a user-selected number of steps (33 consecutive steps by default); *ii)* one of the distance counters reaches a maximum value of 19 (*i.e.* the dcm$_{(L\text{-}R)}$ lies in the same region for 11.5 nonconsecutive nanoseconds).

SuMD-Analyzer Tool

The SuMD-Analyzer is a standalone tool written in python, tcl, and bash to analyze the SuMD trajectories (Figure 2B). The tool is integrated with VMD[28] and UCSF Chimera[38] for the graphical visualization and exploits Wordom[39] and Gnuplot[37] functionalities. The provided analyses cross over four different aspects: *i)* the ligand position, *ii)* the IE, *iii)* the per residue interactions, and *iv)* the replicas comparison.

When the SuMD-Analyzer is launched, the trajectories produced by SuMD are merged and aligned to the starting reference structure using the RMSD tool in VMD by using alpha-carbon atoms for the superposition. The merged trajectory is subjected to a striding procedure picking one frame every 5 through the VMD animate module.

Ligand Position. Two analyses follow the coordinates explored by the ligand during the SuMD trajectory (Figure 2B, green boxes): *i)* the Root Mean Square Deviation (RMSD); and *ii)* the so-called "Pollicino Analysis". If a reference complex structure is available, the RMSD between the ligand and the reference coordinates supplied is computed along the trajectory. The calculation is performed on the heavy atoms of the ligand using the measure rmsd function implemented in VMD and the data obtained are plotted against the time using Gnuplot[37] (Figure 2B, left green box).

The Pollicino Analysis is a representation that graphically renders the recognition pathway explored by the ligand. At the end of each SuMD step, the coordinates of the ligand mass center are collected and clustered according to their $\mathrm{dcm}_{(L\text{-}R)}$ using a threshold value of 2 Å. The coordinates belonging to the same cluster are averaged and represented by a sphere which radius depends on the population of the cluster. Arrows indicate the chronological order onto which the regions where the sphere reside are approached by the ligand mass center (Figure 2B, right green 22 box).

Interaction Energy. The ligand-protein interaction is analyzed by means of the mdenergy function embedded into VMD. The electrostatic and van der Waals contributions to the potential energy are calculated for each frame and summed to obtain the total IE. With this value, two graphs are derived (Figure 2B, blue boxes): *i)* the "Interaction Energy Landscape", and *ii)* the "Cumulative Interaction Energy". The former chart displays the total IE profile with respect to the $\mathrm{dcm}_{(L\text{-}R)}$ through a colorimetric scale representing the IE value. Each point displayed in the chart represents the last position of the corresponding SuMD step (Figure 2B, left blue box). The latter plot shows the cumulative sum of the total IE values for each frame against the time. Therefore, each point is the sum of all previous IE values. Changes in the observed trend highlight how the variation of ligand conformation/position affects the IE (Figure 2B, right blue box).

Per Residue Interactions.A further set of analyses was developed to highlight the most important residues involved in the ligand recognition pathway (Figure 2B, upper magenta boxes): *i)* the "Protein-ligand Contacts Count", and the *ii)* "Ligand-Protein Recognition Map". In the first graph (Figure 2B, upper left magenta box), the residues more frequently approached by the ligand during the trajectory are reported and for each residue the total number of established contacts is rendered as histograms. In this representation, at each SuMD frame only the residues lying within a distance of 4 Å from any ligand atoms are considered. In the second graph (Figure 2B, upper right magenta box), the residue approached by the ligand are depicted with respect to the simulation time. In particular, each dot in the map represents a trajectory frame colored according to the total number of contacts the ligand has established with a particular residue. White dots means that, at the considered frame, the residue atoms are farther then 4 Å from ligand atoms, while green dots correspond to a contact event and the sum of the contact is coded by the light-green to dark-green scale.

To support the user in the topological localization of the residues mainly interacting with the ligand during the trajectory, molecular 3D representations of the protein are automatically set using USF Chimera[38] (Figure 2B, lower magenta boxes). In particular, the number of ligand- protein contacts is normalized and stored into the B-factor field of the involved residue in the protein pdb file. In the protein 3D representation "Chimera_count" (Figure 2B, lower left magenta box) the ribbons are colored according the so-derived B-factor values. A similar representation, "Chimera_time" (Figure 2B, lower right magenta box), is available with the color code (blue-to-violet) reflecting the chronological order onto which the residues have been approached by the ligand for the first time.

Replicas Analysis. The "Replicas Analysis" (Figure 2B, violet box) is a manager that compares the molecular recognition event occurred in different SuMD replicas. The manager extracts from each trajectory the data regarding the ligand position and the IE, merges the data for each analysis in graphs colored according the replica number to better appreciate the differences.

Results and Discussion

Case Studies Selection

An already anticipated, in this work SuMD applicability domain has been extended using six different case studies, grouped into two major protein classes: *i)* globular systems, and *ii)* transmembrane systems (as summarized in Table 1). Specifically, considering the globular proteins we selected: *a)* the human Caseine Kinase 2 (CK2) in complex with Ellagic acid; *b)* the P1-1 isoform of Glutathione S-transferase (GSTP1-1) in complex with Sulphasalazine (2-hydroxy-(5-{[4-(2-pyridinylamino)sulfonyl]phenyl}azo) benzoic acid, SASP); *c)* the human Peroxiredoxin 5 (PRDX5) in complex with a benzen-1,2-diol; and *d)* the human Serum Albumin (HSA) in complex with (S)-naproxen. Considering the membrane proteins, we selected: *a)* the Leucine transporter (LeuT) from *Aquifex aeolicus* in complex with (S)-fluoxetine; and *b)* the human Adenosine A Receptor (hA$_{2A}$ AR) in complex with the synthetic agonist 5'-N-Ethylcarboxamidoadenosine (NECA). An overview of the structural features of the considered ligand-protein complex is reported in Figure 3 and briefly described in the following.

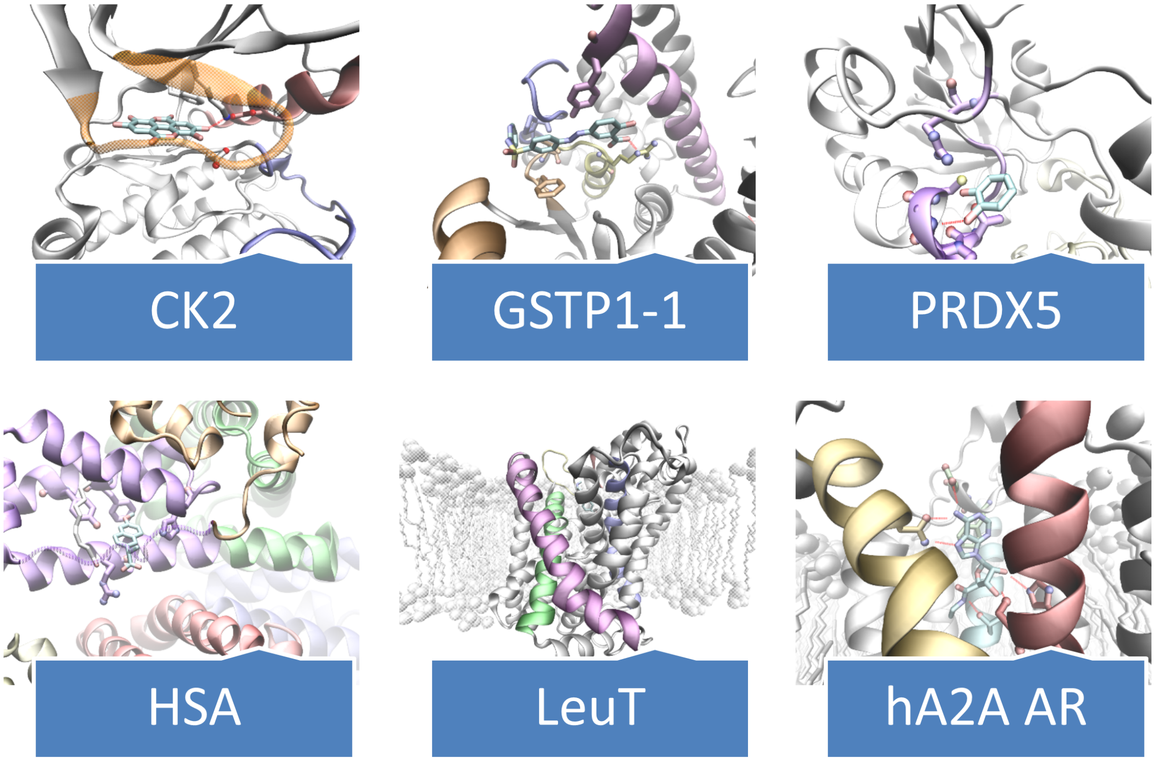


Figure 3 – Overview of the X-ray protein-ligand complexes used as validation cases. On the top: Acid Ellagic-CK2, SASP-GSTP1-1, Benzen-1,2-diol-PDRX5; On the bottom: (S)-naproxen-HAS, (S)-fluoxetin-LeuT, NECA-$hA_{2A}AR$.

CK2 is a ubiquitous and constitutively active serine/threonine kinase (PK) that phosphorylates more than 300 substrates. It is involved in the regulation of numerous cellular process such as cycle progression, apoptosis, transcription and viral infection[40]. The catalytic alpha subunit is composed by two lobes connected by a small loop called “hinge region”. The N- terminal lobe presents five β-strands and the α-helix C involved in the substrate recognition, whereas the C-terminal lobe is composed of α-helices. All PKs present a glycine rich loop (P-loop), an activation loop, and a catalytic loop[40]. The X-ray complex highlights that the inhibitor binds to Lys49, Ser51and His160 as shown in Figure 3A[41].

Glutathione S-transferases (GSTs) are homodimeric phase II detoxification enzymes, active in the bioconjugation of glutathione (GSH) to a wide range of both endogenous and exogenous molecules. The catalytic region of GSTs is topologically subdivided in two different site: *i)* the G-site, selective for GSH recognition and highly conserved crosswise GSTs isoforms, and *ii)* the H-site, less conserved and responsible for the binding of electrophilic molecules[42]. Isoform P1-1 probably represents the most studied GST and has been related to the development of tumors resistance towards numerous anti-cancer drugs[43]. SASP, which is able to inhibit GSTs without acting as a co-substrate for the conjugation reaction with GSH, has been co-crystallized with GSTP1-1 and represents a starting point for structure-based design of new anticancer drugs[44]. The X-ray complex (Figure 3B) highlights that the inhibitor binds to a hydrophobic pocket formed by Phe8, Val10, Val35, Ile104 and Tyr108 side-chains. SASP phenyl ring and salicylic acid moiety are engaged in π−π stacking interactions with the aromatic side-chain of Phe8 and Tyr108, respectively, while the ligand carboxylate is involved in an electrostatic interaction with the Arg13 side chain.

To extend the SuMD capabilities on low affinity ligand we selected the recently solved structure of PRDX5 in complex with a benzen-1,2-diol[45]. PRDX5 belongs to the ubiquitary peroxiredoxin family which role relies on the hydrogen peroxide and alkyl hydroperoxides reduction. PRDX5 plays a remarkable role in post-ischemic inflammations in the brain[46,47].

The catechol was identified by a fragment based screening and the dissociation constant was estimated in the millimolar range (K_d=1.5 +/- 0.5mM). More interestingly, the system was extensively characterized by NMR spectroscopy both with structure-based experiments and ligand-based experiments, resulting in a solid model system for a low-affinity binding event[45]. In the X-ray complex (Figure 3C) the catechol ring is localized to the N-terminus of the second helix establishing a hydrogen-bond network with the backbone nitrogen of Gly46 and Cys47 residues. The sidechain of Arg127 is oriented towards the hydroxyl moiety and contributes to the binding with an additional hydrogen bond. Similarly, the thiol group of Cys47 is faced to the catechol. The Pro40, Leu116 and Phe120 establish hydrophobic interactions with the aromatic ring.

The Human Serum Albumin (HSA) is a deeply investigated protein for its ability in bind a wide range of different molecules in human plasma. (S)-naproxen strongly binds HSA and more interestingly in different sites depending on the presence of other small molecules (e.g. hormones, xenobiotic, fatty acids)[48,49]. The only structure available for this complex was obtained in presence of decanoic acid driving the accommodation of the naproxen molecule in the IB site, a vast and hydrophobic pocket where a multitude of different ligand can be hosted[49]. In the IB site (S)-naproxen inserts its naphthalene scaffold within the hydrophobic pocket and interacts directly with the aliphatic tail of decanoic acid and the residues Ile142, Phe157 and Tyr161 (Figure 3D). The carboxylic group is partially exposed to the solvent but is surrounded by several charged residue forming the entrance of the pocket: Arg145, Lys 190 and in particular Arg186.

Neurotransmitter sodium symporter (NSS) family includes the human serotonin transporter (SERT), norepinephrine transporter (NET) and dopamine transporter (DAT)[50]. To date, there is a lack of focused information about the structure of these important therapeutic targets. In recent past, the crystallographic structure of the LeuT from Aquifex aeolicus (a NSS family member) has been disclosed with the aim of better understand the basis for selective serotonin re-uptake inhibitors (SSRIs) activity towards serotonin transporters[51]. LeuT-(S)-fluoxetine X-ray complex (Figure 3E) highlights hydrophobic contacts between the inhibitor and Leu29, Arg30, Tyr108 and Phe253 side chains. (S)-fluoxetine secondary amino group points towards the extracellular space and engage Asp401 in an electrostatic interaction, while the extracellular gate is locked by the salt bridge between Asp404 and Arg30.

Moving to the last key study, adenosine receptors (ARs) belong to the G protein-coupled receptors (GPCRs) superfamily. The known four subtypes, termed adenosine A_1, A_{2A}, A_{2B} and A_3 receptors, are widely distributed in human body, involved in several physio-pathological processes and represent potential targets for the treatment of several diseases[52]. In the last decade, X-ray structures of the hA_{2A} AR in complex with agonists and antagonists have been released thus offering the basis for molecular modeling investigation[53] including also SuMD simulations[7,54,10] . Here we focus on the complex with NECA[55] (Figure 3F) that features a strong polar interaction between the exocyclic amine group of NECA and the side chain of the conserved Asn253 residue; a hydrogen bond with the nitrogen atom of NECA acetamide moiety and the Thr88 side chain; and an aromatic π-π stacking with the conserved Phe168, located in the second extracellular loop (EL2), and hydrophobic contacts with, among others, the Leu249 side chain.

Globular Systems

Acid Ellagic-CK2 recognition pathway. In the starting geometry the ligand was placed at a distance of 50 Å from the binding site. After the initial randomization step, the distance reduced to 43 Å. As depicted in Figure 4A and shown in Video S1, the first interaction between the ligand and the protein is established after 2 ns of productive trajectory and is mediated Lys49 that directs the ligand to the P-loop of the kinase. As shown by the Pollicino analysis (Figure 4B), the ellagic acid approaches the region of the P-loop and mostly interacts with the Arg47, Lys49, Glu53 and the Lys71 (Figure 4A). These residues describe an interaction site, at 10.5 Å where the ligand resides for about 6 ns. In facts, the ligand RMSD plot (Figure 4C) records stable values in the 2-8 ns time lapse. The IE with the protein in this site is about -20 kcal/mol (Figure 4D at $dcm_{(L-R)}$= 10 Å); the Per Residue Contacts Count graph (Figure 4E) highlights that the above mentioned residues are those establishing the greatest number of contact, whereas the corresponding 3D models helps in identifying their location (Figure 4F) and the chronological order at which they have been approached by the ligand (Figure 5A). Approximately after 7 ns of simulation the ligand moves toward the orthosteric site, where Leu45 stabilizes its conformation and the side-chain of His160 hampers its passage. Through an interaction mediated by Arg43 the ligand overcomes the His160 gate and reaches new interaction site described by Asp120, Arg47, and Met163.

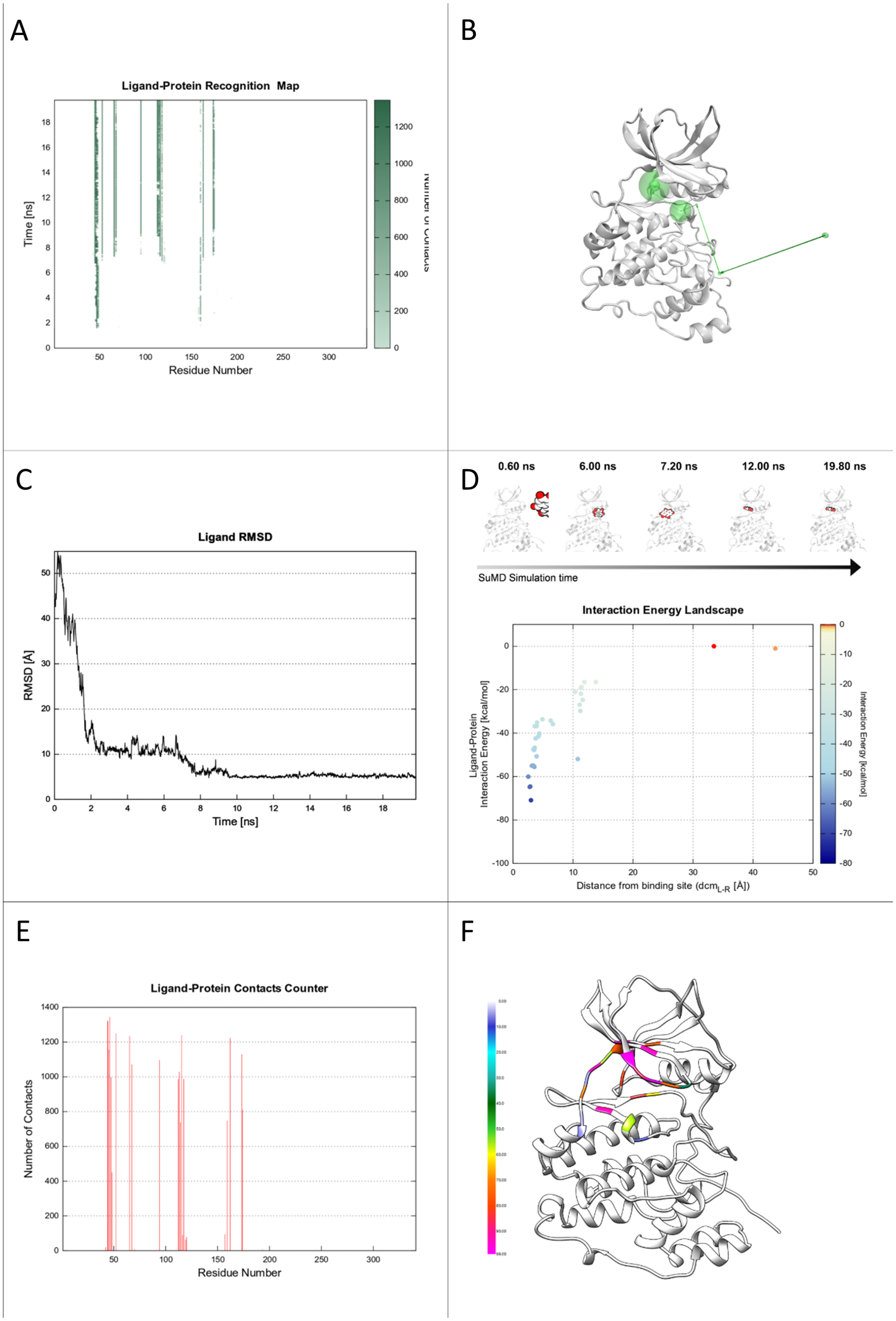


Figure 4 – Acid Ellagic-CK2 recognition pathway. **(A)** Ligand-Protein Recognition Map **(B)** Pollicino Analysis **(C)** Ligand RMSD **(D)** IE Landscape **(E)** Ligand-Protein Contacts Count **(F)** Chimera contacts.

Submitted - Cuzzolin, A.; Sturlese, M.; Deganutti, G.; Salmaso, V.; Sabbadin, D.; Ciancetta, A.; Moro, S. *J. Med. Chem.*

The permanency in this site is about of 2 ns with an interaction energy of -51 kcal/mol (Figure 4C-D). Consistently, the RMSD plot presents another plateau in the time range 8-10 ns (Figure 4C) that corresponds to the swarm of dots in the IE Landscape at dcm$_{(L\text{-}R)}$ = 11 Å (Figure 4D). A further stabilizing interaction with the Asn118 induces a shift in the ligand position that places the ring system parallel to the β7-β8 strands (Video S1).

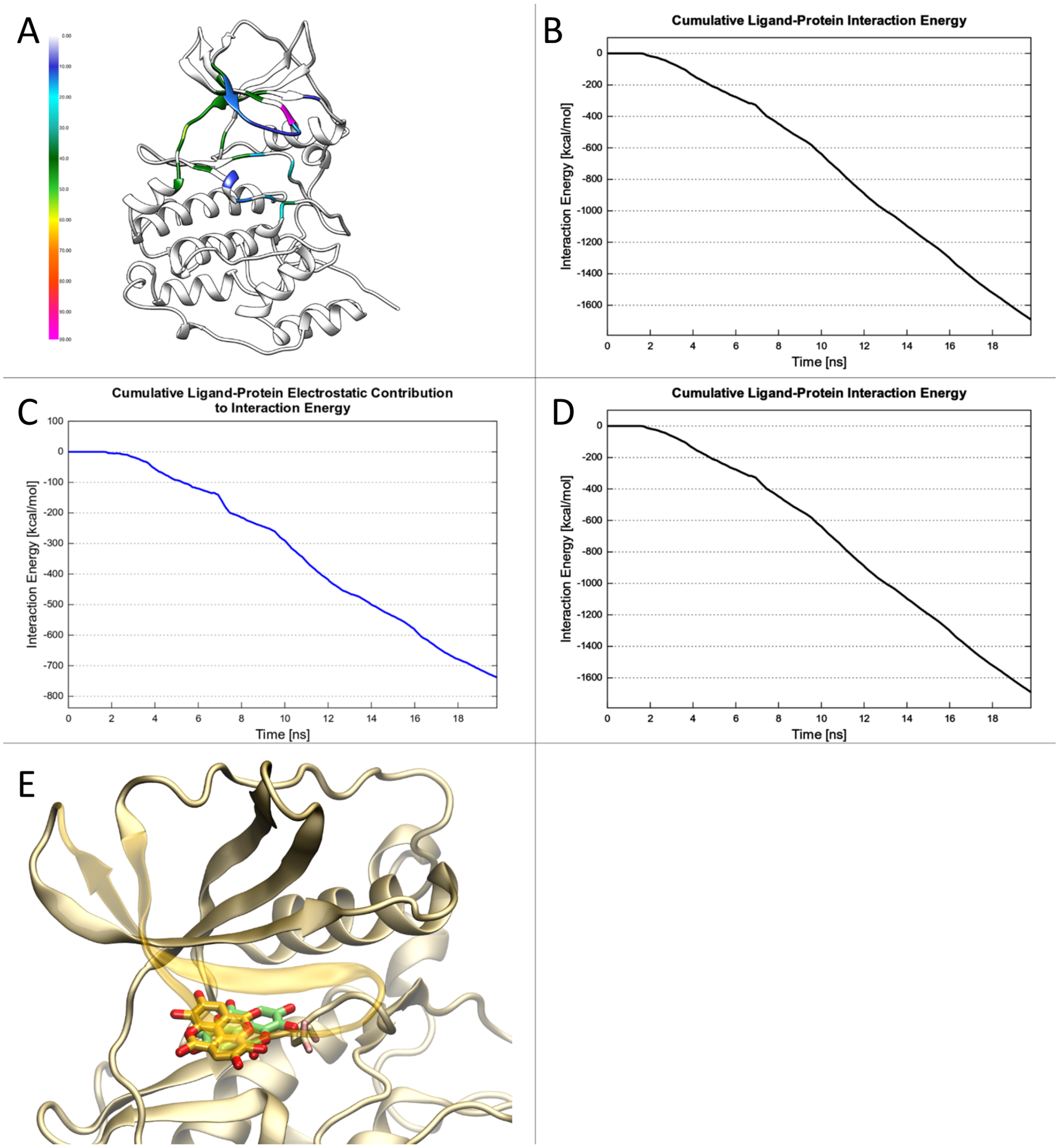


Figure 5 – Acid Ellagic-CK2 recognition pathway. **(A)** Chimera time **(B)** Cumulative IE **(C)** Cumulative IE electrostatic contribution **(D)** Cumulative IE van der Waals contribution **(E)** Superimposition between SuMD endpoint conformation and X-ray binding mode.

Submitted - Cuzzolin, A.; Sturlese, M.; Deganutti, G.; Salmaso, V.; Sabbadin, D.; Ciancetta, A.; Moro, S. *J. Med. Chem.*

As shown in the Cumulative Ligand-Protein IE (Figure. 5B) and its corresponding decomposition into electrostatic and van de Waal contribution, (Figure 5C and D, respectively) the change in the slope indicates that new conformation has a lower interaction energy than the previous one.

In particular, as highlighted by the comparison of the graphs relative to the electrostatic and van der Waals contribution (Figure 5C and D, respectively), the stabilization can be ascribed by the establishment on an electrostatic interaction with Asp175. As result of the new interaction the ligand moves into the orthosteric site (Figure 5E) and interacts with Lys159, Val66, Val117, Val53, His115 and Lys68 by maintaining the same position is maintained till the end of the SuMD simulation. The RMSD plot shows another plateau from 10 ns to the end, whereas the IE Landscape indicates that in this time lapse the ligand is at a distance around 2.5 Å with an IE between -40 to -70 kcal/mol.

The simulation was replicated three times and Replicas Analysis results are reported in Figure 6. In particular, the RMSD plot indicates that one replica does not reach the orthosteric site (Figure 6A, green line), whereas the others reach the same final RMSD value.

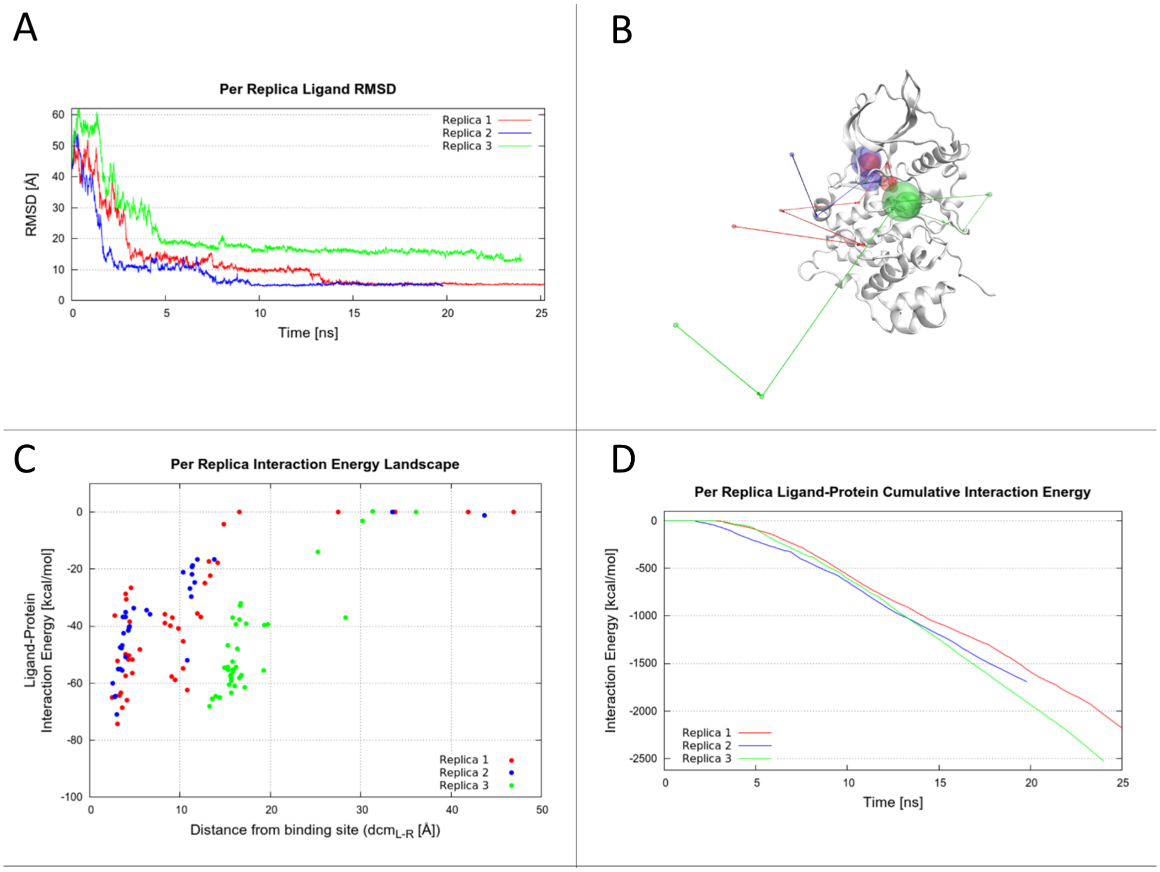


Figure 6 – Acid Ellagic-CK2 recognition pathway **(A)** Per Replica Ligand RMSD **(B)** Per Replica Pollicino Analysis **(C)** Per Replica IE Landscape **(D)** Per Replica cumulative IE.

Submitted - Cuzzolin, A.; Sturlese, M.; Deganutti, G.; Salmaso, V.; Sabbadin, D.; Ciancetta, A.; Moro, S. *J. Med. Chem.*

The same conclusion arises from the investigation of the Pollicino analysis where the ligand pathway of the two replicas converge in proximity of the protein (Figure 6B, red and blue spheres).

The Per Replica IE Landscape helps in explaining why the third replica does not reach the orthosteric site: as indicated by the green dots in Figure 5C the ligand reaches a different interaction site with an IE of -60 kcal/mol, a value close to the IE of the replicas that converge into in the orthosteric site (Figure 6C, red and blue dots). This consideration is confirmed by the trend of the Per Replica Cumulative IE that highlights a more negative slope for the third replica (Figure 6D, green line), indicating a very strong interaction.

SASP-GSTP1-1 recognition pathway. During the SuMD simulation the SASP reaches the GSTP1-1 catalytic H site in less than 6 ns (Video S2). The IE landscape highlights the formation of the first protein-ligand stabilizing interaction when the ligand and protein H site distance is 15 Å (point a, Figure 7A and 7B).

In this preliminary complex, SASP engages the Gly205 backbone oxygen in a hydrogen bond interaction through its sulfamide nitrogen atom and establishes an aromatic π-π stacking interaction between the salicylic moiety and Tyr108 (interactions corresponding to the first continues lines in the Protein-Ligand Recognition Map, Figure 7C). This situation anticipates a ligand positional shift that allows the SASP salicylic carboxylate to approach the positively charged Arg13 side chain, while the benzene ring replaces the salicylic aromatic moiety in the π-π stacking interaction with Tyr108 (point b, Figure 7A).

The energy stabilization of the complex increases and, after 8 ns of simulation, SASP proceeds toward a farther conformation, able to gain a more favourable electrostatic interaction geometry with Arg13 side chain, after the displacement of two water molecules from the solvation sphere of the positively charged residue.

This new pose (point c, Figure 7A and Figure 7B) is retained until the end of SuMD simulation, with the exception of conformational changes occurring to thepyridylsulfamoyl moiety, able to fit in the hydrophobic pocket delimited by Phe8, Val35 Trp38. During the SASP - GSTP1-1 recognition event GSH remain in the catalytic G site of the enzyme, not interacting with the inhibitor. Figure 7D highlights all the residues involved in the interaction with SASP during the SuMD simulations: the selective contacts towards only one enzymatic subunit, as well as the topologically restricted area interested, are well defined by the ribbon colorations.

Considering the SASP crystallographic conformation as geometrical reference, the ligand RMSD analysis (Figure 7E) reaches a minimum after 15 ns of simulation (Figure 7F), before level out at a value of about 5 Å. Figure S2 reports other ligand-protein interaction energy analysis.

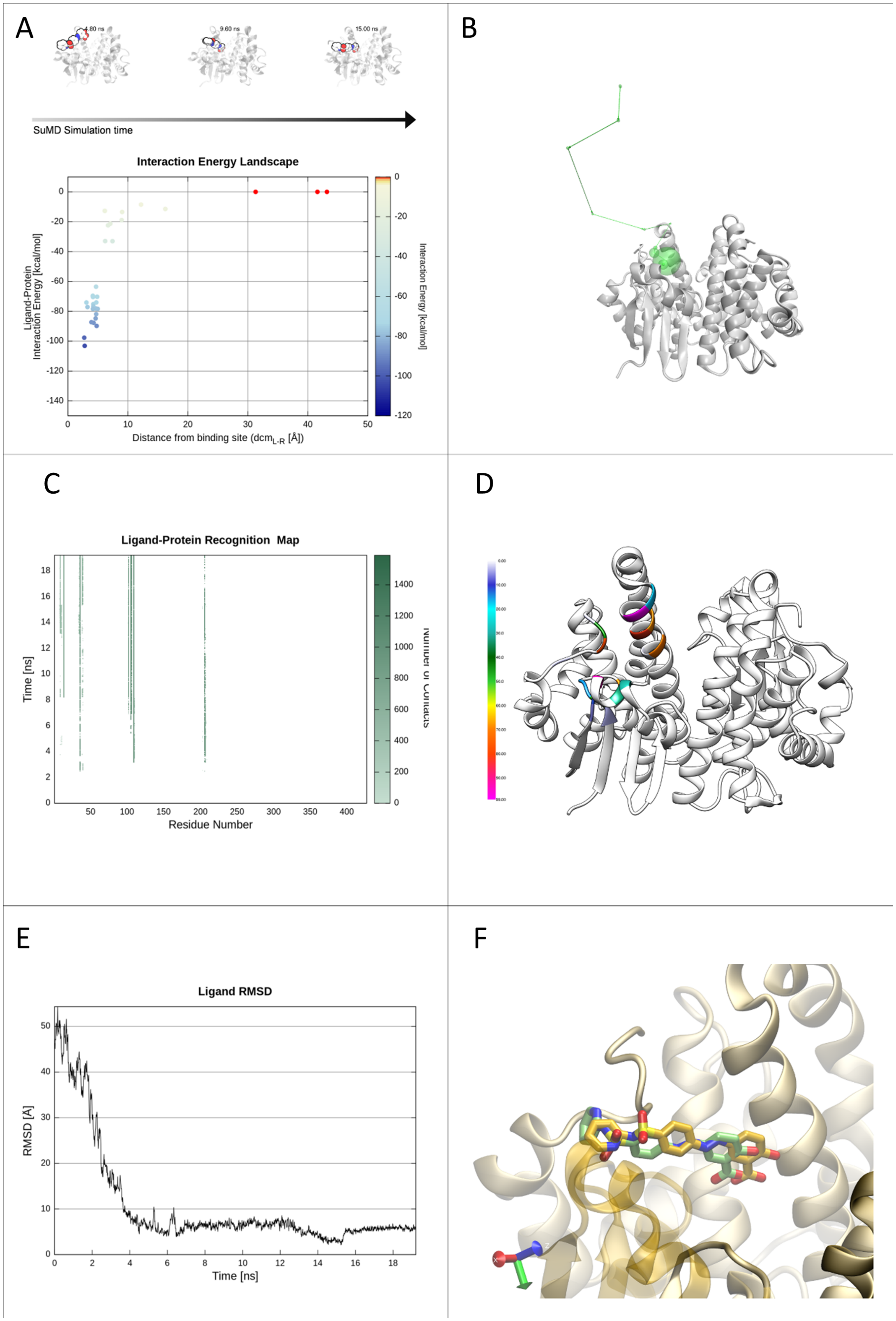


Figure 7 – SASP-GSTP1-1 recognition pathway. **(A)** IE Landscape **(B)** Pollicino Analysis **(C)** Ligand-Protein Recognition Map **(D)** Chimera contacts **(E)** Ligand RMSD **(F)** Superimposition between SuMD endpoint conformation and X-ray binding mode.

Submitted - Cuzzolin, A.; Sturlese, M.; Deganutti, G.; Salmaso, V.; Sabbadin, D.; Ciancetta, A.; Moro, S. *J. Med. Chem.*

The Replicas Analysis (Figure S3) depicts a recognition event with no meta-stable binding sites and characterized by almost a univocal pathway. Nevertheless, in one replica, in the final complex SASP is rotated by 180° (as highlighted by the higher RMSD value) and loses the electrostatic stabilization between its salicylic moiety and Arg13 side chain.

Benzen-1,2-diol−PDRX5 recognition pathway. The simulation were repeated on both monomeric and dimeric form yielding similar results. However, here we will focus on the dimeric form according to solution NMR studies, in which the authors stated the protein as dimer[56] . At the beginning of randomization step the fragment was placed at 78 Å from PDRX5 binding site ($dcm_{(L-R)}$= 78 Å). As reported in figure 8A, 8B (point b) and 8C, after nearly 3 ns the fragment approaches the protein in a region located at around 30 Å from the primary binding site (Video S3).

This meta-binding site lies in the opposite monomeric subunit with respect to the primary binding site and it is defined by residues Leu62, Lys63, Val69, and Val70. As shown by the IE landscape and the Pollicino Analysis (Figure 8A and 8B, respectively), this site engages the ligand with favorable interactions for a couple of nanoseconds. In particular, the formation of a hydrogen bond between the hydroxyl groups of catechol and the carbonyl moiety of the backbone amide of residue Lys95 stabilizes this conformation. After nearly 6 ns the fragment is released by this site and fluctuates to finally reach the primary binding site thought a series of molecular interaction, including residues (chronologically sorted): Glu91, Glu16, Glu18, Phe79 belonging to the first monomer unit (SI figure S4). Finally, the fragment accesses to the binding site where fluctuates experimenting different conformations in accordance with its affinity in the millimolar range. The fluctuations of the fragment in the binding site are also evident in the Protein-Ligand Energy profiles, in which the energy wavers around the value of -20 kcal/mol (SI figure S4).

During the fluctuation, the catechol enters in contact with most of the residue forming the site, in particular (sorted by number of molecular contacts during the trajectory): Thr146, Thr44, Arg127, Phe120, Leu116, Gly46 and Cys47 (Figure 8C, 8D). The main conformation observed corresponds to the crystallographic one, as reported in Figure 8E and 8F where the RMSD reaches a minimum value 0.69 Å at 17.3 ns.

The simulation was repeated in three times randomizing the position of the ligand. The Replicas Analysis is reported in Figure S5. Briefly, in each replica the fragment reached the primary binding site experiencing the conformation reported in the crystallographic data with the best RMSD respectively of 1.12 and 1.24 Å for the replica 2 and 3.

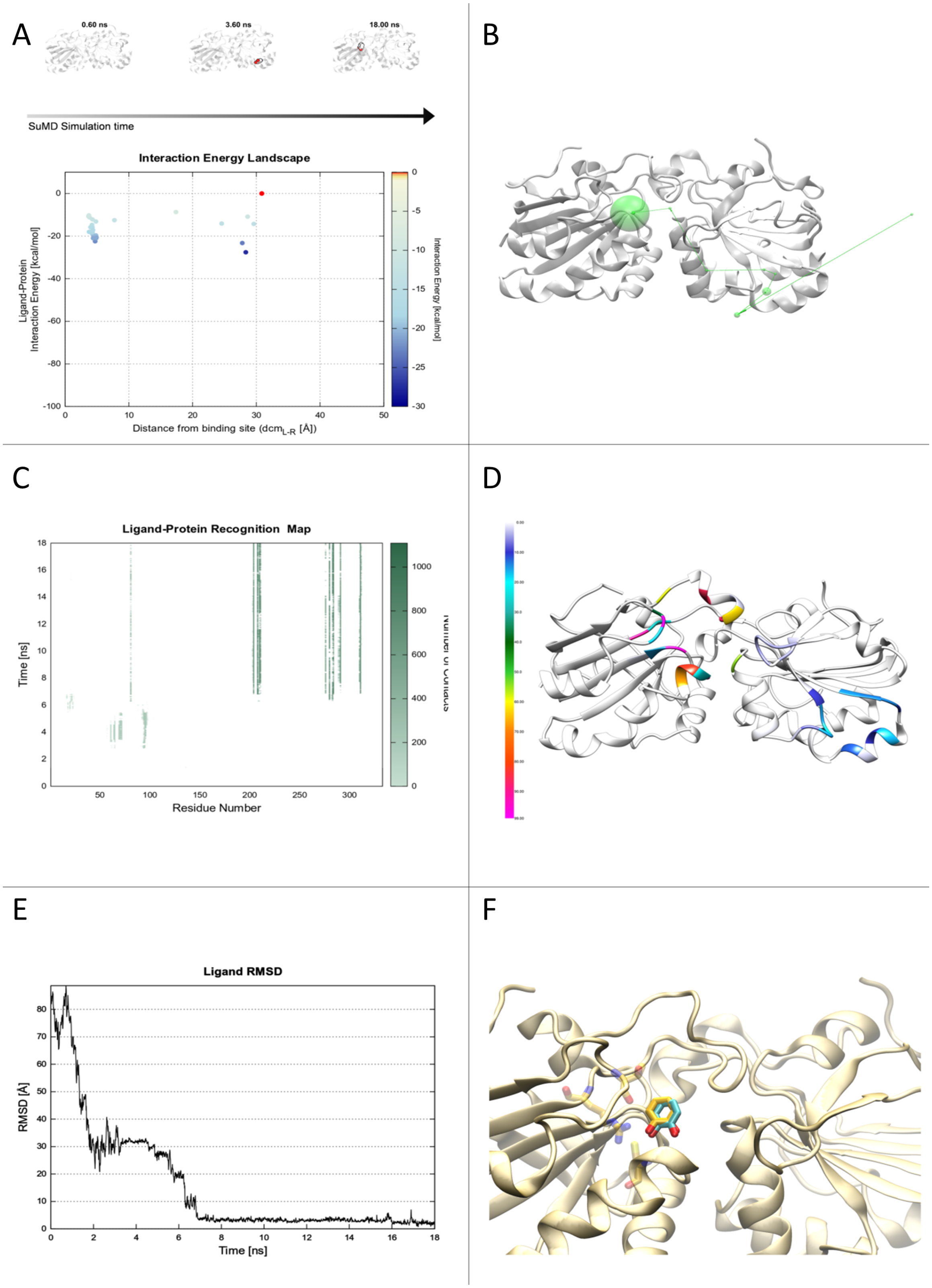


Figure 8 –Benzen-1,2-diol-PDRX5 recognition pathway. **(A)** IE Landscape **(B)** Pollicino Analysis **(C)** Ligand-Protein Recognition Map **(D** Chimera contacts. **(E)** Ligand RMSD **(F)** Superimposition between SuMD endpoint conformation and X-ray binding mode.

Submitted - Cuzzolin, A.; Sturlese, M.; Deganutti, G.; Salmaso, V.; Sabbadin, D.; Ciancetta, A.; Moro, S. *J. Med. Chem.*

(S)-naproxen-HAS recognition pathway. SuMD simulation was performed maintaining decanoic ligand in the IB site according the crystallographic geometries. (S)-naproxen was separated from HSA-decanoid acid complex by 32 Å from IB site (point a in Figure 9A and 9B). In the first SuMD step the ligand fluctuate till 50 Å from the IB site. As reported in Figure 9C after a couple of nanosecond the ligand approaches the first protein site in its trajectory by engaging Lys510 and Thr564 (Video S4). Shortly after, the ligand establishes a network of interaction for 1 ns (from 2.3 ns to 3.2) in a site located at around $dcm_{(L-R)}$= 20 Å (point b in Figure A), defined by residues: Val116, Pro118, Val122, Thr133 and Phe134. Then, the Naproxen molecule approach a second, where it fluctuates for about 3 ns by establishing strong interaction with residues Leu115, Pro118, Lys137, and Ile142 (as also evident from Protein-Ligand Interaction energy in figure S6 in SI).

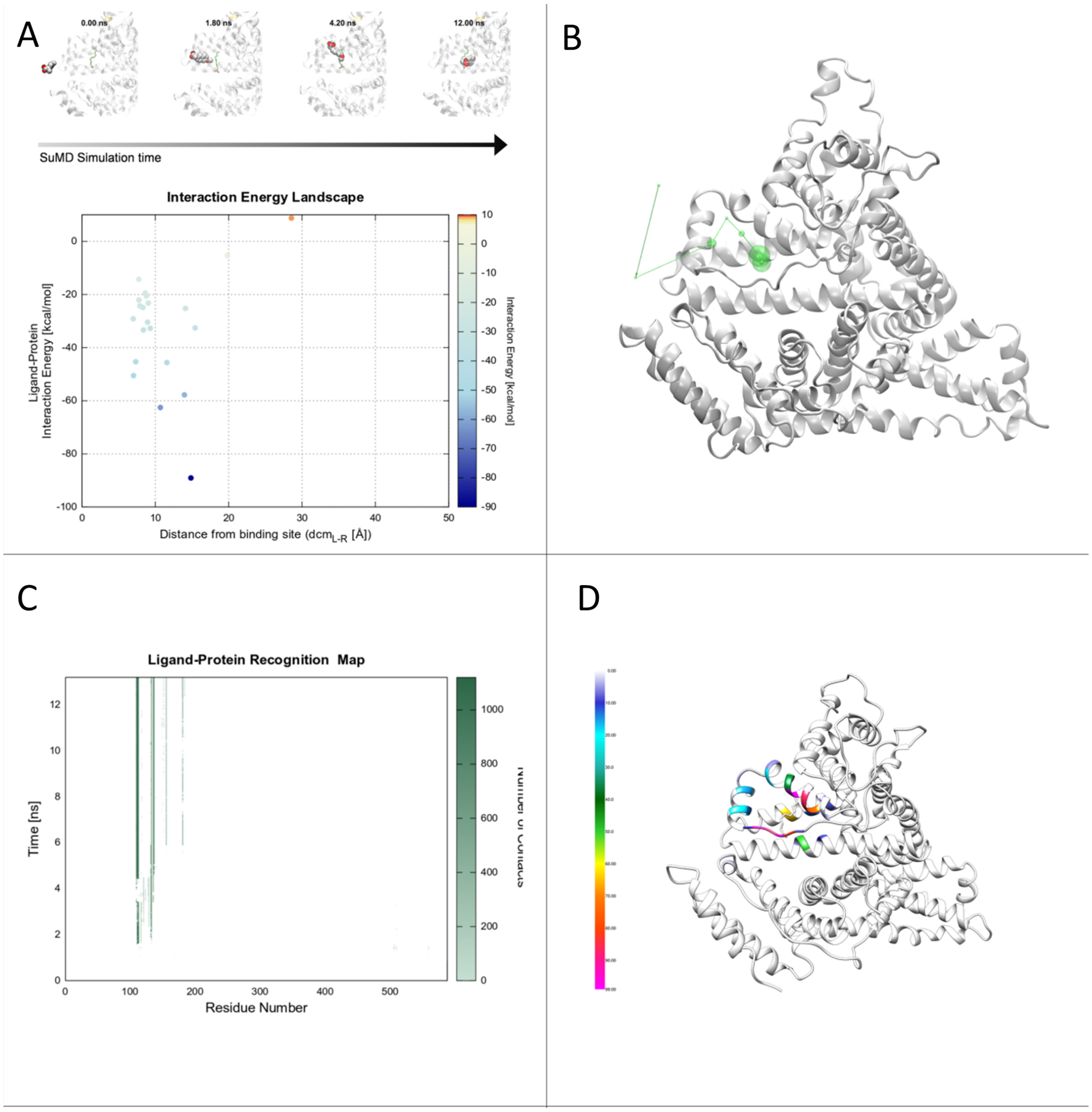


Figure 9 – (S)-naproxen-HAS recognition pathway. **(A)** IE Landscape **(B)** Pollicino Analysis **(C)** Ligand-Protein Recognition Map **(D** Chimera contacts.

Submitted - Cuzzolin, A.; Sturlese, M.; Deganutti, G.; Salmaso, V.; Sabbadin, D.; Ciancetta, A.; Moro, S. *J. Med. Chem.*

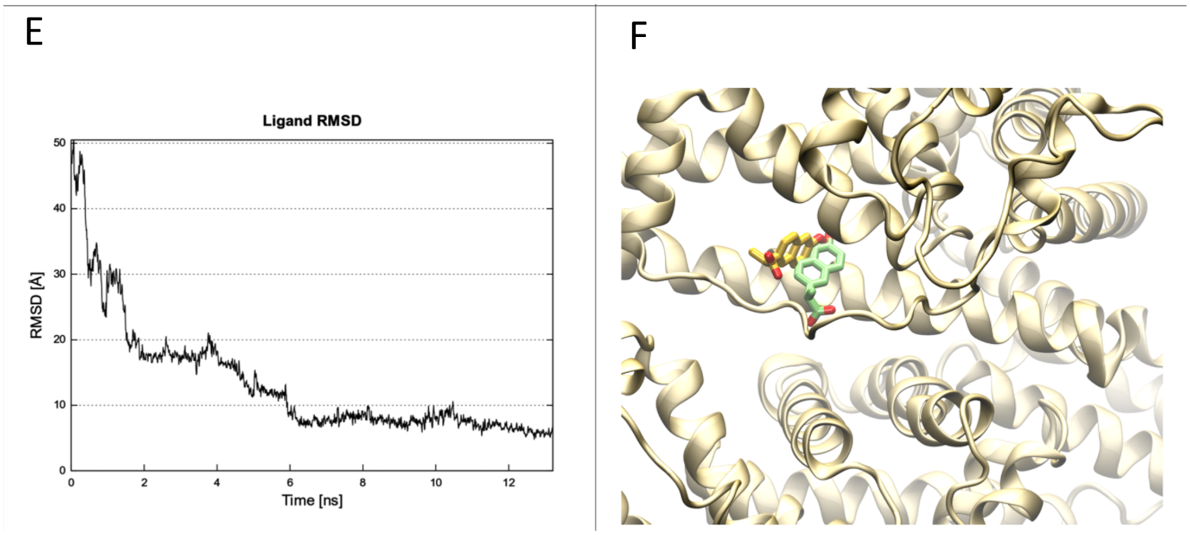


Figure 9 (continuation)**- (E)** Ligand RMSD **(F)** Superimposition between SuMD endpoint conformation and X-ray binding mode.

This meta-binding site is located in front of the principal binding site to which is separated by the presence of a long extended loop (residue 106 to 119) that acts as a gate for the IB site. Finally after 6 ns, (S)-naproxen is able to pass behind the extended loop and reach the IB site (residues Leu115, Ile142, Phe157, Tyr161) as show by Figure 9B and 9E. Within the primary site, the ligand is able to place the methyl ether group in the proximity of Phe152 very similarly to the orientation of crystal structure. On the other hand, the naphthalene core and in particular the carboxylic group adopt a different orientation due to the presence of the extended loop. This different orientation abolishes the ionic interaction between the carboxyl group and the Arg112 observed in the crystallographic structure (Figure 9F). At the end of the simulation the RMSD fluctuates around 5 Å, reaching the lowest value of 4.76 at 12.70 ns (Figure 9E and 9F).

Interestingly, in the other replicas (Figure S7) the ligand reaches the IB site by approaching the extended loop from a different position and occupies a slightly different location in the vast IB site. This suggests the loop might have a crucial role in the recognition process (Figure S7).

Transmembrane Systems

(S)-fluoxetine-LeuT recognition pathway. The (S)-fluoxetine recognition pathway highlights, after 1 ns of SuMD simulation, a first electrostatic interaction between Asp 158 side chain and the ligand charged secondary amine group (Video S5). The energetic stabilization characterizing this complex corresponds to the IE landscape minimum reported in the Figure 10A, point a and Figure 10B. This preliminary complex is able to favor the ligand approach towards an inner pocket of LeuT, topologically defined by Tyr 471 and the aliphatic chains of Lys 474 and Glu 478, reciprocally involved in an ionic interaction. Hydrophobic contacts stabilize this intermolecular complex for about 2 ns, before a conformational change allows (S)-fluoxetine to establish a more favorable electrostatic interaction whit Glu 402 side chain. This scenario anticipates the ligand repositioning inside an inner hydrophobic site, where the ligand engages for almost 7ns Tyr471, Trp406, Ile475 and Phe405 side chains in lipophilic interactions through its phenyl ring (point b, Figure 10A and Figure 10B). During the remaining simulation time, the inhibitor makes contacts with Ala319 (EL4) and the side chains of the key residues Asp404 and Arg30 (point c, Figure 10A and Figure 10B and continuous lines corresponding to the last 4ns of SuMD simulation in Figure 10C), both located at the protein extracellular gate and involved in a ionic lock that 10terically obstructs the SSRIs binding site disclosed by LeuT crystallographic structure.

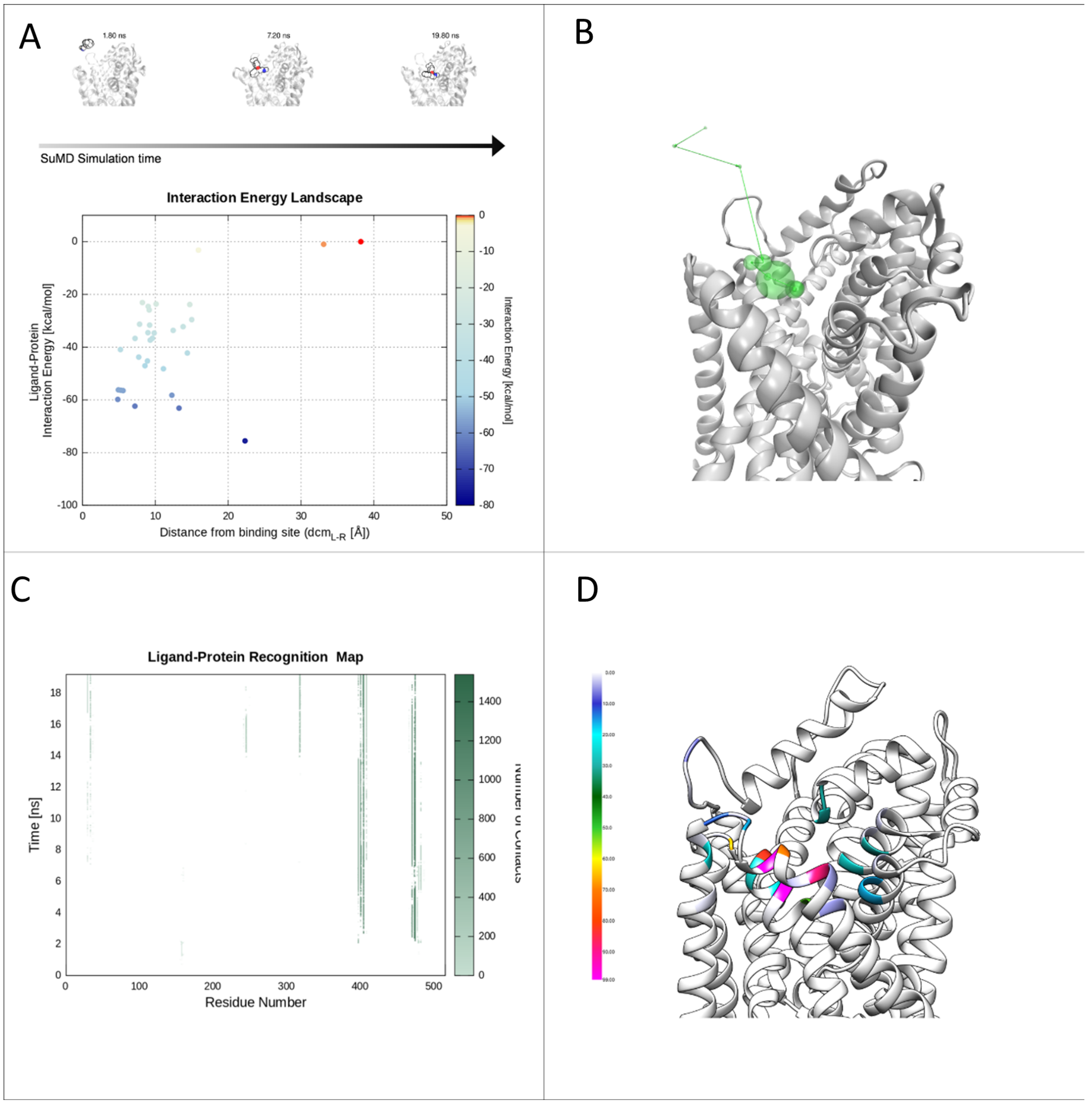


Figure 10 – (S)-fluoxetin-LeuT recognition pathway. **(A)** IE Landscape **(B)** Pollicino Analysis **(C)** Ligand-Protein Recognition Map **(D** Chimera contacts.

Submitted - Cuzzolin, A.; Sturlese, M.; Deganutti, G.; Salmaso, V.; Sabbadin, D.; Ciancetta, A.; Moro, S. *J. Med. Chem.*

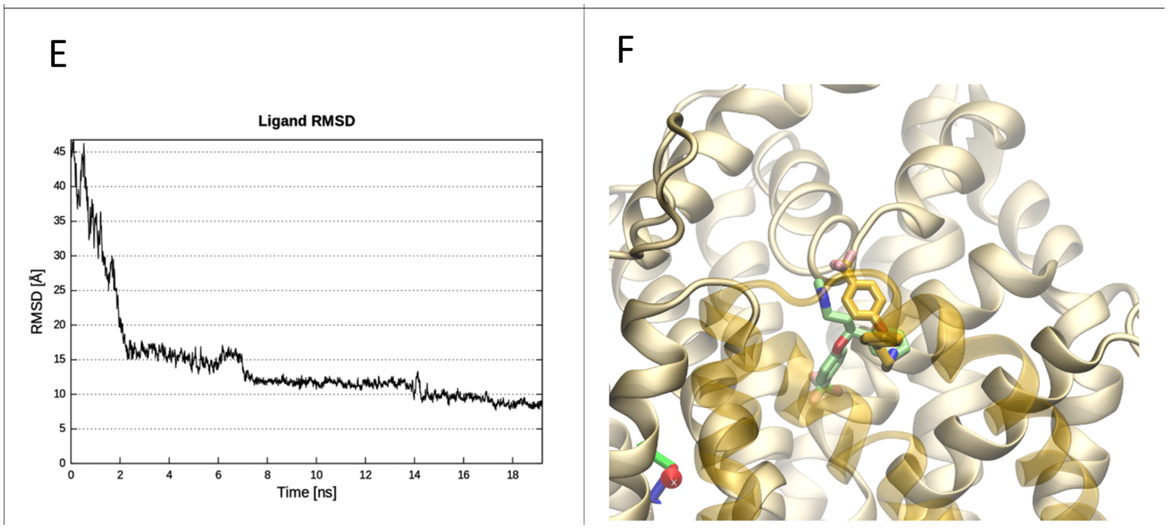


Figure 10 (continuation) **– (E)** Ligand RMSD **(F)** Superimposition between SuMD endpoint conformation and X-ray binding mode.

Figure 10D summarizes all the amino acids involved in the SFX recognition event during the SuMD simulation.

The RMSD plot (Figure 10E) outlines the inhibitor difficulty in reproducing the experimental pose (Figure 10F). Investigation on LeuT crystal structure without co-crystallized inhibitor reveale an alternative conformation of Arg30 side chain, and the absence of the gate ionic lock (Figure S12)[57] : it is possible to speculate that the LeuT extracellular gate, during SuMD simulation timescale, is able to remain in a stable conformation, previously induced by the inhibitor binding and retained even after the removal of the ligand during the system preparation for SuMD.

Replicas analysis (Figure S9) highlights two alternative recognition pathways through the extracellular vestibule, unable to enable SFX to reproduce the binding mode observed in the crystallographic complex and characterized by accentuated energy variations in proximity of the extracellular transporter gate.

NECA−hA$_{2A}$ AR recognition pathway. NECA establishes the first stabilizing contacts with hA$_{2A}$ AR after about 4ns of SuMD simulation (Video S6). During this initial scenario (point a, Figure 11A and Figure 11B) the ligand approaches the protein topological structure defined by ECL2 N-terminus and the residues located at top of TM5 and TM6. More precisely, NECA engages Phe257 (6.59) side chain in a π-π stacking interaction through its purine scaffold and locates the N-ethylcarboxamido moiety towards a pocket delimited by Trp143 (ECL2), Pro173 (ECL2), and Asn175 (TM5) side chains, as highlighted by the first stripes in Figure 11C and the yellow and violet ribbon coloration in Figure 11D.

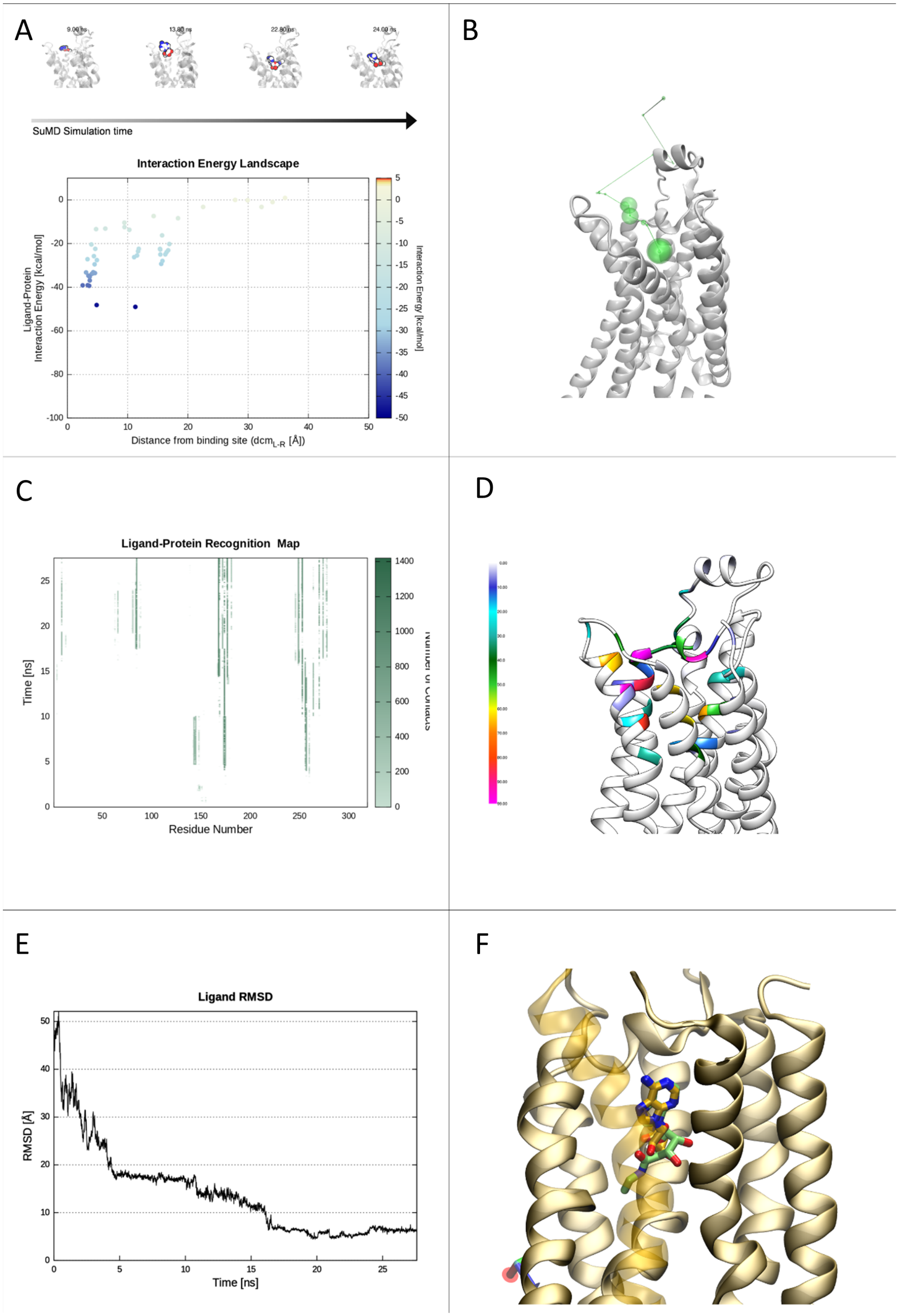


Figure 11 – NECA-hA$_{2A}$ AR recognition pathway. **(A)** IE Landscape **(B)** Pollicino Analysis **(C)** Ligand-Protein Recognition Map **(D)** Chimera contacts **(E)** Ligand RMSD **(F)** Superimposition between SuMD endpoint conformation and X-ray binding mode.

Submitted - Cuzzolin, A.; Sturlese, M.; Deganutti, G.; Salmaso, V.; Sabbadin, D.; Ciancetta, A.; Moro, S. *J. Med. Chem.*

This complex anticipates a repositioning that allows the ligand to reach a meta-stabile binding site, mainly characterized by a π-π stacking interaction with His264 (EL3) side chain, an hydrophobic contact in the direction of Met174 (TM5) side chain, and an hydrogen bond interaction between its C2' hydroxide group and Asn253 (TM6) (point b, Figure 11A and Figure 11B).

During the time slot rising from 14 ns to 20 ns of SuMD simulation, the agonist reaches a deeper position inside the orthosteric binding site and explores different conformations (included a temporary anti-syn transition about the glycoside linkage), until engages Phe168 (ECL2) side chain in a π-π stacking interaction and Asn253 (TM6) side chain in hydrogen bond interactions through its exocyclic amine and the N7 position of the purine scaffold (point c, Figures 11A and Figures 11B). This complex orientation (associated with the minimum RMSD value in Figure 11E, with respect to the NECA crystallographic conformation) is followed by an alternative stabilized conformation (point d, Figure 11A and Figure 11B) which involves also hydrophobic interactions with Leu249 (TM6), Leu85 (TM3) and Val84 (TM3).

During the remaining SuMD simulation time, the protein-ligand complex geometry remain almost unaltered, with the exception of a reorientation of the N-ethylcarboxamidoribose moiety, pointing toward TM4, and the loss of the aromatic π-π interaction due to a conformational change occurring to Phe168 (EL2) side chain. In Figure 10S are reported other ligand-protein energy interaction analysis.

At the minimum RMSD value, NECA pyrimidine scaffold coincides with the crystallographic orientation, while the ribose moiety is oriented in an alternative conformation (Figure 11F).

Replicas Analysis (Figure 11S) highlights also a different NECA recognition pathway, which involves residues located at the ECL2 and characterized by comparable energetic stabilizations.

Conclusion

In the present work, we have demonstrated the general applicability of SuMD simulations using different types of proteins, including both globular and membrane proteins. Moreover, we have presented the SuMD-Analyzer tool that helps, also a non expert user, in the analysis of the SuMD trajectories. Even if various other MD methods have also been used to characterize binding pathways, SuMD has the great advantage of being able to explore the ligand-protein approaching path in nanosecond simulation time scale. Furthermore, SuMD simulations enable the investigation of ligand-protein binding events independently from the starting position, chemical structure of the ligand, and also from its target binding affinity. As described for each key study, SuMD simulations are able to characterize multiple ligand-protein binding pathways identifying a variety of metastable intermediate states (meta-binding sites). These information may be an interesting starting point for further argumentations regarding the pharmacological consequences of that specific ligand-protein recognition process. Moreover, it is worthy to underline that, contrary to expectations, not all SuMD trajectories converge to the structure of the complex obtained crystallographically. Indeed, there are several plausible reasons that may be argued to describe this particular unexpected aspect: *a)* the crystallographically pose of the ligand is not the only minimum of the potential energy surface described by the force field during the SuMD simulations; *b)* the crystallographically conformation of the protein in its bound state is remarkably different respect its apo-form. This could be interpreted as the sign of an important induce-fit process during the ligand recognition; and *c)* the boundary conditions that led to the formation of the crystallographically ligand-protein complex (solvent and co- solvent, pH, ionic strength, or temperature just as a few examples) are not well described during the SuMD simulations. This must always be kept in mind when any conjecture is made starting from the analysis of SuMD trajectories. Currently, a major effort is underway to estimate, from SuMD simulations, binding kinetics properties (in particular on-rate values) in approximate agreement with experimental measurements.

Hopefully, the future of drug design will involve detailed characterization of not only the bound state but also the whole ligand–protein network of recognition pathways, including all metastable intermediate states (meta-binding sites). With such a complete understanding we hope expand our perspectives in several scientific areas from molecular pharmacology to drug discovery.

Submitted - Cuzzolin, A.; Sturlese, M.; Deganutti, G.; Salmaso, V.; Sabbadin, D.; Ciancetta, A.;
Moro, S. *J. Med. Chem.*

Abbreviations

3D	three-dimensional
CK2	caseine kinase 2
CPU	central processing unit
GPCRs	G protein-coupled receptors
GPU	graphics processor unit
GSTP1-1	P1-1 isoform of glutathione S-transferase
hA_{2A}AR	human A_{2A} adenosine receptor
HSA	human serum albumin
IE	interaction energy
K_d	equilibrium dissociation constant
K_{off}	dissociation rate constants
K_{on}	association rate constants
LeuT	leucine transporter
MD	molecular dynamics
NECA	5'-N-ethylcarboxamidoadenosine
PDB	protein data bank
POPC	1-palmitoyl-2-oleoyl-snglycero-3-phosphocholine
PRDX5	eroxiredoxin 5
RMSD	root-mean-square deviation
SASP	sulphasalazine
SuMD	supervised molecular dynamics

Submitted - Cuzzolin, A.; Sturlese, M.; Deganutti, G.; Salmaso, V.; Sabbadin, D.; Ciancetta, A.; Moro, S. *J. Med. Chem.*

Bibliography

1. Protein-Ligand Interactions: From Molecular Recognition to Drug Design; Böhm, H.-J., Schneider, G., Eds.; Methods and Principles in Medicinal Chemistry; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, FRG, 2003.
2. Pan, A. C.; Borhani, D. W.; Dror, R. O.; Shaw, D. E. Molecular Determinants of Drug– receptor Binding Kinetics. Drug Discov. Today 2013, 18 (13–14), 667–673.
3. Moro, S.; Hoffmann, C.; Jacobson, K. A. Role of the Extracellular Loops of G Protein- Coupled Receptors in Ligand Recognition: A Molecular Modeling Study of the Human P2Y1 Receptor. Biochemistry (Mosc.) 1999, 38 (12), 3498–3507.
4. Dror, R. O.; Jensen, M. Ø.; Borhani, D. W.; Shaw, D. E. Exploring Atomic Resolution Physiology on a Femtosecond to Millisecond Timescale Using Molecular Dynamics Simulations. J. Gen. Physiol. 2010, 135 (6), 555–562.
5. Buch, I.; Giorgino, T.; De Fabritiis, G. Complete Reconstruction of an Enzyme-Inhibitor Binding Process by Molecular Dynamics Simulations. Proc. Natl. Acad. Sci. U. S. A. 2011, 108 (25), 10184–10189.
6. Johnston, J. M.; Filizola, M. Beyond Standard Molecular Dynamics: Investigating the Molecular Mechanisms of G Protein-Coupled Receptors with Enhanced Molecular Dynamics Methods. In In G Protein-Coupled Receptors - Modeling and Simulation 796; 95–125, Springer Netherland, 2014; p pp.
7. Sabbadin, D.; Moro, S. Supervised Molecular Dynamics (SuMD) as a Helpful Tool To Depict GPCR–Ligand Recognition Pathway in a Nanosecond Time Scale. J. Chem. Inf. Model. 2014, 54 (2), 372–376.
8. Ciancetta, A.; Sabbadin, D.; Federico, S.; Spalluto, P.; Moro, S. Advances in Computational Techniques to Study GPCR-Ligand Recognition. In Trends in Pharmacol Sci (2015) in press DOI: 10.1016/j.tips.2015.08.006.
9. Sabbadin, D.; Ciancetta, A.; Deganutti, G.; Cuzzolin, A.; Moro, S. Exploring the Recognition Pathway at the Human A2A Adenosine Receptor of the Endogenous Agonist Adenosine Using Supervised Molecular Dynamics Simulations. MedChemComm 2015, 6 (6), 1081–1085.
10. Deganutti, G.; Cuzzolin, A.; Ciancetta, A.; Moro, S. Understanding Allosteric Interactions in G Protein-Coupled Receptors Using Supervised Molecular Dynamics: A Prototype Study Analysing the Human A3 Adenosine Receptor Positive Allosteric Modulator LUF6000. Bioorg. Med. Chem. 2015.
11. Paoletta, S.; Sabbadin, D.; von Kügelgen, I.; Hinz, S.; Katritch, V.; Hoffmann, K.; Abdelrahman, A.; Straßburger, J.; Baqi, Y.; Zhao, Q.; Stevens, R. C.; Moro, S.; Müller, C. E.; Jacobson, K. A. Modeling Ligand Recognition at the P2Y12 Receptor in Light of X- Ray Structural Information. J. Comput. Aided Mol. Des. 2015, 29 (8), 737–756.
12. Harvey, M. J.; Giupponi, G.; Fabritiis, G. D. ACEMD: Accelerating Biomolecular Dynamics in the Microsecond Time Scale. J. Chem. Theory Comput. 2009, 5 (6),

1632– 1639.

13. Case, D.; Babin, V.; Berryman, J.; Betz, R.; Cai, Q.; Cerutti, D.; Cheatham Iii, T.; Darden, T.; Duke, R.; Gohlke, H.; others. Amber14, Version AMBER14; Http://ambermd.org/ (accessed October 2015); University of California, San Francisco, 2014.
14. Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. J. Comput. Chem. 2004, 25 (9), 1157–1174.
15. MacKerell, A. D.; Banavali, N.; Foloppe, N. Development and Current Status of the CHARMM Force Field for Nucleic Acids. Biopolymers 2000, 56 (4), 257–265.
16. Vanommeslaeghe, K.; Raman, E. P.; MacKerell, A. D. Automation of the CHARMM General Force Field (CGenFF) II: Assignment of Bonded Parameters and Partial Atomic Charges. J. Chem. Inf. Model. 2012, 52 (12), 3155–3168.
17. Vanommeslaeghe, K.; MacKerell, A. D. Automation of the CHARMM General Force Field (CGenFF) I: Bond Perception and Atom Typing. J. Chem. Inf. Model. 2012, 52 (12), 3144–3154.
18. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. Nucleic Acids Res. 2000, 28 (1), 235–242.
19. CCG Inc. Molecular Operating Environment (MOE), 2014.09; Http://www.chemcomp.com (accessed October 2015).
20. Labute, P. Protonate3D: Assignment of Ionization States and Hydrogen Coordinates to Macromolecular Structures. Proteins 2009, 75 (1), 187–205. Stewart, J. J. P. MOPAC2012, Version 2012; http://OpenMOPAC.net (accessed October 2015); Stewart Computational Chemistry: Colorado Springs, CO, USA, 2012.
21. Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods V: Modification of NDDO Approximations and Application to 70 Elements. J. Mol. Model. 2007, 13 (12), 1173–1213.
22. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.;

Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. Gaussian 09, Revision B.01; Gaussian, Inc.: Wallingford, CT, 2009. (24)

23. MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiórkiewicz- Kuczera, J.; Yin, D.; Karplus, M. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. J. Phys. Chem. B 1998, 102 (18), 3586–3616.
24. Vanommeslaeghe, K.; MacKerell, A. D. Automation of the CHARMM General Force Field (CGenFF) I: Bond Perception and Atom Typing. J. Chem. Inf. Model. 2012, 52 (12), 3144–3154.
25. Head-Gordon, M.; Pople, J. A.; Frisch, M. J. MP2 Energy Evaluation by Direct Methods. Chem. Phys. Lett. 1988, 153 (6), 503–506.
26. Mayne, C. G.; Saam, J.; Schulten, K.; Tajkhorshid, E.; Gumbart, J. C. Rapid Parameterization of Small Molecules Using the Force Field Toolkit. J. Comput. Chem.2013, 34 (32), 2757–2770.
27. Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. J. Mol. Graph. 1996, 14 (1), 33–38, 27–28.
28. Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of Multiple Amber Force Fields and Development of Improved Protein Backbone Parameters. Proteins 2006, 65 (3), 712–725.
29. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. J. Chem. Phys. 1983, 79 (2), 926.
30. Berendsen, H. J. C.; Postma, J. P. M.; Gunsteren, W. F. van; DiNola, A.; Haak, J. R. Molecular Dynamics with Coupling to an External Bath. J. Chem. Phys. 1984, 81 (8), 3684–3690.
31. Loncharich, R. J.; Brooks, B. R.; Pastor, R. W. Langevin Dynamics of Peptides: The Frictional Dependence of Isomerization Rates of N-Acetylalanyl-N’-Methylamide. Biopolymers 1992, 32 (5), 523–535.
32. Kräutler, V.; van Gunsteren, W. F.; Hünenberger, P. H. A Fast SHAKE Algorithm to Solve Distance Constraint Equations for Small Molecules in Molecular Dynamics Simulations. J. Comput. Chem. 2001, 22 (5), 501–508.
33. Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A Smooth Particle Mesh Ewald Method. J. Chem. Phys. 1995, 103 (19), 8577–8593.
34. Lomize, M. A.; Lomize, A. L.; Pogozheva, I. D.; Mosberg, H. I. OPM: Orientations of Proteins in Membranes Database. Bioinforma. Oxf. Engl. 2006, 22 (5), 623–

625.

35. Grubmüller, H.; Groll, V. Solvate, Version 1.0.1; Http://www.mpibpc.mpg.de/grubmueller/solvate (accessed October 2015); 1996. Williams, T.; Kelley, C. Gnuplot 4.5: An Interactive Plotting Program, Version 4.5; Http://gnuplot.info (accessed October 2015).
36. Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera--a Visualization System for Exploratory Research and Analysis. J. Comput. Chem. 2004, 25 (13), 1605–1612.
37. Seeber, M.; Felline, A.; Raimondi, F.; Muff, S.; Friedman, R.; Rao, F.; Caflisch, A.; Fanelli, F. Wordom: A User-Friendly Program for the Analysis of Molecular Structures, Trajectories, and Free Energy Surfaces. J. Comput. Chem. 2011, 32 (6), 1183–1194.
38. Cozza, G.; Bortolato, A.; Moro, S. How Druggable Is Protein Kinase CK2? Med. Res.Rev. 2010, 30 (3), 419–462.
39. Sekiguchi, Y.; Nakaniwa, T.; Kinoshita, T.; Nakanishi, I.; Kitaura, K.; Hirasawa, A.; Tsujimoto, G.; Tada, T. Structural Insight into Human CK2α in Complex with the Potent Inhibitor Ellagic Acid. Bioorg. Med. Chem. Lett. 2009, 19 (11), 2920–2923.
40. Wilce, M. C.; Parker, M. W. Structure and Function of Glutathione S-Transferases. Biochim. Biophys. Acta 1994, 1205 (1), 1–18.
41. Laborde, E. Glutathione Transferases as Mediators of Signaling Pathways Involved in Cell Proliferation and Cell Death. Cell Death Differ. 2010, 17 (9), 1373–1380.
42. Oakley, A. J.; Bello, M. Lo; Nuccetelli, M.; Mazzetti, A. P.; Parker, M. W. The Ligandin (non-Substrate) Binding Site of Human Pi Class Glutathione Transferase Is Located in the Electrophile Binding Site (H-Site). J. Mol. Biol. 1999, 291 (4), 913–926.
43. Aguirre, C.; Brink, T. ten; Guichou, J.-F.; Cala, O.; Krimm, I. Comparing Binding Modes of Analogous Fragments Using NMR in Fragment-Based Drug Design: Application to PRDX5. PLOS ONE 2014, 9 (7), e102300.
44. Declercq, J.-P.; Evrard, C.; Clippe, A.; Stricht, D. V.; Bernard, A.; Knoops, B. Crystal Structure of Human Peroxiredoxin 5, a Novel Type of Mammalian Peroxiredoxin at 1.5 Å resolution1. J. Mol. Biol. 2001, 311 (4), 751–759.
45. Shichita, T.; Hasegawa, E.; Kimura, A.; Morita, R.; Sakaguchi, R.; Takada, I.; Sekiya, T.; Ooboshi, H.; Kitazono, T.; Yanagawa, T.; Ishii, T.; Takahashi, H.; Mori, S.; Nishibori, M.; Kuroda, K.; Akira, S.; Miyake, K.; Yoshimura, A. Peroxiredoxin Family Proteins Are Key Initiators of Post-Ischemic Inflammation in the Brain. Nat. Med. 2012, 18 (6), 911–917.
46. Sjöholm, I.; Ekman, B.; Kober, A.; Ljungstedt-Påhlman, I.; Seiving, B.; Sjödin, T. Binding of Drugs to Human Serum albumin:XI. The Specificity of Three Binding

Sites as Studied with Albumin Immobilized in Microparticles. Mol. Pharmacol. 1979, 16 (3), 767– 777.

47. Lejon, S.; Cramer, J. F.; Nordberg, P. Structural Basis for the Binding of Naproxen to Human Serum Albumin in the Presence of Fatty Acids and the GA Module. Acta Crystallograph. Sect. F Struct. Biol. Cryst. Commun. 2008, 64 (2), 64–69.

48. Kanner, B. I.; Zomot, E. Sodium-Coupled Neurotransmitter Transporters. Chem. Rev. 2008, 108 (5), 1654–1668.

49. Zhou, Z.; Zhen, J.; Karpowich, N. K.; Law, C. J.; Reith, M. E. A.; Wang, D.-N. Antidepressant Specificity of Serotonin Transporter Suggested by Three LeuT-SSRI Structures. Nat. Struct. Mol. Biol. 2009, 16 (6), 652–657.

50. Jacobson, K. A.; Gao, Z.-G. Adenosine Receptors as Therapeutic Targets. Nat. Rev. Drug Discov. 2006, 5 (3), 247–264.

51. Cooke, R. M.; Brown, A. J. H.; Marshall, F. H.; Mason, J. S. Structures of G Protein- Coupled Receptors Reveal New Opportunities for Drug Discovery. Drug Discov. Today 2015.

52. Sabbadin, D.; Ciancetta, A.; Deganutti, G.; Cuzzolin, A.; Moro, S. Exploring the Recognition Pathway at the Human A Adenosine Receptor of the Endogenous Agonist Adenosine Using Supervised Molecular Dynamics Simulations. Med Chem Commun 6,1081-1085, 2015.

53. Lebon, G.; Warne, T.; Edwards, P. C.; Bennett, K.; Langmead, C. J.; Leslie, A. G. W.; Tate, C. G. Agonist-Bound Adenosine A2A Receptor Structures Reveal Common Features of GPCR Activation. Nature 2011, 474 (7352), 521–525.

54. Barelier, S.; Linard, D.; Pons, J.; Clippe, A.; Knoops, B.; Lancelin, J.-M.; Krimm, I. Discovery of Fragment Molecules That Bind the Human Peroxiredoxin 5 Active Site. PLoS ONE 2010, 5 (3), e9744.

55. Krishnamurthy, H.; Gouaux, E. X-Ray Structures of LeuT in Substrate-Free Outward- Open and Apo Inward-Open States. Nature 2012, 481 (7382), 469–474.

56. Kober, A.; Sjöholm, I. The Binding Sites on Human Serum Albumin for Some Nonsteroidal Antiinflammatory Drugs. Mol. Pharmacol. 1980, 18 (3), 421–426.

3.5 Exploring the recognition pathway at the human A_{2A} adenosine receptor of the endogenous agonist adenosine using supervised molecular dynamics simulations

Davide Sabbadin, Antonella Ciancetta, Giuseppe Deganutti, Alberto Cuzzolin and Stefano Moro*

Abstract

Adenosine is a naturally occurring purine nucleoside that exerts a variety of important biological functions through the activation of four G protein-coupled receptor (GPCR) isoforms, namely the A_1, A_2, A_{2B} and A_3 adenosine receptors (ARs). Recently, the X-ray structure of adenosine-bound hA_{2A} AR has been solved, thus providing precious structural details on receptor recognition and activation mechanisms. To date, however, little is still known about the possible recognition pathway the endogenous agonist might go through while approaching the hA_{2A} AR from the extracellular environment. In the present work, we report the adenosine-hA_{2A} AR recognition pathway through the analysis of a series of Supervised Molecular Dynamics (SuMD) trajectories. Interestingly, a possible energetically stable meta-binding site has been detected and characterized.

Introduction

Adenosine is a naturally occurring purine nucleoside that forms primarily from the metabolism of adenosine triphosphate (ATP), both intracellularly and extracellularly[1]. Consequently, the extracellular levels of adenosine are regulated by its synthesis, metabolism, release and uptake[1,2]. Adenosine exerts pleiotropic functions throughout the body. In the central nervous system (CNS), the nucleoside plays important functions, such as modulation of neurotransmitter release, synaptic plasticity and neuroprotection in ischemic, hypoxic and oxidative stress events[1,3,4]. In addition, adenosine plays different roles in a large variety of tissues. In the cardiovascular system, adenosine produces either vasoconstriction or vasodilation of veins and arteries. Moreover, adenosine regulates T cell proliferation and cytokine production, inhibits lipolysis and stimulates bronchoconstriction[1,3,4].

Adenosine mediates its biological effects by recognizing four G protein-coupled receptor (GPCR) isoforms, namely the A_1, A_{2A}, A_{2B} and A_3 adenosine receptors (ARs). Each subtype has a unique pharmacological profile, tissue distribution and effector coupling[1,4]. Considering receptor sequence similarity, among the human ARs (hARs), the most similar are the A_1 and A_3 ARs (49% similarity), and the A_{2A} and A_{2B} ARs (59% similarity). Conversely, the A_1, A_{2A} and A_3 ARs possess relatively high affinity for adenosine whereas the A_{2B} AR shows relatively lower affinity for adenosine, as summarized in Table 1.

Published - Sabbadin, D.; Ciancetta, A.; Deganutti, G.; Cuzzolin, A.; Moro, S. *Med Chem Commun* **2015**, *6* (6), 1081–1085.

Recently, the crystallographic structure of adenosine- bound hA$_{2A}$ AR has been solved (PDB code: 2YDO)[5]. Although this structural data is extremely precious to interpret both receptor recognition and activation mechanisms of the endogenous agonist, little is still known about the possible recognition pathway between adenosine coming from the extracellular environment and the hA$_{2A}$ AR embedded in the cytoplasmic membrane.

Table 1 - Adenosine affinities at the four receptor subtypes

	hA$_1$, K$_i$ (nM)	hA$_{2A}$, K$_i$ (nM)	hA$_{2B}$[a]	hA$_3$
Adnosine	*ca.* 100	310	15,000	290

[a] Data from functional studies. [b] ref. 4

In this context, Supervised Molecular Dynamics (SuMD) has been recently presented as an alternative computational method that allows the exploration of ligand–receptor recognition pathway investigations on a nanosecond (ns) time scale[6]. In addition to speeding up the acquisition of the ligand–receptor recognition trajectory, this approach facilitates the identification and the structural characterization of multiple binding events (such as meta-binding, allosteric, and orthosteric sites) by taking advantage of the all-atom MD simulation accuracy of a GPCR–ligand com- plex embedded into an explicit lipid–water environment[6].

In the present study, in order to better understand how adenosine approaches the orthosteric binding site of the hA$_{2A}$ AR, its recognition pathway has been described through the analysis of a series of SuMD trajectories. Interestingly, a possible energetically stable meta-binding site has been detected and characterized. The meta-binding site concept was introduced several years ago to describe those binding events that chronologically anticipate the orthosteric binding event[7].

Results and discussion

As anticipated, recently the crystallographic structure of adenosine-bound hA$_{2A}$ AR has been solved. The attempt to apply MD methodology to address the problem of ligand dissociation from its receptor is subjected to some limitations. First of all, ligand dissociation dynamics is usually a slow event in comparison to the timescales accessible to current simulation techniques and computer resources. This does not mean necessarily that the actual event of ligand dissociation takes so long, but it is clear that conformational sampling cannot be done effectively in a conventional MD simulation. On the other hand, the recognition process between a ligand and its receptor is a very rare event to describe at the molecular level and, even with the recent GPU-based computing resources, it is necessary to carry out classical molecular dynamics (MD) experiments on a long microsecond time scale[6]. For this

Published - Sabbadin, D.; Ciancetta, A.; Deganutti, G.; Cuzzolin, A.; Moro, S. *Med Chem Commun* **2015**, *6* (6), 1081–1085.

reason, in order to better under- stand how adenosine approaches the orthosteric binding site of the hA_{2A} AR, its recognition pathway was explored using a SuMD study (Video 1).

In particular, following the ligand recognition pathway emerged by the analysis of SuMD trajectories (Fig. 1 and Video 1), the third extracellular loop (EL3) of hA_{2A} AR plays an essential role in directing the agonist toward the orthosteric binding site. In particular, His264, Ala265, Pro266 (EL3) and Leu267 (7.32) (Fig. 1, panel A) establish favourable hydrophobic contacts with the adenine core of adenosine. Such interactions orient the ribose ring towards the entrance of the orthosteric binding site. The hydroxyl group in the C3' position of the ribose ring is engaged in a direct hydrogen bond interaction with Glu169 (EL2). Not surprisingly, the described extracellular site corresponds to the previously reported meta-binding site located in EL3[6,7], which enables high-potency hA_{2A} AR antagonists, such as ZM 241385, 6-IJ2,6- dimethylpyridin-4-yl)-5-phenyl-1,2,4-triazin-3-amine (T4G), and 4-IJ3-amino-5-phenyl-1,2,4-triazin-6-yl)-2-chlorophenol (T4E), to reach the orthosteric binding cleft from the extracellular vestibule. As already described, once the antagonists reach the orthosteric binding site, they adopt binding conformations that match the geometric positions observed in the corresponding X-ray structures[6].

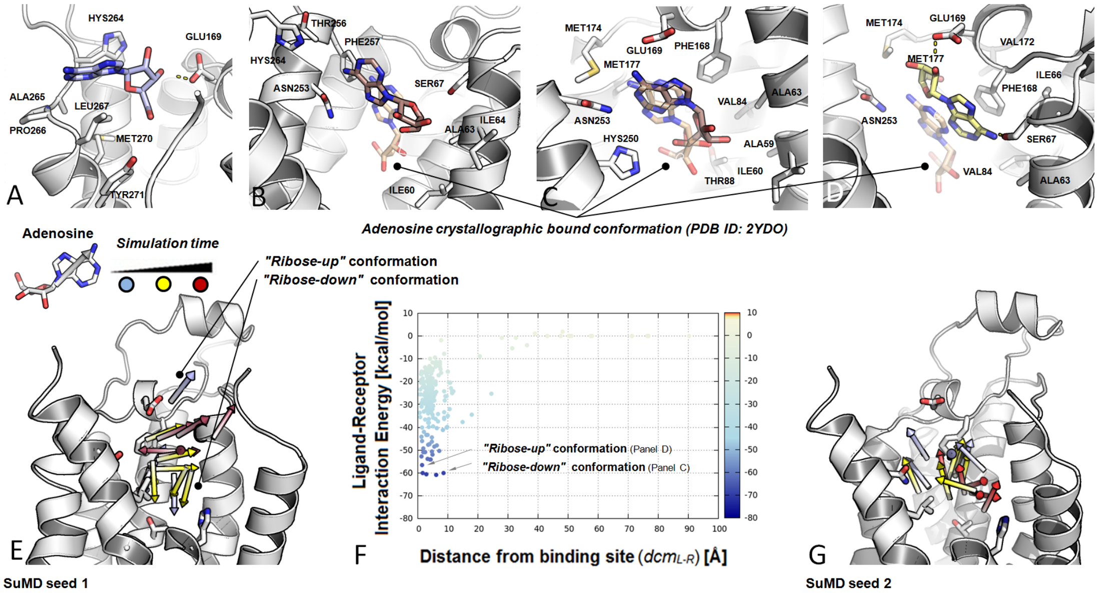


Figure 1 (Panel A to D) - Overview of multiple adenosine binding conformation inside the hA_{2A} AR binding pocket generated from SuMD simulation trajectories in comparison with X-ray crystal structure, PDB ID: 2YDO (wheat sticks). Stick colouring scheme is based on simulation progression (time). Hydrogen atoms are not displayed, whereas hydrogen bond interactions are highlighted as yellow dashed lines. (Panel E and G) Overview of multiple discrete binding states that occur during ligand–receptor recognition. Arrow colouring scheme is based on simulation progression (time). Receptor ribbon representation is viewed from the membrane side facing transmembrane domain 6 (TM6) and transmembrane domain 7 (TM7). (Panel F) Ligand−receptor interaction energy landscape for the nonbiased adenosine-hA_{2A} AR recognition process. The most energetically stable binding conformations of adenosine inside the hA_{2A} AR binding pocket are highlighted by arrow. Interaction energy values are expressed in kcal mol^{-1}.

Published - Sabbadin, D.; Ciancetta, A.; Deganutti, G.; Cuzzolin, A.; Moro, S. *Med Chem Commun* **2015**, *6* (6), 1081–1085.

By approaching the orthosteric binding site, adenosine explores receptor-bound states that only partially overlap – RMSD < 3.5 Å – (Fig. 1, panel B–C) with the crystallographic bound conformation. In such conformational states, the ribose moiety explores the bottom part of the binding pocket ("ribose-down" conformation) and is in close contact with Thr88 (3.36). Glu169 (EL2) and Asn53 (6.55) are involved in key polar interactions with the endo and exocyclic nitrogen atoms of the aromatic core. Hydrophobic interactions are established with Met174 (5.35), Met177 (5.38), Ala59 (2.57), Ala63 (2.61), Val84 (3.32) and Ile160 (El2). In particular, Phe168 (EL2) is involved in π-stacking interaction with the adenine core.

Notably, the role of several key residues (such as Thr88 (3.36), Phe168 (EL2) and Met177 (5.38)) herein highlighted is consistent with the available mutagenesis data for agonist binding, which have been recently analysed and clarified by means of MD/FEP calculations[8].

As reported in Fig. 1, panel D–F, once inside the orthosteric site, adenosine dynamically flips between two different binding modes: the one above reported – the so-called “ribose-down” conformation – and the “ribose-up” conformation (Fig. 1, panel D) where the ribose moiety is directed towards the extracellular space. The hydroxyl group, attached at the C2' position of the ribose ring, establishes a hydrogen- bond interaction with Glu169 (EL2) and the exocyclic nitro- gen atom of the adenine ring interacts with the Ser67 (2.65) side chain. The agonist aromatic ring is involved in a π-stacking interaction with Phe168 (EL2). Val84 (3.32), Ala63 (2.61) and Met174 (5.35) are responsible of the majority of non-polar ligand–receptor contacts.

Therefore, although the described ligand–receptor contacts provide sufficient energetic protein–ligand complex stabilization to reach the global protein–ligand interaction energy minimum (Fig. 1, panel F), the recognition of the agonist is not accompanied by subsequent stabilization of the ligand conformation within the orthosteric site, as adenosine dynamically flips between the “ribose down” and “ribose up” binding modes (Fig. 1, panel E). Therefore, as also elucidated by a clustering analysis of the space explored by adenosine during the binding pathway, the agonist recognition process does not show the same behaviour of potent hA_{2A} AR antagonists. The adenosine binding profile, instead, is more similar to the one observed for a weak binder such as caffeine (Fig. 1, panel F)[6]. Moreover, this peculiar conformational landscape along with the emerged major interaction sites, which anticipate the orthosteric binding site, is independent from ligand placement and orientation at the beginning of the SuMD simulation (Fig. 1, panel G).

Published - Sabbadin, D.; Ciancetta, A.; Deganutti, G.; Cuzzolin, A.; Moro, S. *Med Chem Commun* **2015**, *6* (6), 1081–1085.

Experimental

General

The numbering of the amino acids follows the arbitrary scheme by Ballesteros and Weinstein: each amino acid identifier starts with a helix number, followed by the position relative to a reference residue among the most conserved amino acids in that helix, to which the number 50 is arbitrarily assigned[9].

Trajectory analysis and figure and video generation have been performed using several functionalities implemented by Visual Molecular Dynamics[10], WORDOM[11], the PyMOL Molecular Graphics System, Version 1.5.0.4 Schrödinger, LLC (http://www.pymol.org/) and the Gnuplot graphic utility (http://www.gnuplot.info/). Ligand-hA$_{2A}$ AR interaction energies were calculated by extrapolating the non-bonded energy interaction term of CHARMM27 Force Field[12] using NAMD[13].

Computational facilities

All computations were performed on a hybrid CPU/GPU cluster. Molecular dynamics simulation has been performed with a 2 NVIDIA GTX 680 and 3 NVIDIA GTX 780 GPU cluster engineered by Acellera[14].

Human A$_{2A}$ adenosine receptor–ligand complex preparation

The selected agonist-bound crystal structures (PDB IDs: 2YDO[5]) and the FASTA sequence of the hA$_{2A}$ AR (Uniprot ID: P29274) were retrieved from the RCSB PDB database[15] (http://www.rcsb.org) and the UniProtKB/Swiss-Prot[16,17], respectively. The co-crystallized ligand structure was extracted from the orthosteric binding site and randomly placed in the space above the receptor, at least 40 Å away from protein atoms. Ionization states and hydrogen positions were assigned by using the MOE-sdwash utility (pH 7.0). The FASTA sequence was aligned, using BLAST (Blosum 62 matrix)[18], with the template sequence. Backbone and conserved residue coordinates were copied from the template structure, whereas newly modelled regions and non-conserved residue side chains were modelled and energetically optimized by using CHARMM 27 force field[12] until a r.m.s. of conjugate gradient <0.05 kcal mol^{-1} Å^{-1} was reached. Missing loop domains were constructed by the loop search method implemented in the Molecular Operating Environment (MOE, version 2012.10) program[19] on the basis of the structure of compatible fragments found in the Protein Data Bank. N-terminal and C-terminal were deleted if their lengths exceeded those found in the crystallographic template. The "Protonate-3D" tool[20] was used to appropriately assign ionization states and hydrogen positions to the build models. Then, the structures were subjected to energy minimization with CHARMM 27 force field[12] until the r.m.s. of conjugate gradient was <0.05 kcal mol^{-1} Å^{-1}. Protein stereochemistry evaluation was then performed by employing several tools (Ramachandran and *χ* plots measure *j/ψ* and *χ1/χ2* angles, clash contact reports) implemented in the MOE suite[19].

Receptor membrane embedding and system preparation

Receptors were embedded in a 1-palmitoyl-2-oleoyl-sn-glycero-3 phosphocholine (POPC) lipid bilayer (85 × 85 Å wide) and placed into the membrane according to the suggested orientation reported in the "Orientations of Proteins in Membranes (OPM)" database[21] for the hA_{2A} AR in a complex with the antagonist T4G (PDB ID: 2YDV[7]). Overlapping lipids (within 0.6 Å) were removed upon insertion of the protein. The prepared systems were solvated with TIP3P water[22] using the program Solvate 1.0[23] and neutralized with Na^+/Cl^- counterions to a final concentration of 0.154 M. The total number of atoms per system was approximately 75,000. Membrane MD simulations were carried out on a GPU cluster with the ACEMD program using the CHARMM27 Force Field[18] and periodic boundary conditions. Initial parameters for the ligands were derived from the CHARMM General Force Field for organic molecules[24,25]. The system was equilibrated using a stepwise procedure. In the first stage, to reduce steric clashes due to the manual setting up of the membrane–receptor system, a 500 step conjugate-gradient minimization was performed. Then, to allow lipids to reach equilibrium and water molecules to diffuse into the protein cavity and to avoid ligand–receptor interaction in the equilibration phase, protein and ligand atoms were restrained for the first 8 ns by a force constant of 1 kcal mol^{-1} $Å^{-2}$. Then side chains were set free to move, while gradually reducing the force constant to 0.1 kcal mol^{-1} $Å^{-2}$ to the ligand and alpha carbon atoms up to 9 ns. Temperature was maintained at 298 K using a Langevin thermostat with a low damping constant of 1 ps^{-1}, and the pressure was maintained at 1 atm using a Berendsen barostat. Bond lengths involving hydrogen atoms were constrained using the M-SHAKE algorithm[26] with an integration time step of 2 fs. Harmonical constraints were then removed and Supervised MD was conducted in a NVT ensemble. Long-range Coulomb interactions were handled using the particle mesh Ewald summation method (PME)[27] with grid size rounded to the approximate integer value of cell wall dimensions. A non-bonded cutoff distance of 9 Å with a switching distance of 7.5 Å was used. In order to assess the biophysical validity of the built systems, the average area per lipid headgroup (APL) and bilayer thickness measurements for each built system were measured using Grid-MAT-MD[28]. The corresponding calculated averaged area per lipid headgroup of the extracellular and intracellular leaflet during the production phase for all simulations was in agreement with the experimental values measured for 1-palmitoyl-2- oleoyl-sn-glycero-3-phosphocholine (POPC) lipid bilayers[29].

Conclusions

In the present work, we have carried out SuMD experiments to elucidate the recognition pathway of the naturally occurring purine nucleoside adenosine by the hA_{2A} AR. The analysis of the SuMD trajectories revealed that residues located in the third extracellular loop play an essential role in orienting the ribose ring of agonist toward the entrance of the orthosteric site, thus representing a possible energetically stable meta-binding site.

Our analysis has also revealed that, once the orthosteric site is reached, adenosine experiences a dynamic flip between two different binding modes: the "ribose-down" and the “ribose-up” conformation, with the ribose moiety pointing towards the intracellular and extracellular space, respectively. Consequently, the adenosine binding profile resulting from our analysis resembles that of a weak binder rather than the one previously observed for potent hA_{2A} AR antagonists.

Further work is underway in our lab to better elucidate the role of the meta-binding site that has been detected and characterized in this study. In particular, SuMD simulations with adenosine-hA_{2A} AR 2:1 stoichiometry are currently under evaluation. Moreover, we are carrying out a comprehensive SuMD exploration of the recognition pathway of adenosine against all other adenosine receptor subtypes to clarify the experimental selectivity profile provided by the natural agonist.

Published - Sabbadin, D.; Ciancetta, A.; Deganutti, G.; Cuzzolin, A.; Moro, S. *Med Chem Commun* **2015**, *6* (6), 1081–1085.

Bibliography

1. Jacobson, K. A. & Gao, Z.-G. Adenosine receptors as therapeutic targets. *Nat. Rev. Drug Discov.* **5,** 247–264 (2006).

2. Latini, S. & Pedata, F. Adenosine in the central nervous system: release mechanisms and extracellular concentrations. *J. Neurochem.* **79,** 463–484 (2001).

3. Sebastião, A. M. & Ribeiro, J. A. Fine-tuning neuromodulation by adenosine. *Trends Pharmacol. Sci.* **21,** 341–346 (2000).

4. Fredholm, B. B., IJzerman, A. P., Jacobson, K. A., Linden, J. & Müller, C. E. International Union of Basic and Clinical Pharmacology. LXXXI. Nomenclature and classification of adenosine receptors--an update. *Pharmacol. Rev.* **63,** 1–34 (2011).

5. Lebon, G. *et al.* Agonist-bound adenosine A2A receptor structures reveal common features of GPCR activation. *Nature* **474,** 521–525 (2011).

6. Sabbadin, D. & Moro, S. Supervised molecular dynamics (SuMD) as a helpful tool to depict GPCR-ligand recognition pathway in a nanosecond time scale. *J. Chem. Inf. Model.* **54,** 372–376 (2014).

7. Moro, S., Hoffmann, C. & Jacobson, K. A. Role of the extracellular loops of G protein-coupled receptors in ligand recognition: a molecular modeling study of the human P2Y1 receptor. *Biochemistry (Mosc.)* **1,** 3498–3507

8. Keränen, H., Gutiérrez-de-Terán, H. & Åqvist, J. Structural and Energetic Effects of A2A Adenosine Receptor Mutations on Agonist and Antagonist Binding. *PLoS ONE* **9,** e108492 (2014).

9. Juan A. Ballesteros, H. W. Integrated Methods for the Construction of Three-Dimensional Models and Computational Probing of Structure-Function Relations in G-Protein-Coupled Receptors. *Methods Neurosci.* **25,** 366–428 (1995).

10. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graph.* **14,** 33–38, 27–28 (1996).

11. Seeber, M. *et al.* Wordom: A user-friendly program for the analysis of molecular structures, trajectories, and free energy surfaces. *J. Comput. Chem.* **32,** 1183–1194 (2011).

12. MacKerell, A. D., Banavali, N. & Foloppe, N. Development and current status of the CHARMM force field for nucleic acids. *Biopolymers* **56,** 257–265 (2000).

13. Phillips, J. C. *et al.* Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **26,** 1781–1802 (2005).

Published - Sabbadin, D.; Ciancetta, A.; Deganutti, G.; Cuzzolin, A.; Moro, S. *Med Chem Commun* **2015**, *6* (6), 1081–1085.

14. Acellera. *Acellera* at <https://www.acellera.com/>

15. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28,** 235–242 (2000).

16. UniProt Consortium. The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.* **38,** D142–148 (2010).

17. Jain, E. *et al.* Infrastructure for the life sciences: design and implementation of the UniProt website. *BMC Bioinformatics* **10,** 136 (2009).

18. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402 (1997).

19. Chemical Computing Group. at <http://www.chemcomp.com/>

20. Labute, P. Protonate3D: Assignment of ionization states and hydrogen coordinates to macromolecular structures. *Proteins Struct. Funct. Bioinforma.* **75,** 187–205 (2009).

21. Lomize, M. A., Lomize, A. L., Pogozheva, I. D. & Mosberg, H. I. OPM: orientations of proteins in membranes database. *Bioinforma. Oxf. Engl.* **22,** 623–625 (2006).

22. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79,** 926–935 (1983).

23. Grubmuller, H. & Groll, V. Solvate. (1996). at <http://www.mpibpc.mpg.de/grubmueller/solvate>

24. Vanommeslaeghe, K. & MacKerell, A. D. Automation of the CHARMM General Force Field (CGenFF) I: Bond Perception and Atom Typing. *J. Chem. Inf. Model.* **52,** 3144–3154 (2012).

25. Vanommeslaeghe, K., Raman, E. P. & MacKerell, A. D. Automation of the CHARMM General Force Field (CGenFF) II: assignment of bonded parameters and partial atomic charges. *J. Chem. Inf. Model.* **52,** 3155–3168 (2012).

26. Kräutler, V., van Gunsteren, W. F. & Hünenberger, P. H. A fast SHAKE algorithm to solve distance constraint equations for small molecules in molecular dynamics simulations. *J. Comput. Chem.* **22,** 501–508 (2001).

27. Essmann, U. *et al.* A smooth particle mesh Ewald method. *J. Chem. Phys.* **103,** 8577–8593 (1995).

28. Allen, W. J., Lemkul, J. A. & Bevan, D. R. GridMAT-MD: a grid-based membrane analysis tool for use with molecular dynamics. *J. Comput. Chem.* **30,** 1952–1958 (2009).

29. Kucerka, N., Tristram-Nagle, S. & Nagle, J. F. Structure of fully hydrated fluid phase lipid bilayers with monounsaturated chains. *J. Membr. Biol.* **208,** 193–202 (2005).

Published - Sabbadin, D.; Ciancetta, A.; Deganutti, G.; Cuzzolin, A.; Moro, S. *Med Chem Commun* **2015**, *6* (6), 1081–1085.

3.6 Understanding allosteric interactions in G protein-coupled receptors using Supervised Molecular Dynamics: a prototype study analysing the human A_3 adenosine receptor positive allosteric modulator LUF6000

Giuseppe Deganutti, Alberto Cuzzolin, Antonella Ciancetta, Stefano Moro*

Abstract

The search for G protein-coupled receptors (GPCRs) allosteric modulators represents an active research field in medicinal chemistry. Allosteric modulators usually exert their activity only in the presence of the orthosteric ligand by binding to protein sites topographically different from the orthosteric cleft. They therefore offer potentially therapeutic advantages by selectively influencing tissue responses only when the endogenous agonist is present. The prediction of putative allosteric site location, however, is a challenging task. In facts, they are usually located in regions showing more structural variation among the family members. In the present work, we applied the recently developed Supervised Molecular Dynamics (SuMD) methodology to interpret at the molecular level the positive allosteric modulation mediated by LUF6000 toward the human adenosine A_3 receptor (hA_3 AR). Our data suggest at least two possible mechanisms to explain the experimental data available. This study represent, to the best of our knowledge, the first case reported of an allosteric recognition mechanism depicted by means of molecular dynamics simulations.

Introduction

Besides the orthosteric site, which conventionally recognizes endogenous ligands, most G protein-coupled receptors (GPCRs) possess topographically distinct allosteric sites that can be recognized by small molecules and accessory cellular proteins. Pharmacologically speaking, an allosteric modulator does not have any activity by itself, thus needing the orthosteric binder to exhibit its action. Although the modulatory character of allosteric binders is not always clear-cut, true allosteric modulators increase or decrease the action of an agonist or an antagonist recognising the allosteric site(s) on the receptor. In facts, ligand binding to allosteric sites promotes a conformational reorganization in the GPCR that can alter orthosteric ligand affinity and/or efficacy. Although an allosteric modulator may not possess efficacy by itself, it can provide a powerful therapeutic advantage over orthosteric ligands, as they selectively influence tissue responses only when the endogenous agonist is present. Consequently, allosteric modulation of GPCRs has stimulated an intensive identification campaign for new classes of hit-candidates

Published - Deganutti, G.; Cuzzolin, A.; Ciancetta, A.; Moro, S. *Bioorg. Med. Chem.* **2015**, *23* (14), 4065–4071.

different from conventional agonists and antagonists. This has been the subject of several recent reviews[1–3].

However, natural allosteric sites are very difficult to identify because they are usually located far from the orthosteric sites. Moreover, allosteric sites resides in regions of the receptor that show more structural variation among family members and, consequently, this implies a general lack of success in predicting the locations of potential binding regions. Albeit the crystallographic structure of the M2 receptor simultaneously bound to the orthosteric agonist iperoxo and the positive allosteric modulator LY2119620 has been recently reported[4], little is known about the possible allosteric control regarding the activation mechanism of other GPCRs.

Within this framework, we have recently reported on an alternative computational method – the Supervised Molecular Dynamics (SuMD) – that allows to investigate the ligand– receptor recognition pathway in a nanosecond (ns) time scale[5]. In addition to speeding up the acquisition of the ligand–receptor recognition trajectory, this approach facilitates the identification and the structural characterization of multiple binding events (such as meta-binding, allosteric, and orthosteric sites) by taking advantage of the all-atom MD simulations accuracy of GPCR– ligand complexes embedded into explicit lipid–water environment[5].

Interestingly, adenosine receptors (ARs) were among the first GPCRs discovered to be allosterically regulated and, in particular, allosteric enhancers for A1 and A3 ARs have been widely investigated[1,2,6]. Among the most interesting allosteric enhancers for the A3 AR, *N*-(3,4-dichlorophenyl)-2-cyclohexyl- 1*H*-imidazo[4,5-*c*]quinolin-4-amine (LUF6000, see Fig. 1) has been deeply characterized[7,8]. LUF6000 potentiates the maximum efficacy of the agonist Cl-IB-MECA by 45–50%, enhances agonist efficacy in functional assays and decreases the agonist dissociation rate without influencing agonist potency. Moreover, LUF6000 has been reported to act as allosteric enhancer of the maximal effect exerted by structurally diverse agonists at the A_3 AR, being more effective for low-efficacy than for high-efficacy agonists.

Very recently, in vivo studies have reported the ability of LUF6000 to act as allosteric modulator of rat and mice A_3 ARs by allowing the endogenous ligand adenosine to bind to the receptor with higher affinity[9].

With the aim to interpret at the molecular level the positive allosterism mediated by LUF6000 toward the human A_3 AR (hA_3 AR), possible recognition pathways have been explored by performing SuMD simulations in the absence and in presence of the natural agonist adenosine (Fig. 1). Interestingly, our results suggest two possible mechanisms by which LUF6000 might exert its positive allosteric modulator effects: according to the outcomes of our simulations, the ligand might either induce a loop rearrangement that stabilizes agonist placement into the orthosteric site, or form a ternary complex with the agonist bound receptor state, thus acting as orthosteric pocket cap.

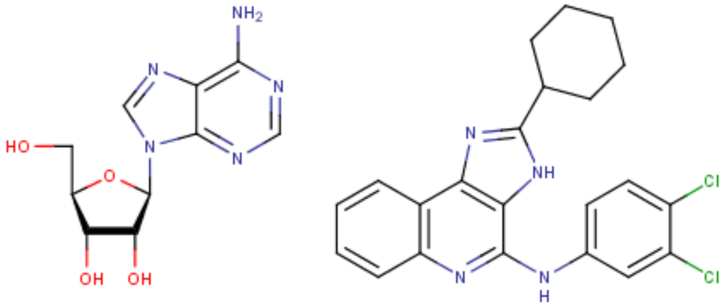


Figure 1 – Structures of the endogenous hA3 AR agonist adenosine (left) and the positive allosteric modulator LUF6000 (right).

SuMD Simulations

LUF6000-hA$_3$ AR recognition mechanism.

The imidazoquinolinamine allosteric modulator LUF6000 enhances agonist efficacy in functional assays and decreases agonist dissociation rate without influencing agonist potency[7,8]. Besides, LUF6000 presents a weak antagonist activity (ca. 45% inhibition at 10 μM)[7]. To explore LUF6000 attitude to recognize the orthosteric binding site of the hA$_3$ AR, we analysed its recognition pathway by performing SuMD experiments.

During the SuMD simulations, LUF6000 reached the orthosteric binding site in less than 20 ns. The corresponding energy landscape (Fig. 2A) highlights two major interaction sites (**a** and **b**). Prior reaching the orthosteric site, LUF6000 interacts with residues located in a region between the second and third extracellular loops (EL2 and EL3, respectively). Met151 (EL3), Thr154 (EL2), Met174 (5.35) side chains and the aliphatic portion of Arg173 (EL2) establish hydrophobic contacts with the imidazoloquinoline core of LUF6000, whereas the ligand exocyclic nitrogen atom is involved in a hydrogen bond interaction with the backbone of Lys152 (EL2) (**a** in Fig. 2A, Fig. 2B, Video S1). While reaching the orthosteric site, the Val169 (EL2) side chain facilitates ligand reorientation providing favourable hydrophobic contacts. The most stable conformation observed (**b** Fig. 2A, Fig. 2C, Video S1) is characterized by hydrophobic contacts with Phe168 (EL2), Met174 (5.35), Leu246 (6.51), Leu264 (7.35), Leu268 (7.39) and Trp243 (6.48), whereas Asn250 (6.55) is engaged in a hydrogen bond with the exocyclic nitrogen of the ligand. At the maximum energetic stabilization, the formed complex is characterized by energetic values of about -50 kcal/mol, a value previously observed for weak ARs binders. This is consistent with the LUF6000 antagonist activity observed through functional assays at the hA$_3$ AR[8].

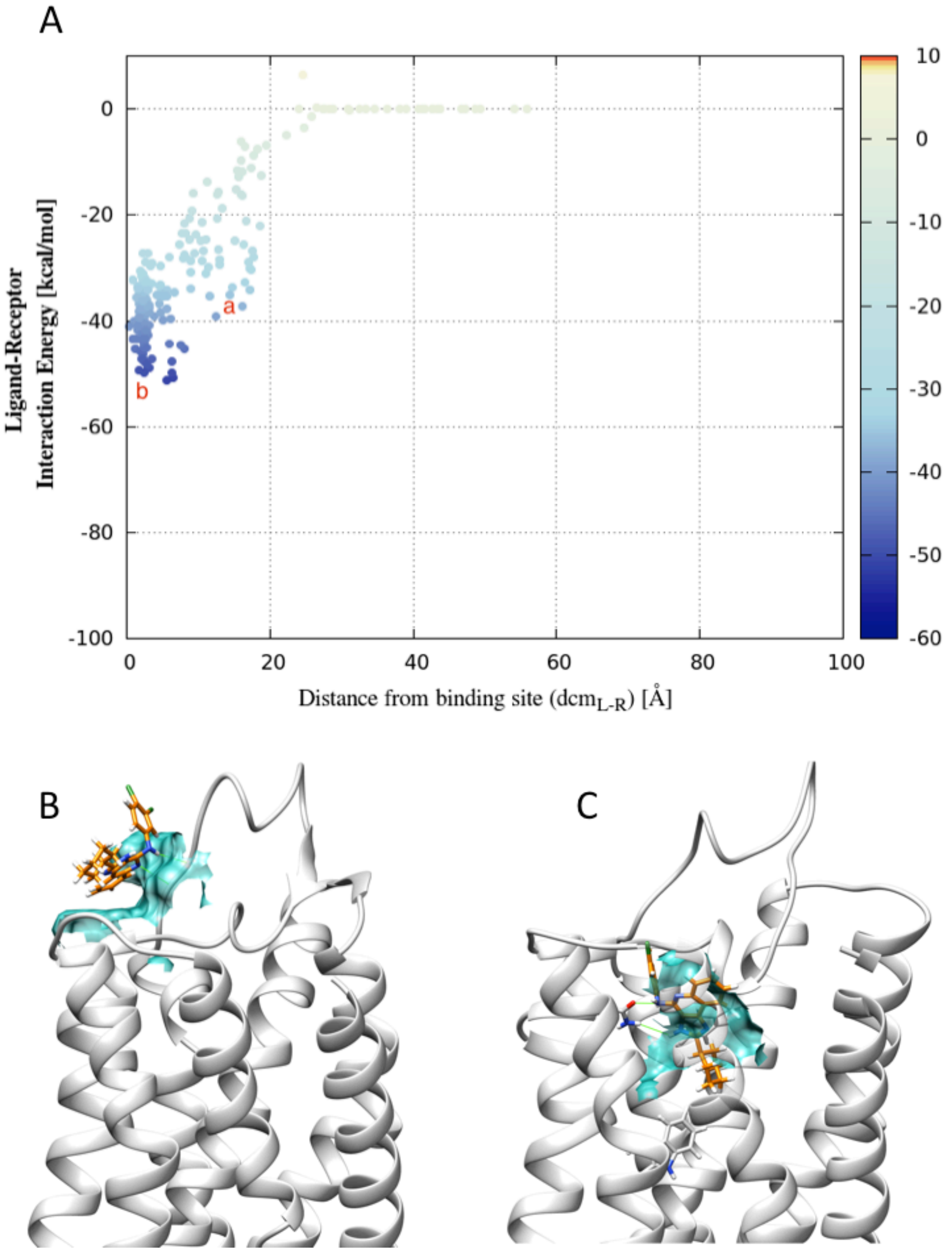


Figure 2 – (A) Interaction Energy landscape for the recognition pattern of LUF6000 by the hA$_3$ AR.**(B)** LUF6000 binding mode in the meta-binding site.**(C)** LUF6000 binding mode in the orthosteric binding site. Ligand is displayed as orange stick, side chains of residues interacting through hydrogen bond or π-π stacking are depicted as grey stick, whereas side chains of residues interacting through hydrophobic contacts are rendered as coloured surfaces.

Published - Deganutti, G.; Cuzzolin, A.; Ciancetta, A.; Moro, S. *Bioorg. Med. Chem.* **2015**, *23* (14), 4065–4071.

Adenosine-hA$_3$ AR recognition mechanism

Recently, the crystallographic structure of adenosine in complex with the hA$_{2A}$ AR has been solved[10]. This structural piece of information aid elucidating both the recognition and activation mechanisms of the ARs by their endogenous agonist adenosine. Although in principle the interaction pattern of the adenosine-hA$_{2A}$ AR complex can be transferred to the other ARs subtypes, in order to better understand how adenosine approaches the orthosteric binding site of the hA$_3$ AR, its recognition pathway as described by the obtained SuMD trajectories has been analysed.

In our SuMD experiments, adenosine reached the orthosteric binding site in less than 20 ns. The corresponding energy landscape is reported in Fig. 3A. During the recognition pathway (Video S2), EL3 engages the ligand ribose moiety in favourable hydrogen bonds mainly through Val259 (EL3) and Gln261 (EL2) backbone atoms (**b** in Fig. 3A, Fig. 3B). This situation anticipates a change of adenosine orientation, triggered by hydrophobic contacts between the ligand purine core and Leu264 (7.35), Ile268 (7.39), Ile253 (6.58), and Ile249 (6.54) side chains. Once the orthosteric pocket is reached, adenosine interacts with the side chain of Trp185 (5.46), Leu246 (6.51) and conserved residues Asn250 (6.55), Phe168 (EL2), Trp243 (6.48), Ile268 (7.39) (Fig. 3C, Video S2). As already observed for adenosine recognition pathway by the hA$_{2A}$ AR[11], the agonist explores different conformational states once inside the pocket. In particular, adenosine experiences a dynamic flip between two different binding modes: the above described "ribose-down" and the "ribose-up" conformation, with the ribose moiety pointing towards the intracellular and extracellular space, respectively. The ribose down conformation (**b** in Fig. 3A, Fig. S1) is characterized by additional electrostatic interactions with Glu19 (1.39) and Ser73 (2.65) and represents the most energetically stable ligand-receptor complex observed in the analysed trajectories.

LUF6000-hA$_3$ AR (in complex with adenosine) recognition mechanism

With the aim to reproduce the experimental conditions that allow to measure LUF6000 PAM activity, SuMD simulations were performed considering the hA$_3$ AR receptor in complex with adenosine. LUF6000 was randomly placed 60 Å at least away from the barycentre of the orthosteric binding site. The starting adenosine-hA$_3$ AR complex was extracted from the previously described SuMD trajectory, and selected on the basis of its similarity with the X-Ray crystallographic conformation observed for the complex with the hA$_{2A}$ AR[10].

The LUF6000 recognition energy landscape is reported in Fig. 4A. The pathway described by the SuMD trajectories highlights three main situations: *i)* LUF6000 not interacting with the adenosine-hA$_3$ AR complex (point a in Fig.4A); *ii)* LUF6000 interacting with a meta-binding site (**b** in Fig. 4A); and *iii)* LUF6000 interacting with the orthosteric pocket (**c** in Fig. 4A).

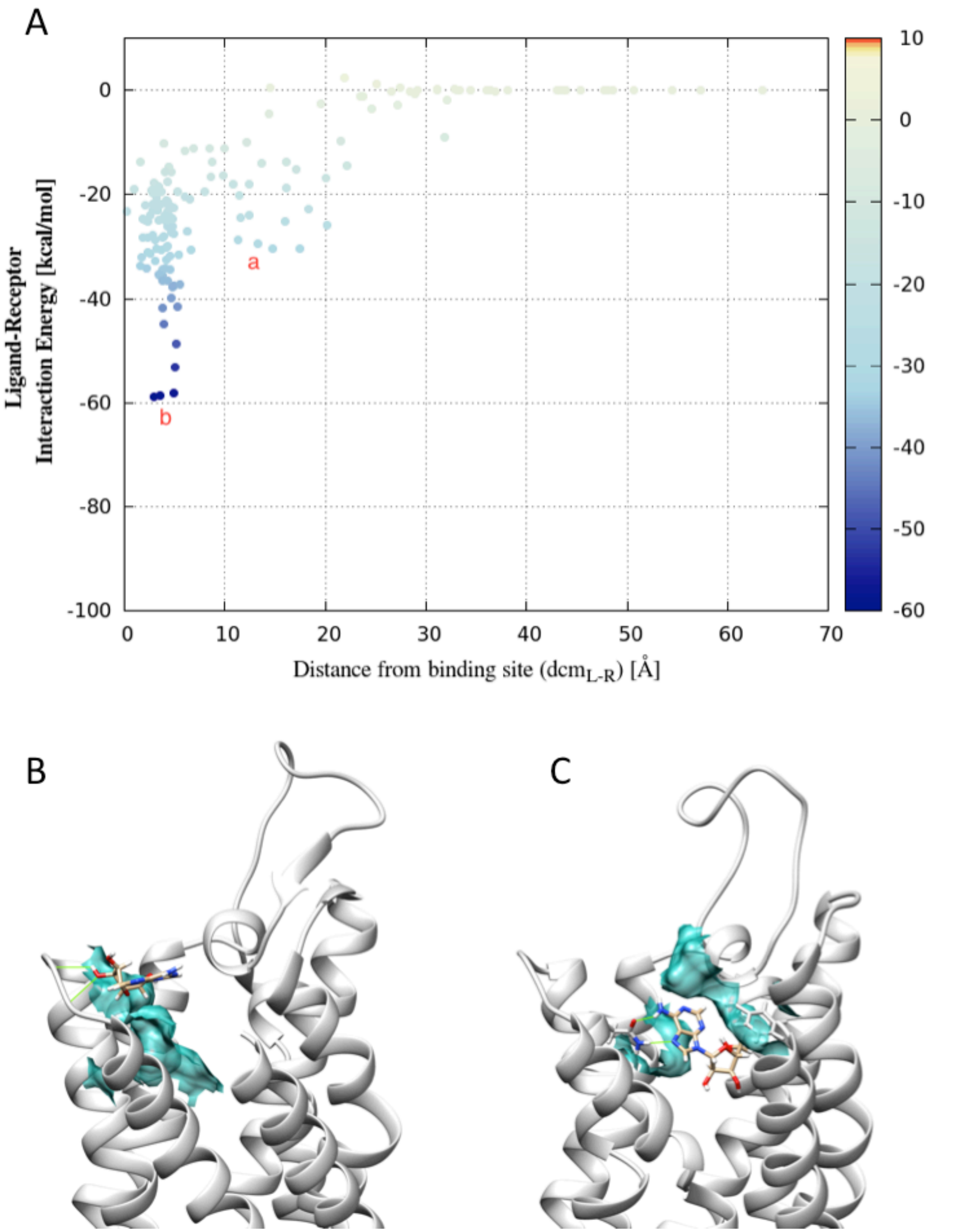


Figure 3 – (A) Interaction Energy landscape for the recognition pattern of adenosine by the hA_3 AR.**(B)** Adenosine binding mode in the meta-binding site.**(C)** Adenosine binding mode in the orhtosteric binding site. Ligand is displayed as tan stick, side chains of residues interacting through hydrogen bond or π-π stacking are depicted as grey stick, whereas side chains of residues interacting through hydrophobic contacts are rendered as coloured surfaces

Published - Deganutti, G.; Cuzzolin, A.; Ciancetta, A.; Moro, S. *Bioorg. Med. Chem.* **2015**, *23* (14), 4065–4071.

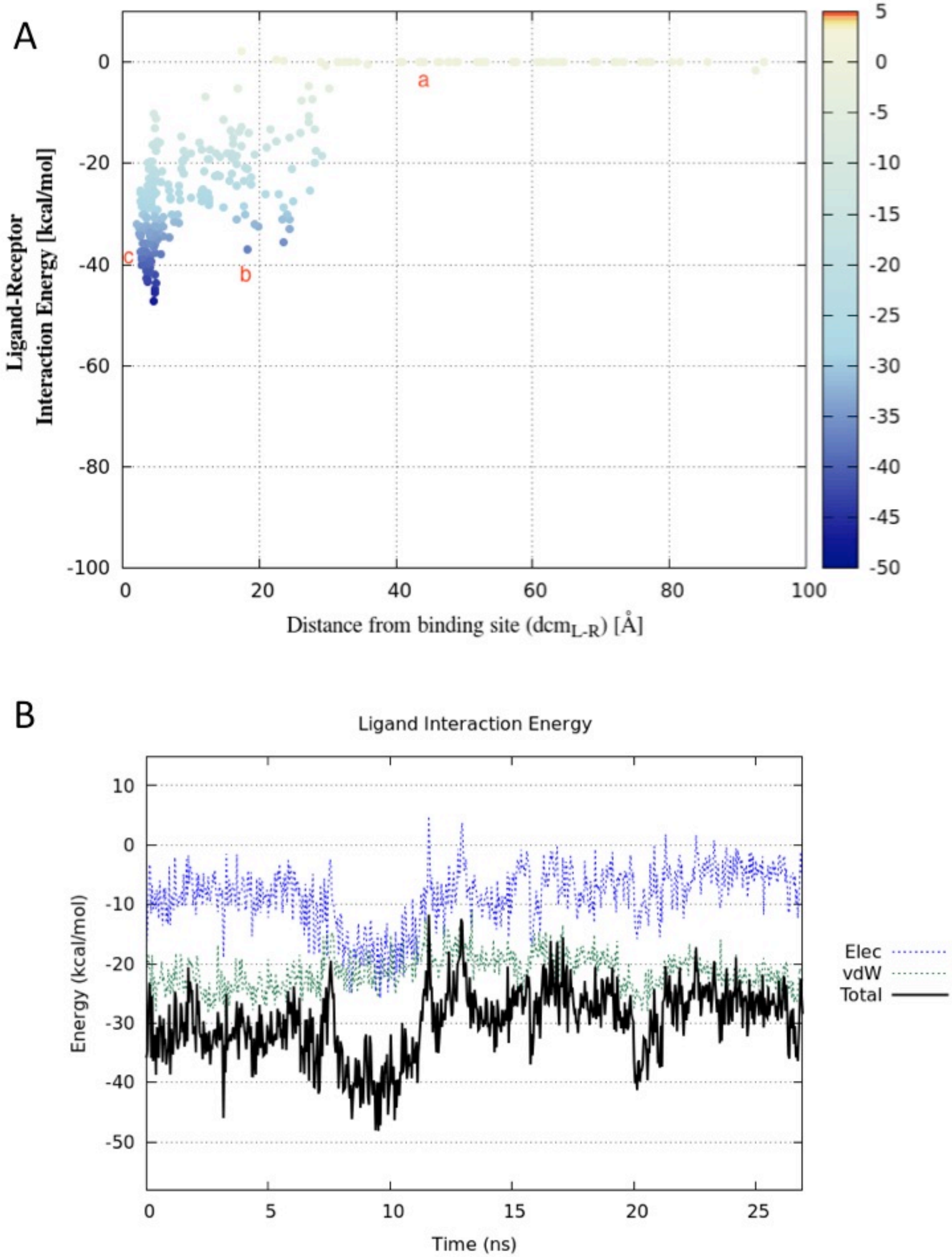


Figure 4 – (A) Interaction Energy landscape for the recognition pattern of LUF6000 by the hA_3 AR-adenosine complex. **(B)** hA_3 AR-adenosine interaction energy.

Published - Deganutti, G.; Cuzzolin, A.; Ciancetta, A.; Moro, S. *Bioorg. Med. Chem.* **2015**, *23* (14), 4065–4071.

Moreover, during the SuMD simulations, the interaction energy between adenosine and the hA_3 AR has been computed (Fig. 4B).

The endogenous agonist reaches a stability maximum approximately after 10 ns of simulation (point B in Fig. 4B), which correspond to LUF6000-hA_3 AR meta binding site complex formation. The ligand approaches EL2 (Fig. 5A) by recruiting Tyr157 (EL2) and His158 (EL2) side chains, and establishing contacts with Arg173 (EL2), Met174 (5.35), Ile253 (6.58), Tyr254 (6.59) (Fig. 4B, Video S3). This loop rearrangement is accompanied by conformational changes in residues located farther in EL2, included Phe168 (EL2), that loses the capability of stabilizing adenosine through π-π stacking interactions. As a consequence, adenosine moves deeper in the orthosteric pocket (Fig. S2) and establishes favourable interactions with Leu246 (6.51), Trp243 (6.48), Ser247 (6.52), and Thr94 (3.36). The overall protein conformational and adenosine positional changes that occur after the interaction between LUF6000 and the hA_3 AR EL2 (*i.e.* system evolution from point a to point b) are reported in Fig. S3.

Once the complex with the meta binding site is formed, approximately after 14 ns of simulations, LUF6000 moves to establish hydrophobic interactions with Tyr254 (6.59), Met174 (5.35), Val169 (EL2) and Ile253 (6.58), located at the top of orthosteric binding site. This allows the ligand to directly interact with adenosine. Simultaneously, the previously evidenced π-π stacking interaction between adenosine and Phe168 (EL2) is restored. In the tertiary complex just formed, the energy interaction between adenosine and the hA_3 AR is stabilized at values slightly lower than the starting complex, with the exception of a transitory stabilization after 21 ns (C in Fig. 4B, Fig. 5B). LUF6000 is therefore able to stabilize the interaction energy between adenosine and the hA_3 AR and to lock the agonist inside the orthosteric pocket for the remaining simulation time. A similar behaviour has been observed also for the tertiary complex formed with a LUF6000 close analogue LUF6069 (See Supplementary Information, Fig. S4-S5 and Video S4).

Conclusion

In the present work, we have utilized SuMD[5], a computational approach we have recently developed, with the aim to characterize and rationalize the activity of LUF6000, a hA_3 AR PAM, at a molecular level. We have analysed the ligand-receptor recognition pattern, both for LUF6000 and the endogenous agonist adenosine separately and also considering the recognition pathway of the PAM by the hA_3 AR in complex with adenosine. This represent, to date, the first case reported of an allosteric mechanism investigated by means of MD simulations.

Our results have highlighted that LUF6000 is able to establish favourable interactions with conserved residues located in the orthosteric binding site of the hA_3 AR, consistently with the experimentally observed weak inhibitor activity at this receptor subtype. The analysis of the interaction pathway of the endogenous agonist adenosine suggests a key role played by residues located in the EL2 in engaging the agonist and energetically promoting its approach to the orthosteric pocket.

The inspection of the interaction pathway obtained by simulating LUF6000 approaching the hA_3 AR in complex with the endogenous agonist adenosine suggests two possible mechanisms to explain the experimentally observed positive allosteric modulation[7,8]. According to our analysis, the ligand could: *i)* trigger conformational changes in the EL2 that would enable the agonist to form more energetically favourable interactions with residues located deeper in the orthosteric binding site; *ii)* establish a ternary complex with the agonist and the receptor, thus acting as orthosteric pocket cap.

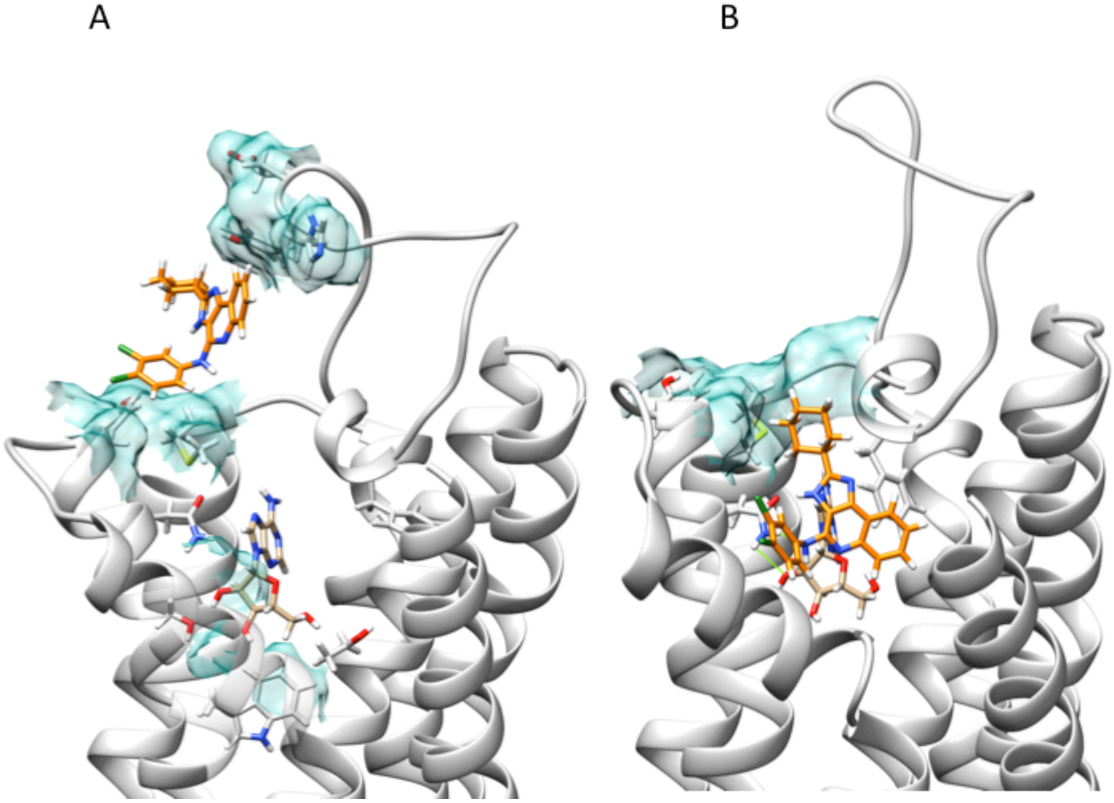


Figure 5 – (A) LUF6000 binding mode in the hA_3 AR meta-binding site.**(B)** LUF6000 binding mode in the hA_3 AR orthosteric binding site occupied by adenosine. LUF6000 and adenosine are displayed as orange and tan stick, respectively. Side chains of residues interacting through hydrogen bond or π-π stacking are depicted as grey stick, whereas side chains of residues interacting through hydrophobic contacts are rendered as coloured surfaces.

Published - Deganutti, G.; Cuzzolin, A.; Ciancetta, A.; Moro, S. *Bioorg. Med. Chem.* **2015**, *23* (14), 4065–4071.

The mutagenesis data available to date[12] apparently confute the first hypothesis, as it has been reported that the mutation of some residues located in the upper region of the receptor does not affect the allosteric activity of the imidazoquinoline compound DU124183 and the pyridinylisoquinoline compound VUF5455[10].

However, it is well accepted that a PAM activity is strictly depending on the structure of the agonist considered to perform the experiments.

Experimental Section

General

All computations were performed on a hybrid CPU/GPU cluster. Molecular dynamics simulation have been performed with GPU cluster equipped with 3 NVIDIA GTX 780 and 3 NVIDIA GTX 980.

Trajectory analysis, Figures and videos generation have been performed using several functionalities implemented by Visual Molecular Dynamics[13], WORDOM[14], the PyMOL Molecular Graphics System, Version 1.5.0.4 Schrödinger, LLC (http://www.pymol.org/) and the Gnuplot graphic utility (http://www.gnuplot.info/). Ligand-hA$_3$ AR interaction energies were calculated extrapolating the non-bonded energy interaction term of CHARMM27 Force Field[15] using NAMD[16]. All molecular dynamics simulations have been carried out using ACEMD engine (http://www.acellera.com/).

The numbering of the amino acids follows the arbitrary scheme proposed by Ballesteros and Weinstein[17]: each amino acid identifier starts with the helix number (1-7), followed by a dot and the position relative to a reference residue among the most conserved amino acids in that helix, to which the number 50 is arbitrarily assigned.

Homology Model of hA$_3$ AR

As, to date, no crystallographic information about the hA$_3$ AR is available, we used a previously build homology model deposited in our web platform dedicated to ARs, Adenosiland[18,19]. In particular, among all the currently available crystallographic structures of the hA$_{2A}$ AR we selected the model built upon the complex with the endogenous agonist adenosine (PDB code: 2YDO, 3.00 Å resolution)[10].

Receptor membrane embedding and system preparation.

Receptors were embedded in a 1-palmitoyl-2-oleoyl-sn- glycero-3-phosphocholine (POPC) lipid bilayer (85x95 Å wide) and placed into the membrane according to the suggested orientation reported in the “Orientations of Proteins in Membranes (OPM)” database[20] for the hA$_{2A}$ AR in complex with the endogenous agonist adenosine (PDB ID: 2YDO)[10]. Overlapping lipids (within 0.6 Å) were removed upon insertion of the protein. The prepared systems were solvated with TIP3P water[21] using the program Solvate 1.0[22] and neutralized by Na^+/Cl^- counter-ions to a final concentration of 0.154 M. The total number of atoms per system was approximately 110000. Membrane MD simulations were carried out on a GPU cluster with the ACEMD program using the CHARMM27 Force Field[15] and periodic boundaries conditions. Initial parameters for the ligands were derived from the CHARMM General Force Field for organic molecules. The system was equilibrated using a stepwise procedure. In the first stage, to reduce steric clashes due to the manual setting up of the membrane-receptor system, a 2500 steps conjugate-gradient minimization was performed. Then, to allow lipids to reach equilibrium, water molecules to diffuse into the protein cavity and to avoid ligand-receptor interaction in the equilibration phase, protein and ligand atoms were restrained for the first 8 ns by a force constant of 1 kcal/mol•$Å^2$. Then, force constant was gradually reduced to 0.1 kcal/mol•$Å^2$ for the next 9 ns. Temperature was maintained at 298 K using a Langevin thermostat with a low damping constant of 1 ps^{-1} and the pressure was maintained at 1 atm using a Berendensen barostat. Bond lengths involving hydrogen atoms were constrained using the M-SHAKE algorithm[23] with an integration timestep of 2 fs.

SuMD simulations

After the equilibration procedure, harmonical constraints were removed and SuMD simulations were conducted in a NVT ensemble. As previously described, the supervision of the trajectory is perpetuated until ligand-receptor distance is lower than 5 Å without introducing bias to the simulations. Long-range Coulomb interactions were handled using the particle mesh Ewald summation method (PME)[24] with grid size rounded to the approximate integer value of cell wall dimensions. A non-bonded cut-off distance of 9 Å with a switching distance of 7.5 Å was used. Ligand parametrization procedure and methodological insights on the quantitative estimate of the electrostatic and hydrophobic occurring ligand-protein interaction maps have been reported previously[5].

Abbreviations

ARs	adenosine receptors
ATP	adenosine triphosphate
hA3AR	human A3 adenosine receptor
EL2	second extracellular loop
EL3	third extracellular loop
LUF6000	*N*-(3,4-dichlorophenyl)-2-cyclohexyl-1*H*-imidazo[4,5-*c*]quinolin-4-amine
LUF6096	*N*-{2-[(3,4- dichlorophenyl)amino]quinolin-4-yl}cyclohexanecarboxamide
GPCRs	G protein-coupled receptors
GPU	graphics processing unit
PAM	positive allosteric modulator
RMSD	root mean square deviation
SAR	structure-affinity relationship;
TM	transmembrane; ZM 241385, 4-[2- [7-amino-2-(2-furyl)-1,2,4-triazolo[1,5-*a*][1,3,5]triazin-5-yl-amino]ethylphenol.

Published - Deganutti, G.; Cuzzolin, A.; Ciancetta, A.; Moro, S. *Bioorg. Med. Chem.* **2015**, *23* (14), 4065–4071.

Bibliography

1. Gao, Z.-G. & Jacobson, K. A. Keynote review: allosterism in membrane receptors. *Drug Discov. Today* **11,** 191–202 (2006).
2. Lewis, J. A., Lebois, E. P. & Lindsley, C. W. Allosteric modulation of kinases and GPCRs: design principles and structural diversity. *Curr. Opin. Chem. Biol.* **12,** 269–280 (2008).
3. Keov, P., Sexton, P. M. & Christopoulos, A. Allosteric modulation of G protein-coupled receptors: a pharmacological perspective. *Neuropharmacology* **60,** 24–35 (2011).
4. Kruse, A. C. *et al.* Activation and allosteric modulation of a muscarinic acetylcholine receptor. *Nature* **504,** 101–106 (2013).
5. Sabbadin, D. & Moro, S. Supervised molecular dynamics (SuMD) as a helpful tool to depict GPCR-ligand recognition pathway in a nanosecond time scale. *J. Chem. Inf. Model.* **54,** 372–376 (2014).
6. Göblyös, A. & Ijzerman, A. P. Allosteric modulation of adenosine receptors. *Biochim. Biophys. Acta* **1808,** 1309–1318 (2011).
7. Kim, Y., de Castro, S., Gao, Z.-G., Ijzerman, A. P. & Jacobson, K. A. Novel 2- and 4-substituted 1H-imidazo[4,5-c]quinolin-4-amine derivatives as allosteric modulators of the A3 adenosine receptor. *J. Med. Chem.* **52,** 2098–2108 (2009).
8. Gao, Z.-G., Ye, K., Göblyös, A., Ijzerman, A. P. & Jacobson, K. A. Flexible modulation of agonist efficacy at the human A3 adenosine receptor by the imidazoquinoline allosteric enhancer LUF6000. *BMC Pharmacol.* **8,** 20 (2008).
9. Cohen, S. *et al.* A3 Adenosine Receptor Allosteric Modulator Induces an Anti-Inflammatory Effect: In Vivo Studies and Molecular Mechanism of Action. *Mediators Inflamm.* **2014,** (2014).
10. Lebon, G. *et al.* Agonist-bound adenosine A2A receptor structures reveal common features of GPCR activation. *Nature* **474,** 521–525 (2011).
11. Sabbadin, D., Ciancetta, A., Deganutti, G., Cuzzolin, A. & Moro, S. Exploring the recognition pathway at the human A2A adenosine receptor of the endogenous agonist adenosine using supervised molecular dynamics simulations. *MedChemComm* **6,** 1081–1085 (2015).
12. Gao, Z.-G. *et al.* Identification of essential residues involved in the allosteric modulation of the human A(3) adenosine receptor. *Mol. Pharmacol.* **63,** 1021–1031 (2003).

13. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graph.* **14,** 33–38, 27–28 (1996).

14. Seeber, M. *et al.* Wordom: A user-friendly program for the analysis of molecular structures, trajectories, and free energy surfaces. *J. Comput. Chem.* **32,** 1183–1194 (2011).

15. MacKerell, A. D., Banavali, N. & Foloppe, N. Development and current status of the CHARMM force field for nucleic acids. *Biopolymers* **56,** 257–265 (2000).

16. Phillips, J. C. *et al.* Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **26,** 1781–1802 (2005).

17. Juan A. Ballesteros, H. W. Integrated Methods for the Construction of Three-Dimensional Models and Computational Probing of Structure-Function Relations in G-Protein-Coupled Receptors. *Methods Neurosci.* **25,** 366–428 (1995).

18. Floris, M., Sabbadin, D., Medda, R., Bulfone, A. & Moro, S. Adenosiland: walking through adenosine receptors landscape. *Eur. J. Med. Chem.* **58,** 248–257 (2012).

19. Floris, M. *et al.* Implementing the 'Best Template Searching' tool into Adenosiland platform. *Silico Pharmacol.* **1,** 25 (2013).

20. Lomize, M. A., Lomize, A. L., Pogozheva, I. D. & Mosberg, H. I. OPM: orientations of proteins in membranes database. *Bioinforma. Oxf. Engl.* **22,** 623–625 (2006).

21. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79,** 926–935 (1983).

22. Grubmuller, H. & Groll, V. Solvate. (1996). at <http://www.mpibpc.mpg.de/grubmueller/solvate>

23. Kräutler, V., van Gunsteren, W. F. & Hünenberger, P. H. A fast SHAKE algorithm to solve distance constraint equations for small molecules in molecular dynamics simulations. *J. Comput. Chem.* **22,** 501–508 (2001).

24. Essmann, U. *et al.* A smooth particle mesh Ewald method. *J. Chem. Phys.* **103,** 8577–8593 (1995).

3.7 ALK Kinase Domain Mutations in Primary Anaplastic Large Cell Lymphoma: Consequences on NPM-ALK Activity and Sensitivity to Tyrosine Kinase Inhibitors

Federica Lovisa*, Giorgio Cozza, Andrea Cristiani, Alberto Cuzzolin, Alessandro Albiero, Lara Mussolin, Marta Pillon, Stefano Moro, Giuseppe Basso, Angelo Rosolen[†] Paolo Bonvini

Abstract

ALK inhibitor crizotinib has shown potent antitumor activity in children with refractory Anaplastic Large Cell Lymphoma (ALCL) and the opportunity to include ALK inhibitors in first- line therapies is oncoming. However, recent studies suggest that crizotinib-resistance mutations may emerge in ALCL patients. In the present study, we analyzed ALK kinase domain mutational status of 36 paediatric ALCL patients at diagnosis to identify point mutations and gene aberrations that could impact on NPM-ALK gene expression, activity and sensitivity to small-molecule inhibitors. Amplicon ultra-deep sequencing of ALK kinase domain detected 2 single point mutations, R335Q and R291Q, in 2 cases, 2 common deletions of exon 23 and 25 in all the patients, and 7 splicing-related INDELs in a variable number of them. The functional impact of missense mutations and INDELs was evaluated. Point mutations were shown to affect protein kinase activity, signalling output and drug sensitivity. INDELs, instead, generated kinase-dead variants with dominant negative effect on NPM-ALK kinase, in virtue of their capacity of forming non-functional heterocomplexes. Consistently, when co-expressed, INDELs increased crizotinib inhibitory activity on NPM-ALK signal processing, as demonstrated by the significant reduction of STAT3 phosphorylation. Functional changes in ALK kinase activity induced by both point mutations and structural rearrangements were resolved by molecular modelling and dynamic simulation analysis, providing novel insights into ALK kinase domain folding and regulation. Therefore, these data suggest that NPM-ALK pre-therapeutic mutations may be found at low frequency in ALCL patients. These mutations occur randomly within the ALK kinase domain and affect protein activity, while preserving responsiveness to crizotinib.

Introduction

Anaplastic Large Cell Lymphoma (ALCL) represents a distinct subset of T-cell non-Hodgkin lymphoma (NHL), accounting for about 10–15% of childhood lymphomas[1]. The relative rarity of this tumour has limited the number of large prospective clinical trials for treatment optimization, and current therapeutic strategies are still based on the use of combined intensive chemotherapy. Despite current treatments achieve an event-free survival around 75%, the out- come of

Published - Lovisa, F.; Cozza, G.; Cristiani, A.; Cuzzolin, A.; Albiero, A.; Mussolin, L.; Pillon, M.; Moro, S.; Basso, G.; Rosolen, A.; Bonvini, P. *PloS One* **2015**, *10* (4), e0121378.

relapsed patients is less than 60%[2] and more effective therapeutic strategies are demanding.

Anaplastic Lymphoma Kinase (ALK) is a receptor tyrosine kinase that was originally described in t(2;5)(p23;q35)-positive ALCL as part of the NPM-ALK fusion protein[3]. Although the physiological function and regulation of full-length ALK receptor is still poorly characterized, aberrant expression of constitutively activated NPM-ALK has been clearly established as the leading cause of ALK-positive ALCL[4]. Tumours bearing ALK gene translocations, amplification or activating point mutations, other than ALCL, have been also identified, including non-small cell lung cancer (NSCLC)[5], Inflammatory Myofibroblastic Tumour (IMT)[6] and neuroblastoma[7]. Compelling studies have indicated that all these malignancies are partially or fully dependent on ALK kinase activity for proliferation and survival[7,8,9], as inhibition of ALK or downregulation of its expression yields potent anti-tumour efficacy both in vitro and in vivo[10].

In this context, the ALK kinase inhibitor crizotinib has been approved for the treatment of ALK-rearranged malignancies, and is now considered the standard of care for both early- and advanced-stage NSCLC patients[11,12]. More recently, crizotinib has entered Phase I/II clinical trial for the treatment of young patients with relapsed or refractory solid tumors and ALCL (ClinicalTrials.gov, NCT00939770, Children's Oncology Group, United States), given to the favourable toxicity profiles and objective response rate demonstrated[13]. For these reasons, the opportunity to use crizotinib as part of first-line therapy in children with ALCL is presently being considered, although failure after treatment, like that reported in a small number of NSCLC and IMT patients[11,12,14], or described by other previous clinical experiences[15,16,17], cannot be overlooked. With kinases, in fact, relapse may be linked to drug- resistance mutations in the catalytic domain, both when acquired de novo and resulting from selection of pre-existing subdominant clones[18,19,20]. The knowledge gained about drug resistance in cancer has shown that minor mutated cell populations can be identified in patients before the onset of treatment, including those that simply promote tumour progression or con- tribute to resistance[21]. Secondary mutations associated with resistance via reduced inhibitor binding[22,23,24] or increased kinase activity have been described in ALK-positive ALCL as well[25,26,27]. However, their presence at diagnosis has never been investigated, likewise their evolution and impact.

In the present study we performed mutational analysis of NPM-ALK kinase domain in paediatric ALCL tumours, to identify point mutations and gene aberrations that could result in changes of NPM-ALK expression and oncogenic activity. Detection of variants was performed by ultra-deep sequencing, in order to assess, at the time of diagnosis, the presence of subclonal mutations not distinguished by conventional Sanger sequencing.

Published - Lovisa, F.; Cozza, G.; Cristiani, A.; Cuzzolin, A.; Albiero, A.; Mussolin, L.; Pillon, M.; Moro, S.; Basso, G.; Rosolen, A.; Bonvini, P. *PloS One* **2015**, *10* (4), e0121378.

The results of this study demonstrated that aberrations of NPM-ALK gene, although uncommon in naïve patients, included both missense and INDEL mutations, which generated low-active and inactive fusion proteins. Functional validation of selected mutants was per- formed by expressing recombinant proteins in the presence or absence of active NPM-ALK kinase, coupled to structure-based computational analysis of ALK catalytic domain. Biochemical results and molecular modelling data confirmed the predicted silent nature of INDELs, and revealed new insights on ALK conformational changes upon single amino acid substitution. Nevertheless, we also found that INDEL mutations present at the time of diagnosis affected constitutive NPM-ALK kinase activity in vitro, by forming nonfunctional heterocomplexes and increasing the sensitivity to specific inhibition.

Materials and Methods

Patients, samples and cell lines

A total of 36 tissue samples from ALK-positive ALCL patients, enrolled between December 2000 and September 2010 in AIEOP-LNH-97 or ALCL-99 treatment protocols, were included in this retrospective analysis. The study was approved by the ethic committee of Azienda Ospedaliera di Padova. In compliance with the Helsinki Declaration, informed written consent was obtained from parents or legal guardians on behalf of the children enrolled in the study[28]. Diagnosis was centrally reviewed by the AIEOP pathologists and further characterized by means of RT-PCR for t(2;5)(p23;q35) translocation[29]. Median age at diagnosis was 9.1 years (range between 3.6 months to 17.5 years), 23 cases were males and 13 females. Most of the cases represented common type ALCL (42%) and, based on St Jude classification, 92% were stage III-IV[30]. For the functional studies, COS7 and HEK-239T cells were grown in RPMI 1640 and DMEM medium, respectively, supplemented with 10% FCS, 2 mM glutamine (Gibco, Life Technologies Co., Carlsbad, CA, USA), 100U/ml penicillin and 100 µg/ml strepto- mycin (SIGMA-Aldrich Co., St. Louis, MO, USA).

Reagents and antibodies

PF-02341066 (Crizotinib) and NVP-TAE684 (TAE684) were purchased from Selleckchem (Selleck Chemicals, Houston, TX, USA), dissolved in DMSO and stored at -20°C. The antibodies used for Western blot analysis were specific for $STAT3^{Y705}$, ALK^{Y1604}, $ALK^{Y1278/Y1282/Y1283}$, c-myc epitope (rabbit) (used at 1:1000 dilution, Cell Signaling Technology, Inc., Danvers, MA, USA); STAT3 (1:1000, Santa Cruz Biotechnology, Inc., Santa Cruz, CA, USA); ALK, V5 epitope, c-myc epitope (mouse) (1:2000, Invitrogen, Life Technologies Co); γ-tubulin (1:5000, SIGMA-Aldrich). DAPI nucleic acid stain, and fluorophore-conjugated goat anti-rabbit Alexa488 and goat anti-mouse Alexa546 antibodies were bought from Molecular

Probes (1:500, Molecular Probes, Life Technologies Co.). Horseradish peroxidase-conjugated sheep anti-mouse or donkey anti-rabbit antibodies, used at 1:2000 dilution, were purchased from GE Healthcare (GE Healthcare Life Sciences, Uppsala, Sweden). Protein G-sepharose Fast-Flow beads were from GE Healthcare as well. For Western blot analysis, proteins were quantified by BCA protein assay (Pierce Chemical, Co., Rockford, Illinois, USA), transferred to nitrocellulose membranes (Whatman, GE Healthcare Life Sciences) and visualized by using PerkinElmer chemiluminescence reagents (PerkinElmer Inc., Waltham, MA, USA), Amersham Hyper-film ECL (GE Healthcare Life Sciences) and Carestream Kodak Autoradiography chemicals (Sigma-Aldrich).

RT-PCR and amplicon library preparation

Total RNA was isolated using TRIzol reagent (Invitrogen) and RT-PCR was performed as reported previously[29]. ALK kinase domain coding region, corresponding to exons 22–25, was amplified using fusion primers, consisting in a target-specific sequence on the 3'-end, an adapter sequence on the 5'-end and a different MID sequence for each primer pair, according to manufacturer's guidelines (S1 Table and S2 Table). Negative and positive controls for mutated ALK gene used were commercially available human ALCL (KARPAS-299) and neuroblastoma (SH-SY5Y) cell lines, respectively. Amplicon products were quantified using Quantity One software (Bio-Rad Laboratories Inc., Hercules, CA, USA) and pooled at an equimolar ratio. Each sample was run on agarose gel, purified by QIAquick gel extraction kit (Qiagen Co., Hilden, Germany) and diluted to a final concentration of 10^7 PCR fragment molecules/µl.

Next-generation sequencing

Amplicon ultra-deep sequencing was performed using Roche 454 Genome Sequencers GS FLX and GS Junior (Roche Applied Science, Penzberg, Germany). The amplicon-PCR-derived fragments were annealed to carrier beads and clonally amplified by emulsion PCR (emPCR), ac- cording to the manufacturer's protocol. The beads carrying single-stranded DNA templates were enriched, counted and deposited into the PicoTiterPlate for sequencing.

454 sequence data have been deposited in the European Nucleotide Archive (ENA, http:// www.ebi.ac.uk/ena/data/view/) under the accession numbers ERS622534 and ERS622535.

Data analysis and detection of variants

All data were generated using the GS Sequencer software version 2.5.3 (Roche Applied Science), and amplicon pipeline analysis was performed using default settings of the GS Run- Browser software version 2.5.3 (Roche Applied Science). Sequence alignments and variant detection was performed using GS Amplicon Variant Analyzer (AVA) 2.7 (Roche Applied Sci- ence), in combination with a blast-based pipeline for low frequent large INDELs detection (CRIBI Genomics, University of Padova, Padova, Italy). AVA software filters were set to dis- play sequence variances represented even by a single read, using human NPM-ALK kinase mRNA sequence (GenBank U04946.1) for reference. Point mutations (single nucleotide polymorphisms, SNPs) were accepted when present with a frequency of at least 0.5% in both for- ward and reverse reads, whereas INDELs were considered when validated by both software, regardless of frequency. INDEL consensus sequences were analyzed using the mRNA-to-genomic alignment program Spidey (http://www.ncbi.nlm.nih.gov/spidey) and manually reviewed.

Molecular modelling and dynamics simulation

ALK mutants were analyzed through the MOE Protein Align tool with BLOSUM 62 as substitution matrix. The homology models were obtained through the MOE homology modelling tool, using human wild-type (WT) and mutants R1275Q and F1174L ALK crystallographic structures (PDB code 3LCT, 4FNX and 4FNW, respectively) as homologous templates[31]. The models have been generated using AMBER99 forcefield, in the presence of ADP docked to the template active site, while water molecules and other cofactors have been removed.

The protonation state of ALK R308-Ins8 and R308-Ins12 tyrosine kinase models were evaluated with Protonate3D (T = 300K, pH = 7) within MOE and Protonate within Amber- Tools 1.5.

The missing residues of WT and R1275Q structures were built using the same approach described above. We used tLeap and Amber FF99SB to parameterized the 'Ins' protein models and solvated them in TIP3P water boxes, adding counterions (Na^+; Cl^-), whereas the point mutant models were parameterized with CHARMM27 forcefield. ClickMD has been used as molecular dynamic platform for NAMD 2.9 minimization (100,000 step, conjugated-gradient method), equilibration (0.5 ns, alpha carbon positional restrains) and production phase (100 ns NVT, P = 1atm, T = 300K) of the molecular systems through 100,000 conjugated gradients method. ACEMD v2728 has been used as molecular dynamics engine on nVidia GeForce GTX680 computational platform. Finally, the analysis of the resulting trajectories was based on RMSD overtime, RainbowRMSD, heatmaps and distance analysis employing VMD 1.9.1, RMSD Tra- jectory Tools 2.01, RAINBOWRMSD and NRGPLOT.

Generation of mutant constructs

The pcDNA3 plasmid containing WT NPM-ALK was obtained from the original pSRα-tkneo-NPM-ALK plasmid (a kind gift from Dr. S. Morris, S. Jude Research Hospital, Memphis, TN, USA), whereas all the other mutants were generated by site-directed mutagenesis, using the Phusion site-directed mutagenesis kit (Thermo Fisher Scientific Inc., Waltham, MA, USA). To co-express WT and mutant NPM-ALK kinase in COS7 or HEK-293T cells, a double cassette vector pBudCE4.1 (Invitrogen) was used, in which WT NPM-ALK was subcloned into the EF- 1α multiple cloning site (MCS) and fused to the V5 epitope, whereas NPM-ALK mutants were subcloned into the CMV MCS and fused to myc tag.

The full-length human ALK cDNA was purchased from ATCC and subcloned into the mammalian expression vector pcDNA3.1. Point mutations F1174L and R1275Q were introduced by site-directed mutagenesis, as previously indicated.

Transfection, treatments and immunofluorescence

To evaluate the effects of NPM-ALK mutations generated, exponentially growing HEK-239T cells were transiently transfected with WT and/or mutated NPM-ALK constructs using Lipofectamine 2000 reagent (Invitrogen), according to manufacturer's instructions. NPM-ALK expression and activity were analyzed in the presence or absence of the ALK-specific inhibitors crizotinib and TAE684, as described in the manuscript.

In the same manner, HEK-293T were transfected with WT or mutated ALK constructs and full-length receptor expression and activity were measured. To assess localization of WT and mutant NPM-ALK proteins, COS7 cells (0.2 x 10^5) were plated on 8-well chamber slides, transfected with 0.5 µg of respective plasmids and processed for immunofluorescence as described previously[32].

Cell lysis, immunoblotting and immunoprecipitation

To assess protein expression and activity, the cells were washed twice in ice-cold 1X phosphate-buffered saline (PBS) and lysed by addition of Triton X-100 sample buffer as reported previously[33]. Binding of WT NPM-ALK to INDEL mutants was performed by incubating protein lysates with 1–2 µg of specific antibodies (α-V5, α-myc or α-phospho-ALK) at 4°C overnight, and resulting immunocomplexes to 20 µl of Protein G-Sepharose beads for 2 h at 4°C. The immunoadsorbed pellets were washed 4 times with 1% Triton X-100 lysis buffer and heated at 95°C in 1X reducing Laemmli loading buffer. Aliquots of cell lysates (50 µg) and immunoprecipitates were fractionated by 10% SDS-PAGE and transferred to nitrocellulose membranes for Western blot analysis. Proteins were visualized by chemiluminescence. Films were scanned and analyzed by using image analysis software ImageJ (National Institute of Health, Bethesda, MD, USA).

Published - Lovisa, F.; Cozza, G.; Cristiani, A.; Cuzzolin, A.; Albiero, A.; Mussolin, L.; Pillon, M.; Moro, S.; Basso, G.; Rosolen, A.; Bonvini, P. *PloS One* **2015**, *10* (4), e0121378.

Results

NPM-ALK mutational analysis

ALK kinase domain mutational status was investigated in ALCL tumour specimens by 454 amplicon ultra-deep sequencing. About 99,000 sequences aligned with ALK exons 22–25 were obtained (mean of 2,234 ± 734 sequences per sample) and an overall of 686 sequence variants, 300 SNPs and 386 INDELs, were detected by Roche AVA software. Among these, a total of 7 SNPs were represented with the same frequency of at least 0.5% on both forward and reverse reads and, thus, accepted, whereas 10 deletions and 3 insertions (INDELs), with the highest detection frequency, were identified and validated by an additional blast-based pipeline specifically designed for INDELs detection and quantification. Of the 7 SNPs detected, 5 were silent point mutations, while 2, namely R335Q (c.1004G>A) and R291Q (c.872G>A), were missense mutations (Table 1, S1 Fig.). With respect to INDELs, 9 variants represented alternative spliced transcripts. Two of them, a deletion of ALK exon 25 first 2 nucleotides (c.923-924del) and of the whole exon 23 (c.696-825del), were common to all patients (frequency ranges 0.03–4% and 0.1–12.8% respectively), whereas 7 were expressed in a variable number of patients at lower frequency (~0.5%) (Table 1, S1 Fig.). Finally, in 4/9 INDELs the mutation resulted in an out of frame (OOF) transcript.

R291Q and R335Q point mutations affect NPM-ALK activity and drug sensitivity

Deep sequencing of NPM-ALK kinase domain identified 2 missense mutations in 2 distinct samples, representing the amino acid changes R335Q and R291Q. Residue R335 lied within the activation loop of ALK kinase domain and corresponded to amino acid R1275 in full-length ALK receptor, whereas R291, corresponding to ALK R1231, was localized on the C-terminal lobe (Fig. 1A). To find out more about these 2 point mutations, we generated NPM-ALK R291Q and R335Q constructs and transfected HEK-293T cells. We found that NPM-ALK R291Q displayed a tyrosine phosphorylation level similar to its WT counterpart, whereas the R335Q mutation markedly reduced ALK kinase phosphorylation both at C-terminal (Y664) and in the activation loop (Y338/342/343) (Fig. 1B). In contrast, downstream STAT3 target phosphorylation was reduced at similar extent, suggesting that both mutants had a lower signalling potential compared to WT NPM-ALK kinase. Indeed, when exposed to crizotinib, a more pronounced dose- and time-dependent inhibitory effect was observed with respect to WT NPM-ALK, which in turn exhibited a progressive recovery overtime (Fig. 1C and D, p-NPM-ALK and p-STAT3). Conversely, the introduction of F234L mutation markedly reduced NPM-ALK drug sensitivity (Fig. 1C), in line with the activating nature of this mutation[12].

Table 1 - NPM-ALK gene mutations found in ALCL patients by ultra-deep sequencing of exons 22-25.

	cDNA variant	Description	Protein change	Amino Acids	Patients	Freq
1	c.872G>A	Missense point mutation	p.R291Q	680	1/36	0.6
2	c.1004G>A	Missense point mutation	p.R335Q	680	1/36	0.5
3	c.747C>T	Silent point mutation	-	680	1/36	0.8
4	c.825G>A	Silent point mutation	-	680	1/36	0.5
5	c.894G>A	Silent point mutation	-	680	1/36	0.8
6	c.909C>T	Silent point mutation	-	680	1/36	0.5
7	c.975G>A	Silent point mutation	-	680	1/36	0.6
8	c.923-924del	Exon 25 first 2 bases as 3' splice site	p.D309H-OOF	342	36/36	0.03-4
9	c.696-825del	Exon 23 skipping	p.S232R-OOF	273	36/36	0.1-12.8
10	c.696-923del	Exon 23-24 skipping	p.Δ232-307	604	16/36	0.04-0.6
11	c.826-923del	Exon 24 skipping	p.P276R-OOF	310	18/36	<0.5
12	c.733-825del	Exon 23 partial deletion, alternative 5' splice site	p.Δ245-275	649	7/36	<0.5
13	c.826-894del	Exon 24 partial deletion, alternative 3' splice site	p.Δ276-299	657	4/36	<0.5
14	c.924ins24	Intron 24 partial retention, alternative 3' splice site	p.R308ins8	688	10/36	<0.5
15	c.924ins36	Intron 24 partial retention, alternative 3' splice site	p.308ins12	692	4/36	<0.5
16	c.924ins106	Intron 24 partial retention, alternative 3' splice site	p.D309H-OOF	312	10/36	0.04-0.6

Published - Lovisa, F.; Cozza, G.; Cristiani, A.; Cuzzolin, A.; Albiero, A.; Mussolin, L.; Pillon, M.; Moro, S.; Basso, G.; Rosolen, A.; Bonvini, P. *PloS One* **2015**, *10* (4), e0121378.

Exposure of cells to a second ALK inhibitor, NVP-TAE684 (TAE684), confirmed these observations (S2 Fig.), providing additional evidence of the different enzymatic properties imparted by these mutations to NPM-ALK kinase activity.

R1275Q-induced structural changes affect ALK receptor activity

In the wide spectrum of ALK mutations, R1275Q and F1174L are the most frequently reported mutations in cancer that result in dysregulation of ALK activity and signalling[9,34]. However, while F1174L is a point mutation that enhances ALK kinase activity and oncogenic potential *per sé*[12], gain-of-function properties of R1275Q are less clear and somehow depend on the model system employed[9,35]. Therefore, to confirm our findings, we introduced F1174L and R1275Q mutations into full-length ALK background and expressed the corresponding constructs into HEK-293T cells.

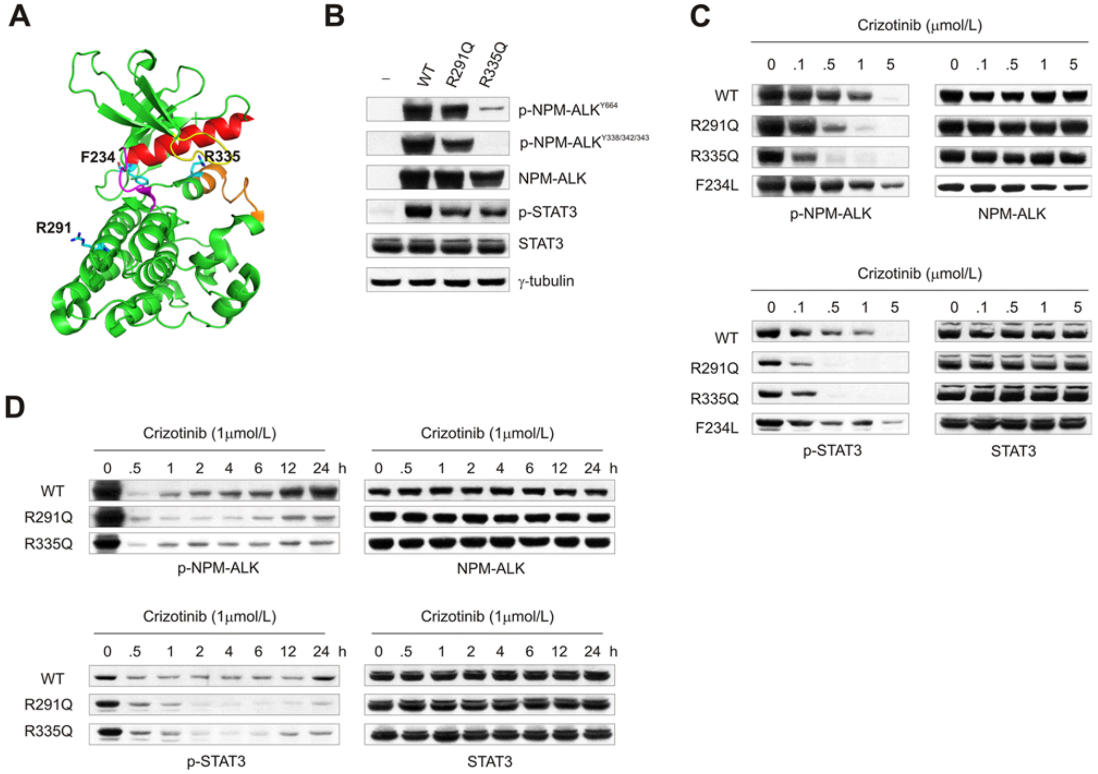


Figure 1 - NPM-ALK point mutations R291Q and R335Q affect kinase activity and sensitivity to crizotinib. **(A)** Cartoon representation of ALK kinase domain, showing positions of NPM-ALK R291 (ALK R1231), R335 (ALK R1275) and F234 (ALK F1174) amino acids (PDB 3LCT). Glycin-rich loop, yellow; activation loop, orange; α-C helix, red; hinge region, magenta. **(B)** Relative protein expression and phosphorylation of wild-type (WT) and mutant (R291Q; R335Q) NPM-ALK in HEK-293T transfected cells. The effects of R291Q and R335Q point mutations were assessed on NPM-ALK and STAT3 phosphorylation (p-NPM-ALK^{Y664} or p-NPM-ALK$^{Y338-342-343}$ and p-STAT3, respectively). γ-tubulin was included as loading control. **(C)** Dose-dependent effect of crizotinib on NPM-ALK expression and activity in HEK-293T cells transfected with WT or mutant (R291Q, R335Q, F234L) NPM-ALK constructs. NPM-ALK and STAT3 phosphorylation levels were determined and compared to total protein expression. **(D)** Time-course analysis of NPM-ALK and STAT3 expression and phosphorylation in HEK-293T cells exposed to 1 μM crizotinib over 24 h.

Published - Lovisa, F.; Cozza, G.; Cristiani, A.; Cuzzolin, A.; Albiero, A.; Mussolin, L.; Pillon, M.; Moro, S.; Basso, G.; Rosolen, A.; Bonvini, P. *PloS One* **2015**, *10* (4), e0121378.

As for F234L and R335Q mutations in NPM-ALK, F1174L ALK exhibited high levels of phosphorylation when ectopically expressed, whereas R1275Q ALK was poorly phosphorylated and less active (p-STAT3) compared to WT and F1174L kinases (Fig. 2A). Besides, ligand-independent receptor activation resulting from protein overexpression, as observed for WT ALK, was not noted in R1275Q ALK expressing cells, providing further evidences of the kinase defective nature of this mutant[36].

Indeed, molecular dynamics simulation analysis of F1174L and R1275Q ALK kinase domain demonstrated significant perturbations in the kinase domain of R1275Q ALK compared to WT (Fig. 2B, R1275Q—WT, closed arrowhead) coupled to a high grade of misfolding of the activation loop moiety (Fig. 2B, cartoon aside, A-loop and Subdomain VII α-helix, open and close arrowheads, respectively; S1 Movie and S2 Movie: F1174L in yellow, R1275Q in red, WT in grey).

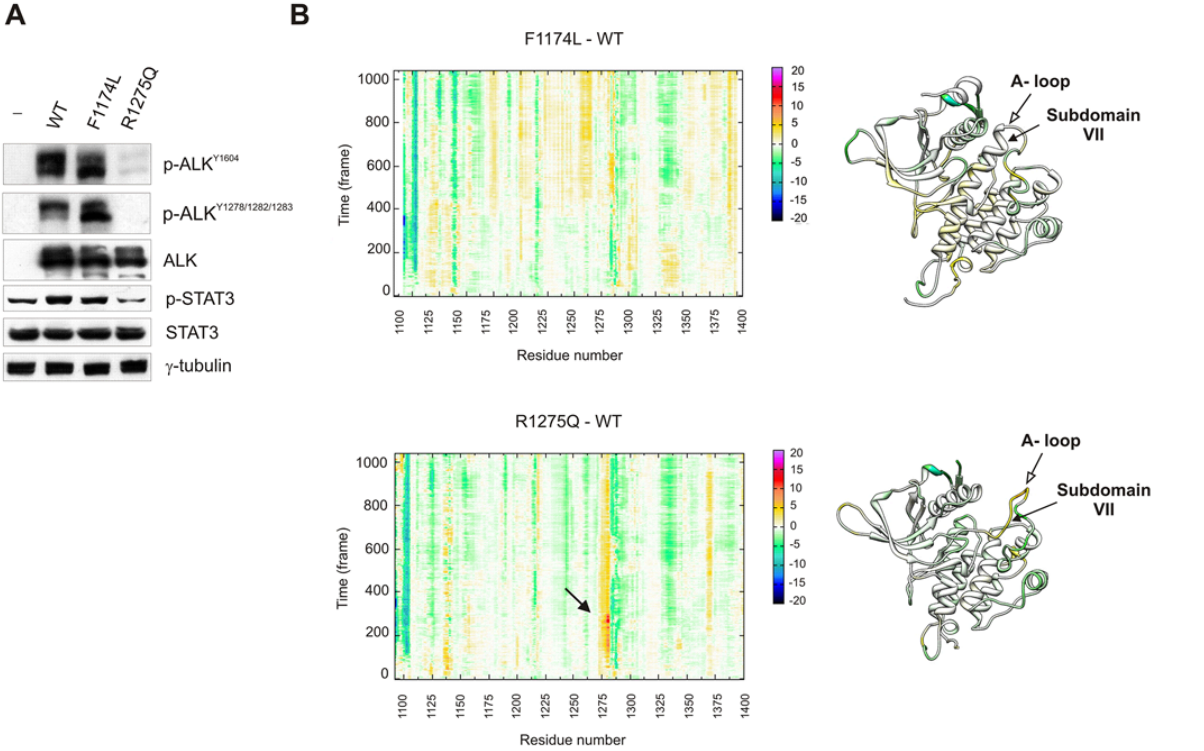


Figure 2 - F1174L and R1275Q molecular dynamics simulation. **(A)** Effect of F1174L (NPM-ALK F234L) and R1275Q (R335Q) point mutations on full-length ALK receptor expression (ALK), phosphorylation (p-ALK^{Y1604} and $^{Y1278-1282-1283}$) and signalling (p-STAT3) in HEK-293T cells. **(B)** Alpha carbon "Rainbow Differential RMSD analysis" between F1174L or R1275Q ALK KD and WT over time (1000 ns) (RMSD_F1174L—RMSD_WT, upper panel; RMSD_R1275Q —RMSD_WT, lower panel). The differential RMSD of each position is encoded by a chromatic scale: negative values (from green to blu) identify regions in which the WT ALK KD is more flexible than the mutant one; positive values (from yellow to magenta) identify regions in which the WT KD is less flexible than mutant KD. Cartoon representation of ALK KD colored by the corresponding differential RMSD values (RMSD_mutant—RMSD_WT) are shown on the right.

Published - Lovisa, F.; Cozza, G.; Cristiani, A.; Cuzzolin, A.; Albiero, A.; Mussolin, L.; Pillon, M.; Moro, S.; Basso, G.; Rosolen, A.; Bonvini, P. *PloS One* **2015**, *10* (4), e0121378.

Molecular modelling and expression analysis of NPM-ALK INDEL variants

By using an approach of cDNA-based amplicon sequencing, we could detect the presence of NPM-ALK alternative spliced transcripts as well. As mentioned above, we identified 13 INDEL mutations, 9 of which resulting from either exon skipping or partial intron retention (Table 1). In 6 of 9 splicing variants, the mutations were associated with extensive deletion of regions crucial for ALK kinase activity, while in 3 variants the structural rearrangements were compatible with ALK enzymatic activity.

We focused on the 2 most common deletions found in pa- tients (c.923-924del, p.D309H-OOF; c.696-825del, p.S232R-OOF) and on the 3 in frame INDELs that conserved most of the residues critical for ATP binding and hydrolysis (c.826-894del, p.Δ276–299; c.924ins24, p.R308-Ins8; c.924ins36, p.R308-Ins12) (S3 Fig.). To examine the functional implication of these variants, each mutant was subcloned and tagged using a double cassette vector designed for simultaneous expression of 2 genes, and expression was assessed in the presence or absence of recombinant WT-V5 NPM-ALK (Fig. 3A). As shown in figure, all mutants were expressed in HEK-293T cells, although low expression levels were observed for extensively deleted S232R-OOF and D309H-OOF INDELs (Fig. 3B, left panels, short and long exposure). Besides, all mutants were catalytically inactive (Fig. 3B, right panels, p-NPM-ALK and p-STAT3), including those initially predicted to be functional based on their conserved sequences. These results were unexpected particularly for R308-Ins8 and R308-Ins12, since the insertions preserved all determinants crucial for ALK kinase activity, including the catalytic Asp308 and the adjacent Arg307 residue (S4A Fig., R and D in colour). Therefore, to provide a molecular explanation to these findings, we performed molecular dynamics simulation on R308-Ins8 and R308-Ins12 mutants and compared time-dependent con- formational changes of mutant KD to that of WT NPM-ALK. While WT kinase showed major conformational changes in the P-loop moiety during the selected time frame (S4B Fig.), R308-Ins8/12 exhibited a pronounced modification of the A-loop in close proximity to the inserted regions (S4C–D Fig., right panels). These structural rearrangements supported novel polar interactions between Ins8/12 motifs and key residues for ATP binding and hydrolysis, having an inhibitory influence on the whole catalytic process (S4C–D Fig., left panels).

myc-tagged INDELs bind and inactivate V5-tagged NPM-ALK kinase

In cells co-expressing native and catalytically inactive alleles, the formation of heterocomplexes may impact on native protein activity. A kinase inactive mutant, in fact, exhibits a dominant- negative effect on the active allele but also can interfere with drug-induced kinase inhibition[37].

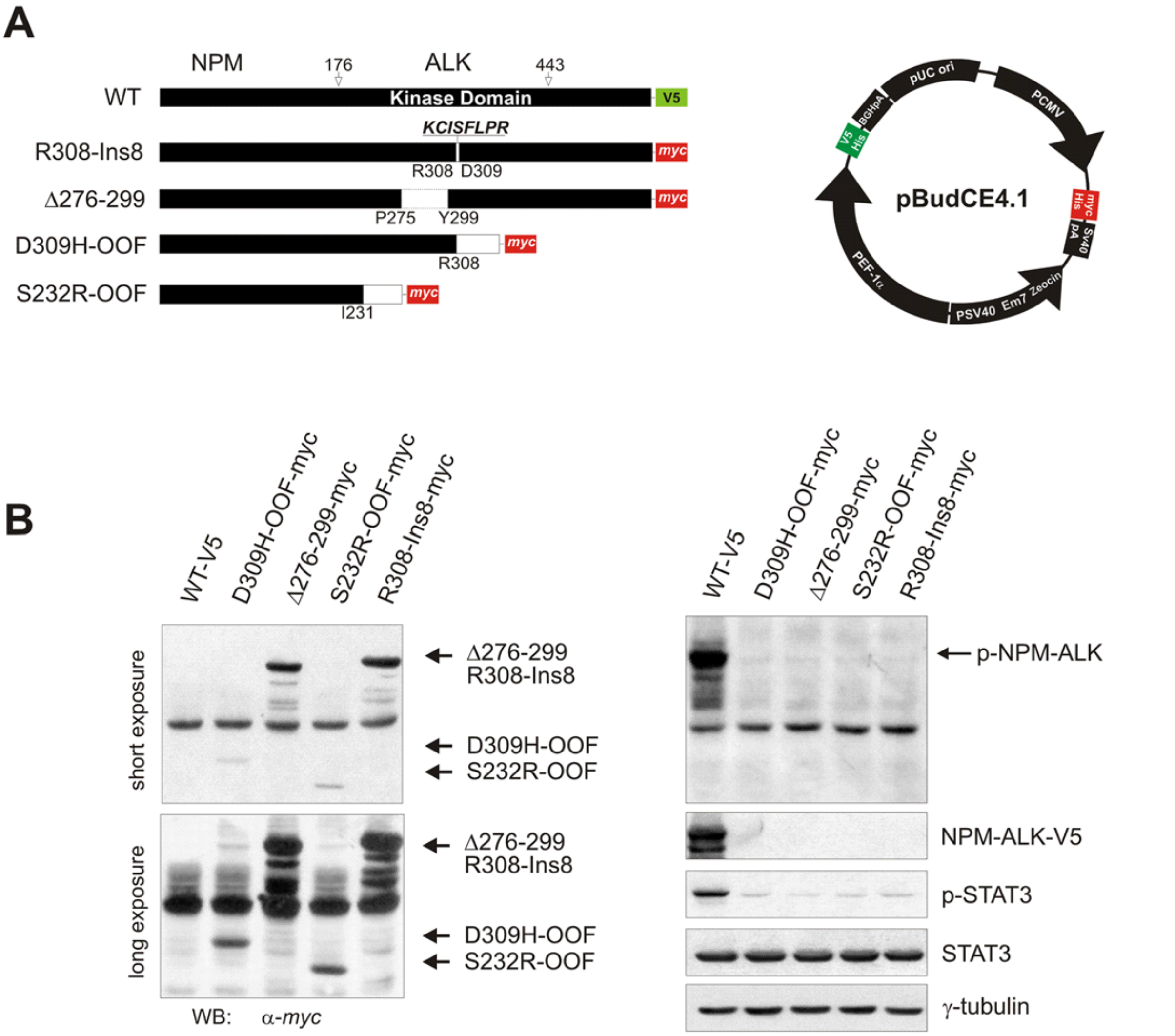


Figure 3 - NPM-ALK INDEL mutants expression and activity in HEK-293T cells. **(A)** Schematic representation of NPM-ALK constructs generated into pBudCE4.1 expression vector according to the in vivo mutational analysis. The figure represents each NPM-ALK construct with the corresponding INDEL (left) and summarizes the features of the pBudCE4.1 vector (right). **(B)** Protein expression analysis of WT and D309H-OOF, Δ276–299, S232R-OOF or R308-Ins8 NPM-ALK constructs in HEK-293T cells by Western blotting. Mutants were detected using an anti-myc specific antibody (NPM-ALK-myc, left panels), whereas WT NPM-ALK was visualized with an anti-V5 antibody (NPM-ALK-V5, right panels). NPM-ALK and STAT3 phosphorylation was also measured by immunoblotting (right panels, p-NPM-ALK and p-STAT3, respectively), using γ-tubulin as loading control.

To functionally test this hypothesis, we introduced WT NPM-ALK and INDEL mutants into HEK-293T cells (Fig. 4A) and assessed NPM-ALK kinase activity upon simultaneous ex- pression. Fresh cell lysates were then immunoprecipitated using either anti-V5 or-myc anti- bodies, and reciprocal immunoblottings were performed. As expected, myc-tagged mutants associated with V5-tagged NPM-ALK (Fig. 4B, upper panel), and vice versa (Fig. 4B, middle panel), irrespective of their mutational status and activity. However, when a phospho-specific ALK antibody was used to purify expressed proteins, the immunocomplexes contained WT-V5 NPM-ALK (Fig.

4C, upper panel, closed arrowhead) but not myc-tagged proteins (Fig. 4C, middle panel, open arrowheads). Indeed, higher levels of phospho-NPM-ALK were found in cells expressing WT NPM-ALK kinase alone (Fig. 4C, lower panel). These data demonstrated that INDELs exert a dominant-negative effect on WT NPM-ALK kinase throughout the formation of inactive protein heterocomplexes.

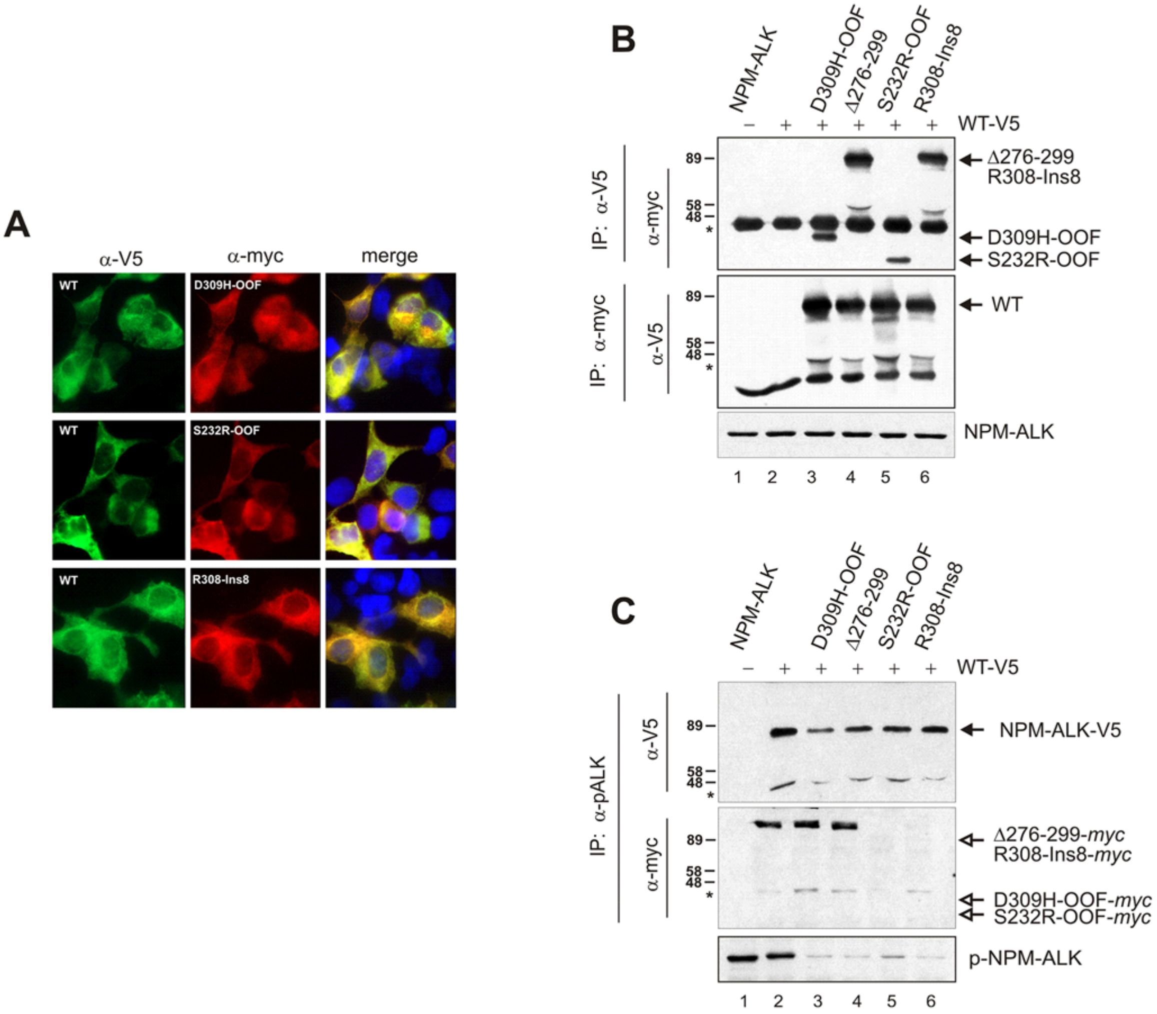


Fig 4. INDEL mutants exert a dominant negative effect on wild-type NPM-ALK kinase activity. **(A)** Subcellular co-localization (merge) of WT NPM-ALK (α-V5, green) and D309H-OOF, S232R-OOF or R308-Ins8 mutants (α-myc, red) in COS7 cells. Cell nuclei are in blue (DAPI dye). **(B)** NPM-ALK WT/INDEL complex formation in HEK-293T co-transfected cells (+), by reciprocal immunoprecipitation. Anti-V5 and anti-myc antibodies were used to precipitate WT NPM-ALK (upper panel) or D309H-OOF, Δ276–299, S232R-OOF and R308-Ins8 mutants (middle panel), respectively, before reciprocal immunoblotting. Untagged WT NPM-ALK (lane 1) was also expressed in these cells and used for quality control of non-specific binding. Total NPM-ALK expression is shown in the lower panel. **(C)** Immunoprecipitation of active NPM-ALK with anti-p-ALK^{Y1604} antibody (α-pALK) in co-transfected (+) HEK-293T cells. Phospho-ALK immunocomplexes were probed for V5- (α-V5, WT NPM-ALK) or myc-tagged (α-myc, INDELs) proteins. Compared to WT NPM-ALK (upper panel, closed arrowhead), INDEL mutants are not phosphorylated (middle panel, lanes 3–6, open arrowheads), thought they reduce basal phosphorylation of WT NPM-ALK kinase (upper and lower panels). Asterisks indicate immunoglobulin (IgG) heavy chain, whereas arrows distinguish relative protein position.

Published - Lovisa, F.; Cozza, G.; Cristiani, A.; Cuzzolin, A.; Albiero, A.; Mussolin, L.; Pillon, M.; Moro, S.; Basso, G.; Rosolen, A.; Bonvini, P. *PloS One* **2015**, *10* (4), e0121378.

NPM-ALK INDELs expression increases native NPM-ALK sensitivity to crizotinib

Finally, to investigate whether INDELs may affect drug sensitivity, HEK-293T cells expressing WT NPM-ALK, either alone or in combination with myc-tagged mutants, were exposed to in- creasing concentrations of crizotinib, and steady-state of total and phosphorylated NPM-ALK was assessed.

Immunoblotting experiments showed a significantly higher sensitivity of WT NPM-ALK to crizotinib in cells co-transfected with mutated NPM-ALK constructs, as co-expression led to a complete inhibition of NPM-ALK phosphorylation (p-NPM-ALK) even at the lowest dose administered (Fig. 5A).

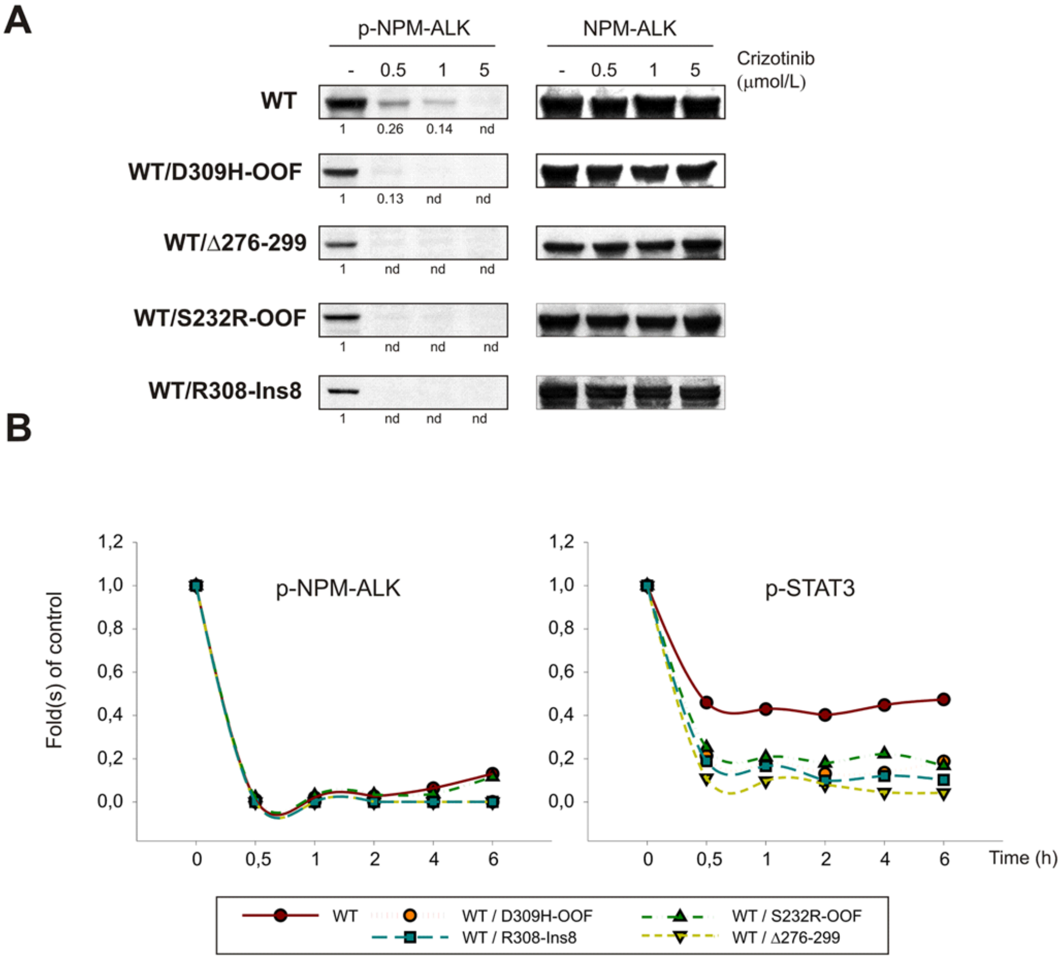


Figure 5 - INDEL mutants increase NPM-ALK sensitivity to crizotinib. HEK-293T cells were transfected with WT NPM-ALK alone or in combination with D309H-OOF, Δ276–299, S232R-OOF and R308-Ins8 mutants. (A) Total and phosphorylated NPM-ALK levels are shown before and after exposure to crizotinib (0.5, 1 and 5 μM) for 6 hours, and band densities, where indicated, are reported as folds of control (-). (B) Crizotinib inhibits NPM-ALK phosphorylation and downstream signalling in a time-dependent manner. NPM-ALK was expressed in HEK-293T cells either alone (WT) or in combination with INDEL mutants (WT/D309H-OOF, WT/Δ276–299, WT/S232R-OOF, WT/R308-Ins8) and exposed to 1 μM Crizotinib for increasing time intervals. Steady-state of phosphorylated NPM-ALK (p-NPM-ALK) and STAT3 (p-STAT3) proteins was determined by Western blotting and reported in graph as fold(s) of control (time exposure 0).

Published - Lovisa, F.; Cozza, G.; Cristiani, A.; Cuzzolin, A.; Albiero, A.; Mussolin, L.; Pillon, M.; Moro, S.; Basso, G.; Rosolen, A.; Bonvini, P. *PloS One* **2015**, *10* (4), e0121378.

Consistently, time-dependent NPM-ALK inhibition showed a brisk reduction of kinase autophosphorylation after 30' of drug exposure, both in the presence or absence of inactive INDELs (Fig. 5B, left). However, a progressive recovery of phosphorylated NPM-ALK was observed in HEK-293T cells expressing either WT alone or in combination with S232R-OOF mutant, providing additional evidence of a stronger NPM-ALK activity when expressed alone or with unstable mutants. In accordance with these observations, drug- induced inhibition of NPM-ALK-dependent STAT3 phosphorylation was more prominent when INDEL mutants were co-expressed, due to the higher amount of non-functional hetero- dimers formed (Fig. 5B, right).

Discussion

Tyrosine kinase inhibitors have become the gold standard therapy of tumour types expressing oncogenic forms of protein tyrosine kinases. However, clinical studies indicate that a significant portion of patients treated with tyrosine kinase inhibitors develop clinical resistance, due to the selection of cancer cells carrying mutations on target kinase.

In NSCLC, primary resistance mechanism of ALK-fusion positive tumours has been mostly secondary mutations within the kinase domain of EML4-ALK, either when compromising drug binding (L1196M, G1269A) or when affecting enzyme conformation and activity (C1156Y, I1171T)[11,38].

In IMT tumours, instead, acquired resistance to ALK inhibitors has shown to occur upon clonal selection of cells harbouring the F1174L activating mutation, in virtue of the important effects this substitution has on ALK tertiary structure[12]. NSCLC and IMT tumours harbouring these 2 types of mutation, however, do not display significant changes in ALK constitutive activity and, therefore, might be positively selected in vivo only in the presence of ALK inhibitors. In contrast, missense mutations associated with familial and sporadic neuroblastoma are mainly activating mutations affecting ALK kinase activity and transforming capability and their detection can occur at diagnosis[39]. The prognostic significance of these mutations, however, is not known yet[13], although for some (G1128A, I1171N, F1174L and F1174V) compound sensitivity has been predicted or experimentally test- ed, or reported in other tumour types (i.e. F1174V in NSCLC and F1174L in IMT)[12,38]. These findings suggest that acquired resistance mutations likely occur in oncogene-driven malignancies, whereas activating site substitutions are more frequent in oncogene-positive tumours. However, extensive clinical experience with BCR-ABL inhibitor imatinib has proven that at least a portion of relapsing patients already harbour the same relapse mutations at diagnosis[20,40,41,42].

In the last few years, the advent of massive parallel next-generation sequencing technologies has greatly enhanced the scope and the speed of molecular cancer research, offering a powerful solution for mutation discovery in tumour samples when mutations are not “fixed” and resistant clones have not emerged yet[21,43]. Early detection of resistance mutations is, therefore, important to predict response to treatment, but also for the impact their identity, frequency and evolution have on clinical course.

Taking advantage of ultra-deep sequencing technology, we amplified exons 22–25 from ALK cDNA, to ensure high sensitive detection of mutations from the expressed translocated NPM-ALK allele in ALCL patients. We aimed at discovering pre-existing NPM-ALK mutations, when also barely expressed, to evaluate their potential impact on NPM-ALK transform- ing activity and drug inhibitor sensitivity.

Two point mutations were identified at diagnosis in 2 distinct cases: a G!A transition at nucleotide 872 and a G!A substitution at nucleotide 1004, which corresponded to R291Q and R335Q amino acid substitution, respectively. R291Q and R335Q corresponded to full-length ALK R1231Q and R1275Q substitutions in neuroblastoma patients[39,44]. We showed here that both residue mutations affected NPM-ALK signalling and drug inhibitor sensitivity. R291Q NPM-ALK kinase displayed catalytic activity similar to WT kinase, whereas R335Q mutation was shown to decrease NPM-ALK autophosphorylation capacity. These data corroborated previous published observations[26,34,35], although they were different from the putative activating nature of this substitution assigned by others[9,44,45]. However, when R1275Q residue mutation was generated in full-length ALK receptor kinase, catalytic activity was reduced compared to WT and F1174L, due to conformational changes in the active site cavity, as sustained by molecular dynamics simulation.

Although our principal aim was the identification of KD mutations potentially relevant for acquired drug resistance, our approach of cDNA-based ultra-deep sequencing provided information on alternative-spicing events in NPM-ALK kinase domain as well. Herein, we identified INDEL mutations resulting from exons skipping or partial intron retention events, including common deletions and more rare in-frame variants previously described in other tumour types[46,47]. An important question addressed in this study was how INDELs expression could affect NPM-ALK activity and drug responsiveness, since oncogenic function of NPM-ALK is strictly dependent on its own dimerization and trans-phosphorylation capacity. Previous studies have shown that insertion/deletion aberrations may have profound effects on protein function, affecting kinase activity or drug-binding affinity. In NSCLC tumours, recurrent EGFR mutations are localized within the catalytic domain and comprise both INDELs and point mutations. However, whereas point mutations are usually activating, INDELs may favor or not the active state of EGFR kinase, depending on size and position. Indeed, whereas NSCLC patients with exon 19 insertions or deletions are responsive to EGFR inhibitors, in- frame insertions in exon 20 confer drug resistance both in vitro and in vivo[48,49]. Both types of mutation are likely to stabilize the active conformation of the kinase. However, while exon 19 insertions increase EGFR affinity for its inhibitors, insertions in exon 20 result in a significant reduction of it[50]. As a consequence, patients with INDELs resulting in reduced EGFR kinase activity are among the best responders to EGFR inhibitors and have a favourable out- come during treatment[51,52].

In this scenario, we found that all INDELs had a dominant-negative effect on WT NPM-ALK and, by forming nonfunctional heterodimers, they significantly reduced the overall kinase activity while increasing sensitivity to specific inhibition. In cells expressing WT NPM-ALK, signal processing was strongly reduced, albeit minimally maintained. In cells co-expressing NPM-ALK kinase and INDEL variants, instead, cell signalling was totally averted both in a time- and dose-dependent manner.

In summary, our study demonstrated that NPM-ALK mutations are uncommon in ALCL patients at diagnosis. These mutations result in single amino acid substitutions or more complex structural rearrangements of NPM-ALK kinase. Whether these subclonal mutations coexist in the same cell with the WT allele, expand or are lost for natural selection is not know yet. However, their identification and characterization may be helpful to identify the most appropriate therapy for each patient, preventing either over-treatment or relapse.

Published - Lovisa, F.; Cozza, G.; Cristiani, A.; Cuzzolin, A.; Albiero, A.; Mussolin, L.; Pillon, M.; Moro, S.; Basso, G.; Rosolen, A.; Bonvini, P. *PloS One* **2015**, *10* (4), e0121378.

Bibliography

1. Ferreri AJ, Govi S, Pileri SA, Savage KJ. Anaplastic large cell lymphoma, ALK-positive. Crit Rev Oncol Hematol. 2012; 83: 293–302. doi: 10.1016/j.critrevonc.2012.02.005 PMID: 22440390
2. Woessmann W, Zimmermann M, Lenhard M, Burkhardt B, Rossig C, Kremens B, et al. Relapsed or re- fractory anaplastic large-cell lymphoma in children and adolescents after Berlin-Frankfurt-Muenster (BFM)-type first-line therapy: a BFM-group study. J Clin Oncol. 2011; 29: 3065–3071. doi: 10.1200/JCO.2011.34.8417 PMID: 21709186
3. Stein H, Mason DY, Gerdes J, O'Connor N, Wainscoat J, Pallesen G, et al. The expression of the Hodg- kin's disease associated antigen Ki-1 in reactive and neoplastic lymphoid tissue: evidence that Reed- Sternberg cells and histiocytic malignancies are derived from activated lymphoid cells. Blood. 1985; 66: 848–858. PMID: 3876124
4. Wan W, Albom MS, Lu L, Quail MR, Becknell NC, Weinberg LR, et al. Anaplastic lymphoma kinase ac- tivity is essential for the proliferation and survival of anaplastic large-cell lymphoma cells. Blood. 2006; 107: 1617–1623. PMID: 16254137
5. Soda M, Choi YL, Enomoto M, Takada S, Yamashita Y, Ishikawa S, et al. Identification of the transform- ing EML4-ALK fusion gene in non-small-cell lung cancer. Nature. 2007; 448: 561–566. PMID: 17625570
6. Coffin CM, Patel A, Perkins S, Elenitoba-Johnson KS, Perlman E, Griffin CA. ALK1 and p80 expression and chromosomal rearrangements involving 2p23 in inflammatory myofibroblastic tumor. Mod Pathol. 2001; 14: 569–576. PMID: 11406658
7. Soda M, Takada S, Takeuchi K, Choi YL, Enomoto M, Ueno T, et al. A mouse model for EML4-ALK- positive lung cancer. Proc Natl Acad Sci U S A. 2008; 105: 19893–19897. doi: 10.1073/pnas. 0805381105 PMID: 19064915
8. Butrynski JE, D'Adamo DR, Hornick JL, Dal Cin P, Antonescu CR, Jhanwar SC, et al. Crizotinib in ALK- rearranged inflammatory myofibroblastic tumor. N Engl J Med. 2010; 363: 1727–1733. doi: 10.1056/ NEJMoa1007056 PMID: 20979472
9. George RE, Sanda T, Hanna M, Frohling S, Luther W, Zhang J, et al. Activating mutations in ALK pro- vide a therapeutic target in neuroblastoma. Nature. 2008; 455: 975–978. doi: 10.1038/nature07397 PMID: 18923525
10. La Madrid AM, Campbell N, Smith S, Cohn SL, Salgia R. Targeting ALK: a promising strategy for the treatment of non-small cell lung cancer, non-Hodgkin's lymphoma, and neuroblastoma. Target Oncol. 2012; 7: 199–210. doi: 10.1007/s11523-012-0227-8 PMID: 22968692
11. Choi YL, Soda M, Yamashita Y, Ueno T, Takashima J, Nakajima T, et al. EML4-ALK mutations in lung cancer that confer resistance to ALK inhibitors. N Engl J

Med. 2010; 363: 1734–1739. doi: 10.1056/ NEJMoa1007478 PMID: 20979473

12. Sasaki T, Okuda K, Zheng W, Butrynski J, Capelletti M, Wang L, et al. The neuroblastoma-associated F1174L ALK mutation causes resistance to an ALK kinase inhibitor in ALK-translocated cancers. Can- cer Res. 2010; 70: 10038–10043. doi: 10.1158/0008-5472.CAN-10-2956 PMID: 21030459
13. Mosse YP, Lim MS, Voss SD, Wilner K, Ruffner K, Laliberte J, et al. Safety and activity of crizotinib for paediatric patients with refractory solid tumours or anaplastic large-cell lymphoma: a Children's Oncolo- gy Group phase 1 consortium study. Lancet Oncol. 2013; 14: 472–80. doi: 10.1016/S1470-2045(13) 70095-0 PMID: 23598171
14. Katayama R, Shaw AT, Khan TM, Mino-Kenudson M, Solomon BJ, Halmos B, et al. Mechanisms of ac- quired crizotinib resistance in ALK-rearranged lung Cancers. Sci Transl Med. 2012; 4: 120ra117.
15. Shah NP, Nicoll JM, Nagar B, Gorre ME, Paquette RL, Kuriyan J, et al. Multiple BCR-ABL kinase do- main mutations confer polyclonal resistance to the tyrosine kinase inhibitor imatinib (STI571) in chronic phase and blast crisis chronic myeloid leukemia. Cancer Cell. 2002; 2: 117–125. PMID: 12204532
16. Tamborini E, Bonadiman L, Greco A, Albertini V, Negri T, Gronchi A, et al. A new mutation in the KIT ATP pocket causes acquired resistance to imatinib in a gastrointestinal stromal tumor patient. Gastro- enterology. 2004; 127: 294–299. PMID: 15236194
17. Kobayashi S, Boggon TJ, Dayaram T, Janne PA, Kocher O, Meyerson M, et al. EGFR mutation and re- sistance of non-small-cell lung cancer to gefitinib. N Engl J Med. 2005; 352: 786–792. PMID: 15728811
18. Branford S, Rudzki Z, Walsh S, Parkinson I, Grigg A, Szer J, et al. Detection of BCR-ABL mutations in patients with CML treated with imatinib is virtually always accompanied by clinical resistance, and mu- tations in the ATP phosphate-binding loop (P-loop) are associated with a poor prognosis. Blood. 2003; 102: 276–283. PMID: 12623848
19. Willis SG, Lange T, Demehri S, Otto S, Crossman L, Niederwieser D, et al. High-sensitivity detection of BCR-ABL kinase domain mutations in imatinib-naive patients: correlation with clonal cytogenetic evolu- tion but not response to therapy. Blood. 2005; 106: 2128–2137. PMID: 15914554
20. Pfeifer H, Wassmann B, Pavlova A, Wunderle L, Oldenburg J, Binckebanck A, et al. Kinase domain mu- tations of BCR-ABL frequently precede imatinib-based therapy and give rise to relapse in patients with de novo Philadelphia-positive acute lymphoblastic leukemia (Ph+ ALL). Blood. 2007; 110: 727–734. PMID: 17405907
21. Soverini S, De Benedittis C, Machova Polakova K, Brouckova A, Horner D, Iacono M, et al. Unraveling the complexity of tyrosine kinase inhibitor-resistant

populations by ultra-deep sequencing of the BCR- ABL kinase domain. Blood. 2013; 122: 1634–1648. doi: 10.1182/blood-2013-03-487728 PMID: 23794064

22. Ceccon M, Mologni L, Bisson W, Scapozza L, Gambacorti-Passerini C. Crizotinib-resistant NPM-ALK mutants confer differential sensitivity to unrelated Alk inhibitors. Mol Cancer Res. 2013; 11: 122–132. doi: 10.1158/1541-7786.MCR-12-0569 PMID: 23239810
23. Zdzalik D, Dymek B, Grygielewicz P, Gunerka P, Bujak A, Lamparska-Przybysz M, et al. Activating mu- tations in ALK kinase domain confer resistance to structurally unrelated ALK inhibitors in NPM-ALK- positive anaplastic large-cell lymphoma. J Cancer Res Clin Oncol. 2014; 140: 589–598. doi: 10.1007/ s00432-014-1589-3 PMID: 24509625
24. Lu L, Ghose AK, Quail MR, Albom MS, Durkin JT, Holskin BP, et al. ALK mutants in the kinase domain exhibit altered kinase activity and differential sensitivity to small molecule ALK inhibitors. Biochemistry. 2009; 48: 3600–3609. doi: 10.1021/bi8020923 PMID: 19249873
25. Sasaki T, Koivunen J, Ogino A, Yanagita M, Nikiforow S, Zheng W, et al. A novel ALK secondary muta- tion and EGFR signaling cause resistance to ALK kinase inhibitors. Cancer Res. 2011; 71: 6051–6060. doi: 10.1158/0008-5472.CAN-11-1340 PMID: 21791641
26. Zhang S, Wang F, Keats J, Zhu X, Ning Y, Wardwell SD, et al. Crizotinib-resistant mutants of EML4- ALK identified through an accelerated mutagenesis screen. Chem Biol Drug Des. 2011; 78: 999–1005. doi: 10.1111/j.1747-0285.2011.01239.x PMID: 22034911
27. Gambacorti Passerini C, Farina F, Stasia A, Redaelli S, Ceccon M, Mologni L, et al. Crizotinib in ad- vanced, chemoresistant anaplastic lymphoma kinase-positive lymphoma patients. J Natl Cancer Inst. 2014; 106: djt378. doi: 10.1093/jnci/djt378 PMID: 24491302
28. Mussolin L, Damm-Welk C, Pillon M, Zimmermann M, Franceschetto G, Pulford K, et al. Use of minimal disseminated disease and immunity to NPM-ALK antigen to stratify ALK-positive ALCL patients with dif- ferent prognosis. Leukemia. 2013; 27: 416–422. doi: 10.1038/leu.2012.205 PMID: 22907048
29. Mussolin L, Pillon M, d'Amore ES, Santoro N, Lombardi A, Fagioli F, et al. Prevalence and clinical impli- cations of bone marrow involvement in pediatric anaplastic large cell lymphoma. Leukemia. 2005; 19: 1643–1647. PMID: 16049513
30. Murphy SB, Fairclough DL, Hutchison RE, Berard CW. Non-Hodgkin's lymphomas of childhood: an analysis of the histology, staging, and response to treatment of 338 cases at a single institution. J Clin Oncol. 1989; 7: 186–193. PMID: 2915234
31. Epstein LF, Chen H, Emkey R, Whittington DA. The R1275Q neuroblastoma

mutant and certain ATP- competitive inhibitors stabilize alternative activation loop conformations of anaplastic lymphoma kinase. J Biol Chem. 2012; 287: 37447–37457. doi: 10.1074/jbc.M112.391425 PMID: 22932897

32. Bonvini P, Zorzi E, Mussolin L, Pillon M, Romualdi C, Peron M, et al. Consequences of heat shock pro- tein 72 (Hsp72) expression and activity on stress-induced apoptosis in CD30+ NPM-ALK+ anaplastic large-cell lymphomas. Leukemia. 2012; 26: 1375–1382. doi: 10.1038/leu.2011.367 PMID: 22289917
33. Bonvini P, Dalla Rosa H, Vignes N, Rosolen A. Ubiquitination and proteasomal degradation of nucleo- phosmin-anaplastic lymphoma kinase induced by 17-allylamino-demethoxygeldanamycin: role of the co-chaperone carboxyl heat shock protein 70-interacting protein. Cancer Res. 2004; 64: 3256–3264. PMID: 15126367
34. Bresler SC, Wood AC, Haglund EA, Courtright J, Belcastro LT, Plegaria JS, et al. Differential inhibitor sensitivity of anaplastic lymphoma kinase variants found in neuroblastoma. Sci Transl Med. 2011; 3: 108ra114. doi: 10.1126/scitranslmed.3002950 PMID: 22072639
35. Schonherr C, Ruuth K, Yamazaki Y, Eriksson T, Christensen J, Palmer RH, et al. Activating ALK muta- tions found in neuroblastoma are inhibited by Crizotinib and NVP-TAE684. Biochem J. 2011; 440: 405–413. doi: 10.1042/BJ20101796 PMID: 21838707
36. Boutterin MC, Mazot P, Faure C, Doly S, Gervasi N, Tremblay ML, et al. Control of ALK (wild type and mutated forms) phosphorylation: specific role of the phosphatase PTP1B. Cell Signal. 2013; 25: 1505–1513. doi: 10.1016/j.cellsig.2013.02.020 PMID: 23499906
37. Sherbenou DW, Hantschel O, Turaga L, Kaupe I, Willis S, Bumm T, et al. Characterization of BCR-ABL deletion mutants from patients with chronic myeloid leukemia. Leukemia. 2008; 22: 1184–1190. doi: 10.1038/leu.2008.65 PMID: 18354488
38. Friboulet L, Li N, Katayama R, Lee CC, Gainor JF, Crystal AS, et al. The ALK inhibitor ceritinib over- comes crizotinib resistance in non-small cell lung cancer. Cancer Discov. 2014; 4: 662–673. doi: 10. 1158/2159-8290.CD-13-0846 PMID: 24675041
39. Bresler SC, Weiser AW, Huwe PJ, Park JH, Krytska K, Ryles H, et al. ALK mutations confer differential oncogenic activation and sensitivity to ALK inhibition therapy in neuroblastoma. Cancer Cell. 2014; 26: 682–694. doi: 10.1016/j.ccell.2014.09.019 PMID: 25517749
40. Carella AM, Garuti A, Cirmena G, Catania G, Rocco I, Palermo C, et al. Kinase domain mutations of BCR-ABL identified at diagnosis before imatinib-based therapy are associated with progression in pa- tients with high Sokal risk chronic phase chronic myeloid leukemia. Leuk Lymphoma. 2010; 51: 275–278. doi:

10.3109/10428190903503446 PMID: 20038234

41. Soverini S, Vitale A, Poerio A, Gnani A, Colarossi S, Iacobucci I, et al. Philadelphia-positive acute lym- phoblastic leukemia patients already harbor BCR-ABL kinase domain mutations at low levels at the
42. time of diagnosis. Haematologica. 2011; 96: 552–557. doi: 10.3324/haematol.2010.034173 PMID: 21193419
43. Pfeifer H, Lange T, Wystub S, Wassmann B, Maier J, Binckebanck A, et al. Prevalence and dynamics of bcr-abl kinase domain mutations during imatinib treatment differ in patients with newly diagnosed and recurrent bcr-abl positive acute lymphoblastic leukemia. Leukemia. 2012; 26: 1475–1481. doi: 10. 1038/leu.2012.5 PMID: 22230800
44. Thomas RK, Nickerson E, Simons JF, Janne PA, Tengs T, Yuza Y, et al. Sensitive mutation detection in heterogeneous cancer specimens by massively parallel picoliter reactor sequencing. Nat Med. 2006; 12: 852–855. PMID: 16799556
45. Mosse YP, Laudenslager M, Longo L, Cole KA, Wood A, Attiyeh EF, et al. Identification of ALK as a major familial neuroblastoma predisposition gene. Nature. 2008; 455: 930–935. doi: 10.1038/ nature07261 PMID: 18724359
46. Janoueix-Lerosey I, Lequin D, Brugieres L, Ribeiro A, de Pontual L, Combaret V, et al. Somatic and germline activating mutations of the ALK kinase receptor in neuroblastoma. Nature. 2008; 455: 967–970. doi: 10.1038/nature07398 PMID: 18923523
47. van Gaal JC, Flucke UE, Roeffen MH, de Bont ES, Sleijfer S, Mavinkurve-Groothuis AM, et al. Anaplas- tic lymphoma kinase aberrations in rhabdomyosarcoma: clinical and prognostic implications. J Clin Oncol. 2012, 30: 308–315. doi: 10.1200/JCO.2011.37.8588 PMID: 22184391
48. Fleuren ED, Roeffen MH, Leenders WP, Flucke UE, Vlenterie M, Schreuder HW, et al. Expression and clinical relevance of MET and ALK in Ewing sarcomas. Int J Cancer. 2013; 133: 427–436. doi: 10.1002/ ijc.28047 PMID: 23335077
49. Greulich H, Chen TH, Feng W, Janne PA, Alvarez JV, Zappaterra M, et al. Oncogenic transformation by inhibitor-sensitive and -resistant EGFR mutants. Plos Medicine. 2005; 2: 1167–1176.
50. He M, Capelletti M, Nafa K, Yun CH, Arcila ME, Miller VA, et al. EGFR Exon 19 Insertions: A New Family of Sensitizing EGFR Mutations in Lung Adenocarcinoma. Clinical Cancer Research. 2012; 18: 1790–1797. doi: 10.1158/1078-0432.CCR-11-2361 PMID: 22190593
51. Arcila ME, Nafa K, Chaft JE, Rekhtman N, Lau C, Reva BA, et al. EGFR Exon 20 Insertion Mutations in Lung Adenocarcinomas: Prevalence, Molecular Heterogeneity, and Clinicopathologic Characteristics. Molecular Cancer Therapeutics. 2013; 12: 220–229. doi: 10.1158/1535-7163.MCT-12-0620 PMID:

23371856

52. Pao W, Miller V, Zakowski M, Doherty J, Politi K, Sarkaria I, et al. EGF receptor gene mutations are common in lung cancers from "never smokers" and are associated with sensitivity of tumors to gefitinib and erlotinib. Proc Natl Acad Sci U S A. 2004; 13306–13311. PMID: 15329413
53. Chung KP, Wu SG, Wu JY, Yang JC, Yu CJ, Wei PF, et al. Clinical outcomes in non-small cell lung can- cers harboring different exon 19 deletions in EGFR. Clin Cancer Res. 2012, 18: 3470–3477. doi: 10. 1158/1078-0432.CCR-11-2353 PMID: 22510346

CONCLUSION AND FUTURE PERSPECTIVES

In this PhD project different methodologies and techniques were applied with the aim to correctly deal with the proposed problems. Moreover, in most of the cases these approaches were slightly modified and improved to enhance their usefulness for the cases study. This was the example for the ALK project in which two mutants (F1174L and R1275Q) of the protein were investigated through MD simulations. Then, the Root Mean Square Deviation (RMSD) was computed for these two mutants and the wild type (wt). The results were presented by using a graphical plots derived from the RainbowRMSD, that was called delta-RainbowRMSD. In this new representation, the x-axis presents the protein residues, the y-axis presents the simulation time and the colorimetric scale indicates the RMSD difference between the mutants and the wt. The analysis provided a plausible explanation, from an atomistic point of view, for the different phosphorylation performance of the mutants kinase, from an atomistic point of view.

Besides these practical projects, we were more focused in the developments of methods, algorithms and software which have also been integrated in other projects and now they are in-home used as canonical procedure. Two of the developed software reached the last stage of validation: DockBench-v1.0 and SuMD-v1.0. The former has already been released and it is a tool to perform a docking benchmark simulations. The software handles seven docking software for a total of 17 docking protocols. The docking simulations and the relative results are automatically generated for an easy interpretation. Moreover the software allows the user to perform a Virtual Screening (VS) simulation based on the results provided by the software. Recently the tool was applied in D3R Grand Challenge, which foresaw to determine six ligand-protein complexes of the HSP90.

Regarding Supervised Molecular Dynamics (SuMD), this tool is intended to accomplish the study of the ligand-protein recognition mechanism through a cycle of short MD simulations (600 ps length). The software was also integrated with a trajectory manager, which automatically process the generated trajectory and provide plots, graphical representations and raw data in for publishing quality. Despite the tool has not already been released, it was used in different publications proving its powerfulness.

Apart from the stand alone software developed, we proposed a web tool called Adenosiland that provides three-dimensional structure of the adenosine receptors. The platform contains all the solved adenosine structure that can be used in a drug discovery process. Due to the fact that the ligand-driven induced fit of the receptor is a key feature in the ligand-protein interaction, a search algorithm was implemented to select the most suitable structure. This tool was called "Best Template Searching" and suggests a protein structure based on the similarity between a query ligand structure and the co-crystalized ligand with adenosine receptor.

The development of new and alternative methodologies is of utmost importance for the future of the drug discovery process. Indeed, the

Computational Aided Drug Design (CADD) has proved its usefulness in the last decades, however the computational methods accuracy needs to be improved. For this reason, researchers invest most of the efforts in the development of methodologies and algorithms to better support the experimentalists. Moreover, the required time to perform the analysis and the user-friendliness of the proposed new tools, are crucial aspects in any drug discovery process. Consequently, computational efforts were made in order to make easier and to speed up the generation of results. As an example the implementation of the easy-to-use Graphical Unit Interface (GUI), to generate high-quality graphical images (e.g. Pymol or Chimera), has determined a wildly spreading of these pictures in the publications. Based on these considerations, we provided computational tools, which perform complex methodology with a limited relative expertise.

The tendency to developed easy-to-use software will spread the application of sophisticated methodologies to non-expert users.

Nowadays, most of the modelers are hired with the aim to help experimental collaborators by confirming their design idea with computational techniques. Hence, in the next future the CADD scientists should be able to focus their efforts in higher-value-added works, instead of being unfairly treated as ‘docking slaves’[1].

The beneficial scenario described above will lead to the implementation of new methodologies, technologies and techniques, which would provide more accurate predictions. On the other hand, the technology progress is usually faced by the researchers with high expectation, that inevitably gives rise to an overreaction to immature technologies. In 1993 Bezdek proposed a theory that explains the expectations of a new technology along a period of time[2] . In Fig. 1 it is reported the famous curve representing the Bezdek theory which describes the common reaction to a new proposed technology.

The understanding of the tendency described by this theory can be useful to avoid the common overreaction to a new technology experienced by the researchers. Hence, CADD scientists should invest part of their time to better review new computational techniques in order to correctly determine their scope and to reduce their misleading applications. Undoubtedly this situation will reduce the false-expectation reaction experienced by the scientific community and will accelerate the ‘true user benefit’ stage.

[1] [1]Van Drie J. H. – Computer-aided drug design: the next 20 years. – J. Comput Aided Mol Des 2007 21:591-601.

[2] Bezdek, J.C - Fuzzy models—What are they, and why? - Fuzzy Systems, IEEE Transactions on 2011 1:1-6

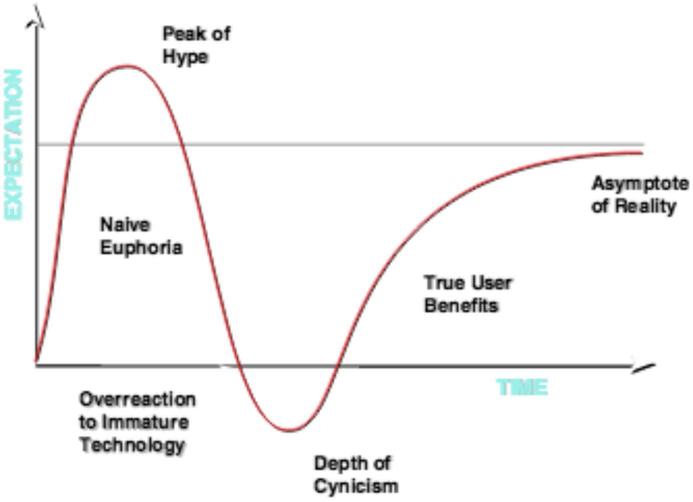


Figure 1 – Bezdek expectation curve: The Bezdek expectation curve shows the common reaction to a new technology.

In the next future we want to focus our efforts into the development of specific Structure-Based techniques and software which will ease routine tasks. In particular we want to create novel methods to be able to deal with certain problems that are still unaffordable with SB approaces. In effect, as it was highlighted in the introduction, in absence of a protein target co-crystalized with a ligand, the SB starts with an important disadvantage. For this reason we would attempt to develop a methodology that will provide an alternative starting point to properly set up SB approaches and consequently reduce or even delete the drawback.

Another possible aspect to be improved would the investigation of the ligand selectivity among a protein family members, with the aim to limit side-effects of drug candidate due to off-target interaction.

Concluding, the available computational resources and methodologies are able to generate results fast and thus, integrate docking and MD approaches easily and efficiently. Indeed, the combination of these two approaches can respectively neutralize the intrinsic defects and limitations of each other, providing more reliable results.