EXPLORING PROTEIN FLEXIBILITY DURING DOCKING
TO INVESTIGATE LIGAND-TARGET RECOGNITION

Coordinator: Ch. mo Prof. Leonard Jan Prins
Supervisor: Ch. mo Prof. Stefano Moro
Ph.D. student: Veronica Salmaso
A mia mamma, mio papà e mia sorella
Abstract

Ligand-protein binding models have experienced an evolution during time: from the lock-key model to induced-fit and conformational selection, the role of protein flexibility has become more and more relevant. Understanding binding mechanism is of great importance in drug-discovery, because it could help to rationalize the activity of known binders and to optimize them. The application of computational techniques to drug-discovery has been reported since the 1980s, with the advent computer-aided drug design. During the years several techniques have been developed to address the protein flexibility issue.

The present work proposes a strategy to consider protein structure variability in molecular docking, through a ligand-based/structure-based integrated approach and through the development of a fully automatic cross-docking benchmark pipeline.

Moreover, a full exploration of protein flexibility during the binding process is proposed through the Supervised Molecular Dynamics. The application of a tabu-like algorithm to classical molecular dynamics accelerates the binding process from the micro-millisecond to the nanosecond timescales. In the present work, an implementation of this algorithm has been performed to study peptide-protein recognition processes.
Sommario


In aggiunta, viene proposta una piena esplorazione della flessibilità proteica durante il processo di legame attraverso la Dinamica Molecolare Supervisionata. L’applicazione di un algoritmo simil-tabu alla dinamica molecolare classica accelera il processo di riconoscimento dalla scala dei micro-millisecondi a quella dei nanosecondi. Nel presente lavoro è stata fatta un’implementazione di questa algoritmica per studiare il processo di riconoscimento peptide-proteina.
INTRODUCTION

1. Ligand-protein binding models

2. Computational methods to study ligand-protein binding
   2.1 Molecular Docking
     2.1.1 Scoring functions
     2.1.2 Sampling
     2.1.3 Semi-flexible docking
     2.1.4 Flexible docking
   2.2 Molecular Dynamics
     2.2.1 Molecular Dynamics and exploration of the phase space
     2.2.2 Advances in classical MD simulations
     2.2.3 Enhanced sampling techniques
       2.2.3.1 Collective Variables-free methods
       Replica Exchange Molecular Dynamics (REMD)
       Accelerated Molecular Dynamics (aMD)
       2.2.3.2 Collective Variables-dependent methods
       Steered Molecular Dynamics (SMD)
       Random Acceleration Molecular Dynamics (RAMD)
       Umbrella Sampling (US)
       Metadynamics
# TABLE OF CONTENTS

## References

30

## AIM OF THE WORK

39

## SCIENTIFIC PUBLICATIONS

43

### Overview of the articles

45

- **DockBench as docking selector tool: the lesson learned from D3R Grand Challenge 2015**
  
- **Combining self- and cross-docking as benchmark tools: the performance of DockBench in the D3R Grand Challenge 2**

- **Sulfonamido-derivatives of unsubstituted carbazole as BACE1 Inhibitors**

- **Synthesis, structure-activity relationships and biological evaluation of 7-phenyl-pyrroloquinolinone 3-amide derivatives as potent antimitotic agents**

- **The role of 5-arylalkylamino- and 5-piperazino- moieties on the 7-aminopyrazolo[4,3-d]pyrimidine core in affecting adenosine A1 and A2A receptor affinity and selectivity profiles**

- **Deciphering the Complexity of Ligand–Protein Recognition Pathways Using Supervised Molecular Dynamics (SuMD) Simulations**

- **Exploring Protein-Peptide Recognition Pathways Using a Supervised Molecular Dynamics Approach**

- **New Trends in Inspecting GPCR-ligand Recognition Process: the Contribution of the Molecular Modeling Section (MMS) at the University of Padova**

## CONCLUSIONS

237
Introduction
1. Ligand-protein binding models

No protein is an island, but exerts its function through recognition with other molecular partners. Ligand-protein interactions are involved in many biological processes, so understanding binding mechanism has been occupying scientists for a long time. Several theories have been developed during the years, with an increasing emphasis on the degree of flexibility of the ligand and protein counterparts.

The first explanation of binding was provided by Emil Fischer in 1894 [1], who firstly proposed the “lock-key” model to explain enzyme specificity: in this model the ligand is natively complementary in shape to its protein binding site, which it recognizes and occupies rigidly as a key to its lock. However, this model could explain neither the behaviour of enzyme noncompetitive inhibition nor allosteric modulation. A modification of the “lock-key” theory was introduced by Koshland [2], who proposed the “induced-fit” theory starting from his observations on enzyme-substrate interactions: according to this binding model, the ligand is able to induce conformational changes to the protein, optimizing their interactions. Later works suggested that many conformational states of the same protein can exist before binding [3, 4], and a ligand can bind preferentially to one of these conformations; this laid the basis for the “conformational selection” model.

Protein-ligand binding is better explained in terms of protein energy landscape, a concept originated to explain protein folding. The free energy landscape model describes the whole conformational space accessible to a system: it consists of the free energy of the system as a function of its conformations (i. e. coordinates of its atoms), resulting in a hypersurface [5]. At a given instant, a protein exists as a single conformational state, but it explores different conformations in time. The probability to occupy a particular region of the phase space is correlated with the energy of that state. In other terms, the energy landscape is conceived as the probability distributions of all conformational states of a protein. The shape of the surface is made of hills and valleys, which respectively stands for low and high probability states. The global minimum and local minima respectively represents native state and metastable states, which are separated by energy barriers (transition states).

The folding energy surface answered to the Levinthal’s paradox [6], which wondered about how protein could fold rapidly given the vastness of the search space. The folding landscape has the shape of a funnel, with the bottom occupied by the folded state; the funnel means that multiple parallel pathways could lead from the multitude of unfolded states to the stable native state [7–11]. The shape of the funnel bottom represents the flexibility of the folded protein: if it is smooth, the protein is rigid, while if it is rugged, with numerous valleys separated by low energy barriers, the protein is flexible.

On this basis, the binding process could be explained in terms of binding energy landscape [5, 12]. Binding can be interpreted as a folding without chain connectivity, and thus with a greater number of degrees of
freedom; this makes the binding energy landscape even more complicated than the folding one. Starting from the simpler example, where both the binding counterparts are rigid (i.e. smooth folding-funnel), recognition process can be conceived as the search for the global minimum of the free energy surface described by the three translational and rotational degrees of freedom of the ligand and the protein. If the interacting counterparts are both flexible (i.e. rugged folding-funnel bottom), they exist as an ensemble of different conformations, and each conformation of the ligand may in principle interact with each conformation of the protein, making the binding landscape really complex [13].

The binding event can stabilize one of the conformations of the protein, either the native or a different one; in the second case, a shift is induced in the equilibrium of protein populations [14].

“Population-shift” (“conformational selection”) differs from “induced-fit” in the chronological sequence of events that bring to binding, and have different ranges of applicability [15–17]; according to the first, the ligand picks one of the already available conformations of the protein, while, according to the second, the ligand induces the conformational rearrangement on the protein. These two models can be used to explain the behaviour of agonists, partial-agonists, antagonists and inverse agonists of G-protein coupled receptors, with rhodopsin activation by all-trans retinal well approximated by “induced fit”, while “conformational selection” is suggested for binding of agonists to β2-adrenergic receptor [15–17].

However, new models are emerging, trying to merge the aforementioned models: an “extended conformational selection model”, for example, has proposed a “conformational selection” followed by conformational adjustment (“induced fit”) [18].
2. Computational methods to study ligand-protein binding

The study of binding mechanism, beyond its epistemological purpose, has concrete applications in modern medicinal chemistry and drug design.

Since 1980s, computer technologies have been applied to the drug discovery process [19], giving rise to Computer-Aided Drug Design (CADD). This technique earned great interest soon and deserved a cover article on October 5, 1981 Fortune magazine, entitled “Next Industrial Revolution: Designing Drugs by Computer at Merck” [19]. CADD techniques are used principally for three reasons: virtual screening, hit/lead optimization and design of novel compounds. In virtual screening a huge database of compounds is examined searching for binding capacity for a target and a subset of compounds is picked out and suggested for *in vitro* testing; the purpose is to increase the hit rate of novel drugs by reducing the number of compounds to test experimentally. The second application of CADD is the optimization of *hit/lead* compounds driven by the rationalization of structure-activity relationship. After the individuation of key elements to bind a target, design of new compounds may be attempted.

CADD methods may be classified as ligand-based (LB) and structure-based (SB), depending on the availability and employment of the target structure [20]. In the framework of CADD, structure-based drug design (SBDD) methods take advantage of the abundance of experimentally solved structures in the Protein Data Bank [21], which can possibly be used also as templates for homology models if the structure of interest is lacking. SBDD is based on the premise that the knowledge of the target structure can help to rationalize and optimize binding, since ligand-target interactions are mediated by their complementarity. With the evolution of the binding models it is clear that speaking of “target structure” is an approximation, given that proteins fluctuate among an ensemble of structures [12].

The possibility to predict ligand binding modes and to interpret binding processes is valuable to individuate, optimize and suggest novel ligands, and for this reason the scientific community has been putting great efforts in developing new computational techniques since the 1980s.

In the following paragraphs we will present an excursus of the main structure-based computational techniques employed in drug-discovery, making a parallelism between them and the aforementioned ligand-protein binding models. In fact, during years, molecular modeling techniques have been experiencing a progressive inclusion of flexibility features in conformational sampling, moving from a static to a dynamic view. The more the flexibility, the higher the number of degrees of freedom of the system, and consequently the computational effort, so this new dynamic view has been made possible by the improvement of hardware technologies. In any case, because of computational limitations, different techniques are used to maximize the balance between efficacy and efficiency, depending on the purpose of their application: for example, in
a virtual screening speed is essential, while slower but higher accurate techniques can be used to hit/lead optimization.

2.1 Molecular Docking

The primary aim of molecular docking is the prediction of the best matching binding mode of a ligand to a macromolecular partner (here just proteins are considered). It consists of the generation of a number of possible conformations-orientations, i.e. poses, of the ligand within the protein binding-site. For this reason, the availability of the three-dimensional structure of the molecular target is a necessary condition; it can be an experimentally solved structure (such as by X-ray crystallography or NMR) or a structure obtained by computational techniques (such as homology modeling).

Molecular docking is composed mainly by two stages: an engine for orientational/conformational sampling and a scoring function, which associate a score to each predicted pose [22–24]. The sampling process should effectively search the conformational space described by the free energy landscape, where energy, in docking, is approximated by the scoring function. The scoring function should be able to associate the native bound-conformation to the global minimum of the energy hypersurface.

2.1.1 Scoring functions

Scoring functions play the role of poses selector, being used to discriminate putative correct binding modes and binders from non-binders in the pool of poses generated by the sampling engine.

There are essentially three types of scoring functions:

1. **Force-field based scoring functions**:

   Force-field is a concept typical of molecular mechanics (see Box 1) which approximates the potential energy of a system through a combination of bonded (intramolecular) and nonbonded (intermolecular) components. In molecular docking generally the nonbonded components are taken into account, with possibly the ligand bonded terms, especially the torsional components. Intermolecular components include the van der Waals term, described by the Lennard-Jones potential, and the electrostatic potential, described by the Coulomb function, where a distance-dependent dielectric may be introduced to consider the solvent effect. However, additional terms have been added to the force-field scoring functions, such as solvation terms. [25]

   Examples of force field based scoring functions are: GoldScore [26], AutoDock [27], GBVI/WSA [28].
2. *Empirical scoring functions:*

These functions are the sum of various empirical energy terms such as van der Waals, electrostatics, hydrogen bond, desolvation, entropy, hydrophobicity, *etc.*, which are weighted by coefficients optimized to reproduce binding affinity data of a training set by least squares fitting [22].

The LUDI [29] scoring function was the first example of an empirical one. Other empirical scoring functions are: GlideScore [30, 31], ChemScore [32], PLANTS CHEMPLP [33].

3. *Knowledge-based scoring functions:*

These methods assume that ligand-protein contacts statistically more observed are correlated with favorable interactions. Starting from a database of structures, the frequencies of ligand-protein atom pairs contacts are computed and converted into an energy component. When evaluating a pose, the aforementioned tabulated energy components are summed up for all ligand-protein atom pairs, giving the score of the pose.

DrugScore [34] [35] and GOLD/ASP [36] are examples of knowledge-based scoring functions.

4. *Consensus Scoring:*

This strategy consists of the combination of multiple scoring functions [37].

In addition, new scoring function have been developed; for example there are application of machine learning, interaction fingerprints, attempts with quantum mechanical scores [38].
INTRODUCTION

Box1

Molecular mechanics is a method which approximates the treatment of molecules with the laws of classical mechanics, in order to limit the computational cost required for quantum mechanical calculations [131]. Atoms are considered as charged spheres connected by springs, neglecting the presence of electrons, in accordance with Born-Oppenheimer approximation [132]. The potential energy is approximated by a simple function, the force-field, which is the sum of bonded (intramolecular) and nonbonded energy terms. The basic form of the function comprises, in the bonded portion, bond stretching and bending described by harmonic potential, and torsional potential described by a trigonometric function. Nonbonded terms consist of van der Waals and Coulomb electrostatic interactions between couples of atoms.

As an example, the basic components of the CHARMM [76] force field are reported in the following equations

\[ V = V_{\text{bonded}} + V_{\text{nonbonded}} \]

\[ V_{\text{bonded}} = \sum_{\text{bonds}} K_b (b - b_0)^2 + \sum_{\text{angles}} K_\theta (\theta - \theta_0)^2 + \sum_{\text{dihedrals}} K_\chi (1 + \cos(n\chi - \delta)) \]

\[ V_{\text{nonbonded}} = \sum_{\text{nonbonded pairs } ij} \frac{q_i q_j \varepsilon r_{ij}}{r_{ij}} + \sum_{\text{nonbonded pairs } ij} \varepsilon_{ij} \left[ \left( \frac{R_{\text{min},ij}}{r_{ij}} \right)^{12} - 2 \left( \frac{R_{\text{min},ij}}{r_{ij}} \right)^6 \right] \]

where \( K_b, K_\theta \) and \( K_\chi \) are the bond, angle and torsional force constants; \( b, \theta \) and \( \chi \) are bond length, bond angle and dihedral angle (0-subscript for the equilibrium values); \( n \) is the multiplicity and \( \delta \) the phase of the torsional periodic function; \( r_{ij} \) is the distance between atoms \( i \) and \( j \); \( q_i \) and \( q_j \) are the partial charges of atoms \( i \) and \( j \); \( \varepsilon \) is the effective dielectric constant; \( \varepsilon_{ij} \) is the Lennard-Jones well depth and \( R_{\text{min},ij} \) is the distance between atoms \( i \) and \( j \) at Lennard-Jones minimum.

These terms may appear slightly different in different force-fields, and anharmonicity and cross-terms are generally added.

The parameters of the force field are obtained by fitting quantum mechanical or experimental values.
2.1.2 Sampling

The first molecular docking algorithm was developed in the 1980s by Kuntz et al. [39]; the receptor was approximated by a series of spheres filling its surface clefts, and the ligand by another set of spheres defining its volume. A search was made to find the best steric overlap between binding site and receptor spheres, neglecting any kind of conformational movement.

This method is a fully-rigid docking, according to the classification of docking techniques based on the degrees of flexibility of simulated molecules [40]:

1. **Rigid docking:**
   
   both ligand and protein are considered as rigid entities, and just the three translational and three rotational degrees of freedom are considered during sampling. This approximation is analogous to the “lock-key” binding model and is mainly used for protein-protein docking, where the number of conformational degrees of freedom is too high to be sampled.

   Generally in these methods the binding site and the ligand are approximated by “hot” points and the superposition of matching points is evaluated [41]. An example is the Triangle Matcher approach of MOE Dock [28].

2. **Semi-flexible docking:**
   
   just one of the molecules, the ligand, is flexible, while the protein is rigid. Thus, the conformational degrees of freedom of the ligand are sampled, in addition to the six translational plus rotational ones. These methods assume that a fixed conformation of a protein may correspond to the one able to recognize the ligands to be docked. This assumption, as already reported, is not always verified.

3. **Flexible docking:**
   
   it introduces the concept that a protein is not a passive rigid entity during binding, but considers both ligand and protein as flexible counterparts. Different methods have been introduced during years, some rested on the induced fit binding model and others on conformational selection.

The increase of flexibility describes a system with a great number of degrees of freedom, in other terms, the potential energy surface is function of numerous coordinates. Consequently, the computational effort required to perform a docking calculation is augmented. Both sampling and scoring should be optimized to give a good balance between accuracy and speed. In fact, the industry-friendly application of docking in virtual screening campaign of millions of compounds depends on the velocity of docking calculations. For this
reason, more and more improvements are made in the development of new algorithms, able to deeply search the phase space, but not at the expense of velocity.

2.1.3 Semi-flexible docking

Several docking algorithms have been developed since the 1980s. Often it is difficult to classify clearly each docking software, because different algorithms may be integrated in a multi-phase approach. However, docking algorithms can be classified as follows [22, 24]:

1. **Systematic search techniques**:

   In systematic search a set of discretized values is associated to each degree of freedom, and all the values of each coordinate are explored in a combinatorial way [25]. These methods are subdivided into:
   
   a. **Exhaustive search**:

      it is a systematic search in the strict sense, since all the rotatable bonds of the ligands are examined in a systematic way. A number of constraints and termination criteria is generally established to limit the search space and avoid a combinatorial explosion. The docking pipeline of the software Glide [30, 42] involves a stage of exhaustive search.

   b. **Fragmentation**:

      the first implementation of ligand flexibility into docking was introduced by Desjarlais [43], who proposed a method made of fragmentation of the ligand, rigid docking of the fragments into the binding site, and subsequent linking of the fragments. In this way, partial flexibility is implemented at the joints between the fragments. Other methods, defined as incremental construction, dock one fragment first and then attach incrementally the others. Examples of methods utilizing fragmentation are FlexX [44] and Hammerhead [45].

   c. **Conformational Ensemble**:

      Rigid docking algorithms can be easily enriched by a sort of flexibility if an ensemble of previously generated conformers of the ligand is docked to the target, in a sort of conformational selection fashion on the ligand counterpart. Examples are offered by FLOG [46], EUDOC [47], MS-DOCK [48].

2. **Stochastic methods**:

   Stochastic algorithms change randomly, instead of systematically, the values of the degrees of freedom of the system. The advantage of these techniques is the speed, so they could potentially find the optimal solution really fast. As a drawback, they do not ensure a full search of the conformational space, so the true
solution may be missed. The lack of convergence is partially solved by increasing the number of iterations of the algorithm. The most famous stochastic algorithms are [22]:

a. **Monte Carlo (MC) methods:**

   Monte Carlo methods are based on the Metropolis Monte Carlo algorithm, which introduces an acceptance criterion in the evolution of the docking search. In particular, at every iteration of the algorithm, a random modification of the ligand degrees of freedom is performed. Then, if the energy score of the pose is improved, the change is accepted, otherwise it is accepted according to the probability expressed in the following equation:

   \[
P \sim \exp \left[ \frac{-(E_1 - E_0)}{k_B T} \right]
   \]

   where \(E_1\) and \(E_0\) are the energy score before and after the modification, \(k_B\) the Boltzmann constant, and \(T\) the temperature of the system.

   This is the original form of Metropolis algorithm, but it is implemented in different variants within docking softwares. Some example are provided by the earlier versions of AutoDock [49, 50], ICM [51], QXP [52], MCDOCK [53], AutoDock Vina [54], ROSETTALIGAND [55].

b. **Tabu search methods:**

   The aim of these algorithms is to prevent the exploration of already sampled zones of the conformational/positional space. Random modifications are performed on the degrees of freedom of the ligand at each iteration. The already sampled conformations are registered, and when a new pose is obtained, it is accepted only if not similar to any pose already explored. PRO_LEADS [56] and PSI-DOCK [57] are two examples of this category.

c. **Evolutionary Algorithms (EA):**

   These algorithms are based on the idea of biological evolution, with the most famous ones coinciding with Genetic Algorithms (GAs). The concept of gene, chromosome, mutation and crossover are borrowed from biology. In particular, the degrees of freedom are encoded into genes, and each conformation of the ligand is described by a chromosome (collection of genes), which is assigned a fitness score. Mutations and crossovers occur within a population of chromosomes, and chromosomes with higher fitness survive and replace the worst ones. The most famous examples are GOLD [58, 59], AutoDock 3 & 4 (which implement a different version of GA, the Lamarckian GA) [27], PSI-DOCK [57], rDock [60].
d. *Swarm optimization (SO) methods:*

These methods take inspiration by swarm behaviour. The sampling of the degrees of freedom of a ligand is guided by the information deposited by good poses already sampled. For example, PLANTS [61] adopts an ACO (Ant Colony Optimization) algorithm mimicking the behaviour of ants, that communicate the easiest way to reach a source of food through the deposition of pheromone. Here, each degree of freedom is associated to a pheromone. Virtual ants choose conformations considering the values of pheromones, and successful ants contribute in pheromone deposition.

Other examples of SOs are: SODOCK [62], pso@autodock [63], PSOVina [64].

3. *Simulation methods:*

The most famous example of this category is Molecular Dynamics, a method that describes the time evolution of a system. A wider explanation will be given in the paragraph 2.2.1.

Energy minimization methods can be inserted in this category, but generally they are not used as stand-alone search engines [24]. Energy minimization is a local optimization technique, used to bring the system to the closest minimum on the potential energy surface.
2.1.4 Flexible docking

Some attempts have been made to introduce protein flexibility into docking calculations. These methods take advantage of different degrees of approximation and can be divided into approaches that consider single protein or multiple protein conformations [65].

1. **Single Protein Conformation**:
   
   a. **Soft docking**:
      
      This method, first described by Jiang and Kim [66], consists of an implicit and rough treatment of protein flexibility. The van der Waals repulsion term employed in force field scoring functions is reduced, allowing small clashes that permit a closer ligand-protein packing. In this way, a sort of induced-fit is simulated. As a drawback, this approach approximates just feeble protein movements, and could implicate unreal poses. [67, 68]

   b. **Sidechain flexibility**:
      
      This strategy introduces alternative conformations for some protein side chains [69]. This is generally done exploiting databases of rotamer libraries. Some docking methods, such as GOLD, sample some side chains degrees of freedom within their own search engine. Obviously, considering side chain flexibility, huge conformational variations of the protein are neglected by these methods.

2. **Multiple Protein Conformations**:

Multiple experimental structures may be available for the same target. Moreover, an ensemble of protein conformations can be obtained via computational techniques, such as Monte Carlo or Molecular Dynamics simulations. The idea of multiple protein conformations docking is to take into account all the diverse structures, following different possible strategies:

   a. **Average grid**:
      
      The structures of the ensemble are used to construct a single average-grid, which can be either a simple or weighted average combination of them [70].

   b. **United description of the protein**:
      
      In this case the structures do not collapse into an average grid, but are used to construct the best performing “chimera” protein. For example, FlexE [44] extracts the structurally conserved portions from the structures of the ensemble, and use them to construct an average rigid structure. This portion is fused to the flexible parts of the ensemble in a combinatorial fashion, giving a pool of “chimeras” that are used for docking.
c. *Individual conformations:*

The structures of the ensemble are considered as conformations that can possibly be bound by the ligand, so various docking run are performed, evaluating the ligands of interest on all the target conformations [71].

In light of the last considerations about multiple protein conformations docking, in the following section a deep insight will be given into Molecular Dynamics, firstly as a tool to generate ensemble of conformations, and secondly as a docking method itself.
2.2 Molecular Dynamics

Molecular dynamics (MD) is a computational technique that simulates the dynamic behaviour of molecular systems as a function of time, treating as flexible all the entities in the simulation box (ligand, protein, as well as water if explicit).

It was developed to simulate simple systems, with the first application to study collisions among hard spheres, in 1957 [72]. The first MD simulation of a biomolecule was accomplished in 1977 by McCammon et al. [73]; it was a 9.2 ps simulation of a 58-residues Bovine Pancreatic Trypsin Inhibitor (BPTI), carried out in vacuum with a crude molecular mechanics potential.

Molecular dynamics computes the movements of atoms along time by the integration of the Newton’s equations of motions (classical mechanics), reported in the following equation [74, 75]:

\[
\frac{d^2 r_i(t)}{dt^2} = F_i(t) / m_i
\]

with \(F_i(t)\) force exerted on atom \(i\) at time \(t\), \(r_i(t)\) vector position of atom \(i\) at time \(t\), \(m_i\) mass of atom \(i\).

In particular, time is partitioned into time steps (\(\delta t\)), which are used to propagate the system forward in time. Several integration algorithms are available, which derive Newton’s equations by a discrete-time numerical approximation. The velocity-Verlet integrator is reported in the following equations as an example to compute position and velocity of an atom \(i\) at the time step \(t+\delta t\), starting from step \(t\).

\[
\begin{align*}
 r_i(t + \delta t) &= r_i(t) + v_i(t) \delta t + \frac{1}{2} a_i(t) \delta t^2 \\
 v_i(t + \delta t) &= v_i(t) + \frac{1}{2} [a_i(t) + a_i(t + \delta t)] \delta t 
\end{align*}
\]

where \(r_i(t)\), \(v_i(t)\) and \(a_i(t)\) are respectively position, velocity and acceleration of atom \(i\) at time \(t\), and \(r_i(t+\delta t)\), \(v_i(t+\delta t)\) and \(a_i(t+\delta t)\) are respectively position, velocity and acceleration of atom \(i\) at time \(t+\delta t\).

Acceleration is calculated from the forces acting on atom \(i\) according to Newton’s second law, and forces are computed from the force field, according to the following equation:

\[
a_i(t) = \frac{d^2 r_i(t)}{dt^2} = \frac{F_i(t)}{m_i} = - \frac{dV(r(t))}{dr_i(t)}
\]

where \(V(r(t))\) is the potential energy function retrieved by the force field (see box1).

The most used force field in molecular dynamics are CHARMM [76], AMBER [77], OPLS [78] and GROMOS [79].
2.2.1 Molecular Dynamics and exploration of the phase space

MD trajectories can be considered as sampling engines; in fact, they produce protein conformations usable for Multiple Protein Conformations docking applications. In particular, McCammon et al. developed the so-called Relaxed-Complex Scheme (RCS): mini-libraries of compounds are docked by AutoDock [27] against a large ensemble of snapshots derived from unliganded protein MD trajectories [80–82]. This approach is founded on the conformational selection binding model.

Alternative conformers obtained through MD can also give insights into cryptic or allosteric binding sites [83]. For example, Schame et al. identified an alternative binding site, named “trench”, close to the active site of the HIV-1 integrase [84].

MD has further applications as a docking-coupled technique. In particular it can be used for the assessment of the stability and for the refinement of docking poses; incorrect poses are likely to be unstable and dissociate from the complex, while realistic ones will be stable [65].

Moreover, simulations in explicit solvent may enable to estimate the contribution of water during binding; for example, Klebe et al. developed a method to compute surface water networks that play a role in modulating binding affinities [85].

All the aforementioned applications of molecular dynamics are used as a complement to classic molecular docking techniques, but molecular dynamics itself can be interpreted as a docking method. In principle, the simulation of the complete binding process of a ligand, from the unbound state in bulk solvent to the bound state, would be a fully-flexible docking in explicit solvent. The possibility to investigate the whole binding process could give insights into: metastable states reached by the ligand during the simulation, alternative binding sites, the role of water during binding and conformational rearrangements preceding, concurrent or consecutive to binding.

However, the observation of a binding event during a classical MD simulation is very rare, raising the timescale problem. The timestep in molecular dynamics must be compatible with the fastest motion in the system; in particular, a timestep of 1-2 fs, corresponding to bond vibrations, has to be used. Thus, an high number of MD steps is required to simulate slow processes, such as large domains motions and binding (μs-ms) [86], making the computational effort really hard. In particular, slow timescales are linked to processes that require the overcoming of a high energy barrier [86], corresponding to low populated states in the conformational energy landscape; in this case the simulated system gets trapped in a local minimum, making classical MD inadequate to explore largely the conformational space.
2.2.2 Advances in classical MD simulations

In 1998 Duan and Kollman performed the first 1µs simulation of a protein in explicit solvent, observing the folding of a 36-residue villin headpiece subdomain from a fully unfolded state. This simulation was two orders of magnitude longer than a state-of-the-art simulation of that period, and it was made possible by advances in massively parallel supercomputers and efficient parallelized codes, but still required 2 months of CPU (Central Processing Units) time [87].

Specialized machines have also been designed specifically for MD calculations; for example, a supercomputer named Anton was conceived as a “computational microscope” and was developed with the idea to reach previously inaccessible simulation timescales within a reasonable computation time [88]. This machine has allowed Shaw et al. to characterize the folding of FiP35 WW domain from a fully extended state in a 100 µs simulation and, in addition, to reach the ms timescale in a single simulation of BPTI in the folded-state [89], followed recently by ubiquitin [90]. Moreover, with unbiased simulations in the order of ten µs, Shaw’s group could simulate the complete binding process of beta blockers and agonists to β2-adrenergic receptor [91] and kinase inhibitors to Src kinase [92].

As a drawback, the utilization of supercomputers is an expense that not many research groups can afford. The development of code able to exploit the speed of GPUs (Graphic Processing Units) has given access to long-timescale simulations at relatively low cost [93–95]. In fact, nowadays, simulations of hundreds of nanoseconds are easily performed, and reaching the microsecond timescale is an affordable issue on a GPU-equipped workstation [96].

Moreover, a paradigm shift seems to have been spreading, that is the possibility to simulate long processes using numerous trajectories shorter than the process itself instead of a single long trajectory. This idea has been exploited by the folding@home project, a worldwide distributed computing environment benefitting from the computers of private citizens, when not in use [97]. Since during a classical MD simulation the system is stuck in a minimum, waiting for the fortunate event that triggers the overcoming of an energy barrier, the simulation of many trajectories in parallel would increase the probability to meet the lucky event. Thus, numerous simulations are started from the same initial condition and run in parallel on different computers, and when one escapes from the energy minimum, all the simulations are stopped and started from the new productive configuration [98].

The new paradigm has found its best application in the use of Markov State Models (MSMs) and adaptive sampling. In fact, MSMs are based on an ensemble view of the dynamics, from which statistical properties, such as the probability to occupy a state and the probability to jump from one state to another, are computed. The construction of a Markov model is made of the discretization and projection of a trajectory
INTRODUCTION

into microstates, and of the computation of a transition probability matrix $T(\tau)$ at a given time, the lagtime $\tau$, chosen in a way that the transition is memory-less (Markovian). Each element $T_{ij}(\tau)$ of the transition matrix represents the conditional probability to find the system in state $j$ at time $t+\tau$ while being in state $i$ at time $t$. The transition matrix approximates the dynamic of the system, and enables to compute the free energy from the equilibrium probability distribution and to extrapolate the timescales of the slowest processes, even if they are not directly explored. On a qualitative fashion, the MSM may individuate diverse metastable states and construct multi-states models of the processes [99]. As an example, a MSM was constructed on an aggregate of nearly 500 100ns-trajectories describing benzamidine-trypsin binding (with 37% productive trajectories); this enabled to characterize the binding process individuating three transition states, and to estimate binding free energy with 1 kcal/mol difference from the experimental one (while a higher deviation from experiment was associated with the extrapolated $k_{on}$ and $k_{off}$) [100]. Moreover, the computation of MSM on the collected data can give a feedback about undersampled zones of the phase space, suggesting where to focus further simulation, adapting the sampling (adaptive sampling methods) and increasing the efficiency of simulations [101, 102]. Currently, the major difficulties of this technique are related to the trajectories partition into discrete states, to the choice of the lagtime and to sufficient sampling able to guarantee statistical significance [103].

Several alternative techniques have been developed during the years to overcome the time limitation imposed by classical MD simulations. A first example consists in the Coarse Grained MD simulations, in which groups of atoms are condensed into spheres, reducing the degrees of freedom of the system [104]. This simplifies the conformational landscape of the system, but, as a drawback, the information on the all-atom simulations, that are precious for drug-discovery aim, are lost.

Additional strategies consist of enhanced sampling techniques that apply a bias to molecular dynamics simulations to increase the accessible timescale, enabling the simulation of slow processes like binding, unbinding and folding processes in reduced amount of time.
2.2.3 Enhanced sampling techniques

These methods add a bias force/potential to the system to increase the rate of escape from local minima, entailing an acceleration of conformational sampling. These methods have been conceived primarily to study either folding, binding or unbinding processes, sharing the underlying idea of sampling enhancement and overcoming high energy barriers.

Enhanced sampling techniques can be divided into methods that make use of collective variables to introduce the bias and methods that do not [105].

The employment of a collective variable (CV) is based on the idea that a complex system can be decomposed into one or a combination of reaction coordinates describing the process of interest. These coordinates are named collective variables since it is assumed they can summarize the behaviour of the entire system. After a careful choice of the CVs, the bias is added on these coordinates during the simulation enhancing sampling along the CVs. The phase space is reduced to the space of the collective variables, since the conformational space is projected to the selected CVs, with a consequent dimensional reduction of the free energy surface.

In the following paragraphs few representative enhanced sampling techniques are reported as an example.

### 2.2.3.1 Collective Variables-free methods

**Replica Exchange Molecular Dynamics (REMD)**

This method increases the temperature to accelerate the conformational sampling. The first formulation of Replica Exchange MD[106], also known as Parallel Tempering (PT), consists of the parallel simulation of a number of independent and simultaneous replicas of the same system, starting from the same configuration, but at different temperatures. At regular time intervals, two replicas characterized by neighbour temperatures are switched, or, in other terms, their temperatures are exchanged, with a probability determined by the energy ($E$) and temperature ($T$) of the system. In particular, the transition probability between simulations at temperatures $T_1$ and $T_2$ is determined by the Metropolis criterion:

\[
P(T_1 \rightarrow T_2) = \begin{cases} 
1 & \text{for } (\beta_2 - \beta_1)(E_1 - E_2) \leq 0 \\
\frac{1}{e^{-|\beta_2 - \beta_1|(E_1 - E_2)}} & \text{for } (\beta_2 - \beta_1)(E_1 - E_2) > 0 
\end{cases}
\]

where $\beta=1/k_B T$ (with $k_B$ the Boltzmann constant).

Temperatures are updated by rescaling the velocities of the parent simulations ($v_1$ and $v_2$ to $v_1'$ and $v_2'$) according to the following equation:
The choice of the panel of temperatures is critical, and various strategies have been proposed to guide the selection [107].

REMD have been developed in several directions; for example in Hamiltonian Replica Exchange multiple Hamiltonians are used instead of temperatures [108]. REMD is mainly used to study protein folding, but applications to ligand binding can be found in literature [109].

**Accelerated Molecular Dynamics (aMD)**

Accelerated MD (aMD) facilitates the egress from a low energy basin by adding a bias potential function ($\Delta V(r)$) when the system is trapped in an energy minimum. In particular, when the potential energy ($V(r)$) is lower than a certain cutoff ($E$), the bias is added giving a modified potential ($V^*(r) = V(r) + \Delta V(r)$); otherwise the simulation continues in the true-unbiased potential ($V^*(r) = V(r)$).

The bias function is reported in the following equation:

$$\Delta V(r) = \frac{(E - V(r))^2}{\alpha + (E - V(r))}$$

where $E$ is the potential energy cutoff and $\alpha$ is a tuning parameter determining the depth of the modified potential energy basin.

$E$ has to be at least greater than $V_{\text{min}}$ (the minimum potential energy, close to the starting configuration), while $\alpha = E - V_{\text{min}}$ will allow to maintain the underlying shape of the landscape [110]. A recent application of aMD has enabled the investigation of the significant conformational changes that P2Y1 receptor undergoes after activation [111].
2.2.3.2 Collective Variables-dependent methods

Steered Molecular Dynamics (SMD)

Taking inspiration from atomic force microscopy experiments, in Steered MD (SMD) an external force is applied to a ligand to drive it out of the target binding site [112]. In this way, SMD simulations can lead to the individuation of unbinding pathways, that can be used also to make inference on possible binding roots; in addition, the added force is assumed to be related to the binding strength.

The first implementations of SMD dealt with anchoring the ligand to a virtual spring, that can either have a constant stiffness and be pulled with constant speed on the free end [113], or have an increasing stiffness and be fixed on the free end at a distance larger than the unbinding pathway [114]. Other possibilities involve constant forces or application of forces on different CVs, such as nonlinear coordinates that can help to explore conformational rearrangement of protein domains [115]. SMD relies on an a priori definition of the applied force direction, which can be fixed (for example a simple straight line) or change during the simulation. The choice of the direction is not easy, because a ligand may bump into obstructions during its way out of the protein, but a method evaluating the minimal steric hindrance has been reported [116]. Moreover, integration with the targeted molecular dynamics (TMD) are reported: in TMD a bias force is applied to conduct the system from an initial to a desired final configuration [117], leading to the individuation of a path that can be used as set of directions for a SMD simulation [112].

Random Acceleration Molecular Dynamics (RAMD)

Random Acceleration MD (RAMD), also defined Random Expulsion MD, is an extension of SMD, and, like this, was developed to study the egress of a ligand from its target binding site. It consists in the application of an artificial randomly-directed force on a ligand to accelerate its unbinding. In this way, in comparison with SMD, RAMD avoids the preliminary choice of the force direction; consequently, if some obstructions are found during the exit pathway, the escape direction is switched.

In particular, the direction of the force is chosen stochastically and maintained for a number of MD steps. If during this time interval the average velocity of the ligand is lower than a specified cut-off (or, in other terms, if the distance covered by the ligand is lower than a cutoff distance, r_{min}), meaning that probably a rigid obstruction has been met, a new force direction is assigned to allow the ligand to search for alternative exit pathways [118].
Umbrella Sampling (US)

Umbrella Sampling (US) [119] consists in restraining the system along one or a combination of CVs. Commonly, the CV range of interest is divided into windows, each characterized by a CV reference value ($\xi_{\text{ref}}$).

A bias potential enhances sampling in each window by forcing the system to stay close to the respective CV reference value. The bias is function of the reaction coordinate, and can have different shape, but generally consists in a simple harmonic, as in the following equation:

$$V(\xi) = \frac{k}{2} (\xi - \xi_{\text{ref}})^2$$

Where $k$ is the strength of the potential and $\xi$ is the value of the CV.

The strength of the bias must be high enough to let energy barriers crossing, but sufficiently low to let system distributions of different windows to overlap, which is required for post-processing analysis.

The aim of US is to force sampling in each window to collect sufficient statistics along the whole reaction coordinate. Then the distribution of the system along the CV is calculated, and from this the free energy along the CV [120]. Different post-processing methods can be used to perform combination and analysis of the data coming from the different US windows; the most famous are umbrella integration [121], the weighted histogram analysis method (WHAM) [122], and the more recent Dynamic Weighted Histogram Analysis (DHAM) [123], which can be used also to derive kinetic parameters.

Metadynamics

Metadynamics [124] introduces a bias potential to the Hamiltonian of the system in the form of a Gaussian-shaped function of one or more collective variables. In this case, the bias does not restrain or constrain the system, neither it forces the system along a preferable direction in the CV space. The bias is used to keep memory of the already explored zones of the phase space, and to discourage the system to visit them again [125].

At time $t$, the bias potential ($V_G(S,t)$) is reported in the following equation:

$$V_G(S,t) = \int_0^t dt' \omega \exp \left( - \sum_{i=1}^d \frac{(S_i(R) - S_i(R(t')))^2}{2\sigma_i^2} \right)$$

where $S(R)=(S_1(R),...,S_d(R))$ is a set of $d$ CVs (which ar functions of the coordinates $R$ of the system), $S_i(R(t))$ is the value of the $i$th CV at time $t$, $\sigma_i$ is the Gaussian width for the $i$th CV, and $\omega$ is the energy rate, given by:

$$\omega = \frac{W}{\tau_G}$$

with $W$ the Gaussian height and $\tau_G$ the deposition rate.
Thus, the bias is “history-dependent”, because it is the sum of the Gaussians that have already been deposited in the CV space during time.

The free energy landscape is explored, starting from the bottom of a well, by a random walk; bias-Gaussians are deposited in the CV space with a given frequency, and at each iteration the bias is the sum of the already deposited Gaussians. As time goes by, the system, instead of being trapped in the bottom of a well, is pushed out by the hill of deposited Gaussians, and enters a new minimum. The process continues until all the minima are compensated by the bias potential [126].

In this way, metadynamics enables to enhance sampling and to reconstruct the free energy surface; this can be used to explore binding/unbinding processes [127], and, with the application of funnel metadynamics [128], to the estimation of binding free energy.

Unfortunately, it may occur that the free energy surface is overfilled, but this has been partially solved by well-tempered metadynamics, in which the height of the added Gaussian is rescaled by the already deposited bias [129]. Another issue with metadynamics is the choice of the CVs, which should describe the slowest motions of the system and the initial-final-relevant intermediates. Moreover, a small number of CVs must be used, so a good strategy is the combination with techniques able to enhance sampling along a great number of transverse coordinates [126], such as parallel tempering [130].


INTRODUCTION


INTRODUCTION


103. Pande VS, Beauchamp K, Bowman GR (2010) Everything you wanted to know about Markov State Models but were afraid to ask. Methods 52:99–105


<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
</table>
Aim of the work
As recalled in the introduction, protein flexibility plays a relevant role in ligand-protein recognition, thus it cannot be neglected in drug design. The excursus over the main molecular modeling techniques already available and currently in development has highlighted the urgency to work on new methods able to describe binding in an accessible timescale. The present work has been conceived in this framework, with the aim to contemplate the issue of protein conformational variability both in classical molecular docking and in molecular dynamics.

In the field of molecular docking, the question arises on how to deal with multiple conformations of the same protein. In fact, increasing numbers of structures are experimentally solved nowadays, and it is not uncommon to find tens to hundreds of crystallographic or NMR structures for the same macromolecule. As explained in the introduction, several possibilities have been experimented on how to exploit these data during docking, such as through the construction of average grids or chimera proteins, or using the whole ensemble of structures for different docking runs. Differently, our purpose is to investigate a docking benchmark strategy aimed to associate to each ligand of a database a custom structure chosen from the protein ensemble. The idea is to construct a completely automatic pipeline, able to reduce user intervention and to accelerate a virtual screening process.

Moreover, another part of the work is focused on the development and implementation of a new technique introduced by our research group in 2014*, the so called Supervised Molecular Dynamics, SuMD. SuMD is an algorithm that makes use of classical molecular dynamics (MD) to investigate ligand-protein binding processes. As reported in the introduction, the possibility to explore the recognition process is currently confined to extremely long classical MD simulations and biased MD simulations. Both these methods have drawbacks, linked to the length of the simulation in the first case, and to the introduction of the bias in the second. SuMD decreases the time scale of a binding process from micro-milliseconds to nanoseconds by introducing a tabu-like algorithm, but without biasing the potential. A classical MD simulation is divided into steps, where the distance between the centers of mass of the ligand and the binding site is monitored. The distance values are collected at regular intervals during a SuMD step and are fitted by a line; if the slope of the line is negative, a new SuMD step starts from the current velocities and configuration of the system, otherwise the simulation is restarted from the previous step.

Starting from this, the present work aims to further test and develop SuMD, exploring the pros and cons of this technique. Particular interest is put to broaden the applicability domain of this method, with major concern to peptide-protein binding, due to the pharmaceutical relevance of peptides.

Scientific Publications
Overview of the articles:

The scientific work has been divided into two main fields, one dedicated to molecular docking techniques and the other to molecular dynamics. The most important results are exposed in the current chapter through the scientific publications that have already been published.

1. Molecular Docking work

The first five articles deal with docking methods, and can be divided into two methodological works and three more applicative ones.

1.1 Methodological projects:


The participation to the docking challenge D3R Grand Challenge 2015 is described in this paper. The challenge required the pose prediction for a number of compounds towards heat shock protein 90 (Hsp90) and Mitogen activated protein kinase kinase kinase kinase 4 (MAP4K4). The availability of a high number of protein structures, particularly for Hsp90, raised the development of the present pipeline. Basically, compounds are associated to proteins on the basis of their chemical similarity with the co-crystallized ligand. Then, the DockBench tool is used to conduct a benchmark of different docking protocols and to perform the screening.


The participation to the next D3R Grand Challenge 2 is presented, with the optimization of the aforementioned procedure. In particular, a fully automatization of the pipeline is performed, integrating the preceding idea with a cross-docking benchmark.

1.2 Applicative projects:


In this work a structure-activity relationship study has been performed, rationalizing the binding mode of N-[3-(9H-carbazol-9-yl)-2-hydroxypropyl]-arylsulfonamides on β-Secretase (BACE1)

In this work the interactions of a series of 7-pyrrolo[3,2-f]quinolinones in the colchicine binding site of tubuline is rationalized.


This article focuses on the rationalization of the selectivity profile of a series of 7-aminopyrazolo[4,3-d]pyrimidines on different adenosine receptor subtypes. A novel analysis method is proposed, which performs comparisons of the electrostatic and hydrophobic interaction fingerprints of the docked compounds with those of a known binder.

2. Molecular Dynamics Work


In this paper, the development and testing of SuMD are presented. The functionalities of the SuMD Analyzer tool are shown: a panel of geometric and energetic analysis can be automatically performed using this software.


Here, the implementation of pepSuMD is reported, showing how the supervision can accelerate also the observation of a peptide-protein binding event. The analysis of the interactions on a per-residue scale can give insights into peptidic and proteic aminoacids that are relevant for the recognition.

In the end, a review is reported, summarizing some of the tools developed by our research group in the last years:

DockBench as docking selector tool:
the lesson learned from D3R Grand Challenge 2015

Veronica Salmaso, Mattia Sturlese, Alberto Cuzzolin, Stefano Moro


Abstract

Structure-based drug design (SBDD) has matured within the last two decades as a valuable tool for the optimization of low molecular weight lead compounds to highly potent drugs. The key step in SBDD requires knowledge of the three-dimensional structure of the target-ligand complex, which is usually determined by X-ray crystallography. In the absence of structural information for the complex, SBDD relies on the generation of plausible molecular docking models. However, molecular docking protocols suffer from inaccuracies in the description of the interaction energies between the ligand and the target molecule, and often fail in the prediction of the correct binding mode. In this context, the appropriate selection of the most accurate docking protocol is absolutely relevant for the final molecular docking result, even if addressing this point is absolutely not a trivial task. D3R Grand Challenge 2015 has represented a precious opportunity to test the performance of DockBench, an integrate informatics platform to automatically compare RMDS-based molecular docking performances of different docking/scoring methods. The overall performance resulted in the blind prediction are encouraging in particular for the pose prediction task, in which several complex were predicted with a sufficient accuracy for medicinal chemistry purposes.

1. Introduction

Molecular docking is widely adopted SBDD approach and its impact is clearly demonstrated by the plethora of software developed until now. In the Click2Drug directory [1] more than 50 software are listed, while more than 60 are catalogued on Wikipedia [2]. Considering that several docking algorithms can be coupled to different scoring functions, the number of different docking/scoring combinations is extremely vast.

The primary issue all docking programs try to address is what combination of orientation and conformation (pose) is the most favorable relative to all the other combinations sampled. When applied to screening, the process also requires a comparison of the best pose (or top best poses) of a given ligand with those of the other ligands such that a final ranking (or ordering) can be obtained. However, molecular docking protocols suffer from inaccuracies in the description of the interaction energies between the ligand and the target molecule, and often fail in the prediction of the correct binding mode. In this context, the appropriate selection of the most accurate docking protocol is absolutely relevant for the final molecular docking
performance, even if addressing this point is absolutely not a trivial task for several reasons: (a) each docking protocol has its peculiar input and output file formats, making their managing really tedious when different software are used; (b) input docking parameters can be very different among diverse programs strongly limiting their use in parallel; (c) more and more frequently it is possible that a molecular target was crystallized in more than one form, and it is then necessary to determine which of these is the most suitable for the docking procedure, in particular, when applied to a virtual screening study; and d) last but not least, the fundamental role played by water molecules during the molecular docking simulations.

To overcome these critical issues, we recently developed a tool to support the molecular modeler in identifying the most accurate protocol by an automated and simultaneous comparison of 17 docking/scoring combinations using a self-docking benchmark procedure [3]. In particular, DockBench is an integrate informatics platform to automatically compare RMDS-based molecular docking performances of different docking/scoring methods. An intuitive graphical analysis can help docking users, including non-expert ones, to identify the best docking/scoring combination to perform a docking-based virtual screening campaign. In this contest, D3R Grand Challenge 2015 has represented a precious opportunity to test the performance of DockBench tool in a blind exercise and using high quality ligand–protein complex structures. In particular, D3R Grand Challenge 2015 was organized allowing participants to compete, in a two-stage process, in the prediction of ligand pose and ligand ranking using two very well known therapeutic targets: heat shock protein 90 (Hsp90) and Mitogen activated protein kinase kinase kinase kinase 4 (MAP4K4). Hsp90 is a chaperone protein which has been deeply investigated over the past decades for its crucial role in cancer cells [4], and MAP4K4 is a serine/threonine kinase that has emerged as such a potential therapeutic target for several disorders, in particular for metabolic and cardiovascular diseases [5].

Considering the peculiarity of the DockBench tool in facilitating the prediction of the ligand poses, we decided to concentrate our efforts in determining the best docking method able to reproduce the most accurate pose geometries. The results obtained in the D3R Grand Challenge 2015 (GC2015) revealed a promising capability of our pipeline in pose prediction task. In particular, the mean RMSD obtained in the Hsp90-complexes was 0.86 Å, while for MAP4K4-complexes the mean RMSD showed less accurate value (3.34 Å). The complete pipeline of DockBench used during the two-stage process of the D3R Grand Challenge 2015 GC2015 and a retrospective analysis of its performance will be described in the present study.
2. Experimental section

2.1 Overview of the workflow

The key concept of the workflow adopted in the GC2015 was the identification of the best protocol available in our laboratory in reproducing the crystallographic poses of selected ligands. In detail, given the target and a set of blind ligands, the workflow was articulated into four steps:

(1) Collection of a training set of complexes containing the target from the protein data bank;
(2) Comparing the performance in a self-docking procedure of 17 different docking protocol on the training set;
(3) Selection of one or more suitable protocols according the RMSD;
(4) Evaluation of the similarity of the blind set and the training set of ligands. If significant similarity was found, it drove the selection of the protein conformation;
(5) Docking of the blind ligands;
(6) Selection of the poses using scoring procedures and visual inspection for ambiguous conformations.

In the ranking predictions the protocol was mostly derived from the pose prediction workflow with a further implementation of rescoring procedures.

The procedure of each pose and rank prediction follows all the points depicted above, but tailoring few of them according the set of blind ligands and the protein target (detailed workflow in Figs. 1, 5) and is commented along the results and discussions.

2.2 Hardware

All computational studies were performed on a 200 cores cluster based on Ubuntu operating system (distribution 14.04, 64 bit) under the network file system (NFS) service. MD simulations were carried out by using Acemd [6] on a GPU cluster of 20 NVIDIA GTX graphics cards.

2.3 Ligands preparation

All ligands were prepared following an in-house pipeline previously reported [7]. Briefly, Corina 3.4 was used to generate three-dimensional structures, as well as to neutralize and deprive them of potential counterions [8]. For each compound, the most favorable ionic state was selected by using the “Protonate” tool implemented in MOE suite and based on Generalized Born electrostatics model [9]. MOE was also used to generate the possible tautomeric states, to energy minimize, and to assign the partial charges of each candidate using MMFF94x force field [10]
2.4 Preparation of ligand–protein complexes

The complexes provided by the organizers as well as those retrieved from the Protein Data Bank (PDB) [11] were subjected to the Structure Preparation and “Protonate-3D” tools implemented in MOE2015.10 suite [9], including water molecules if present.

2.5 Molecular docking

Molecular docking calculations were carried out using the following software: AutoDock 4.2.5.1 [12], AutoDock Vina1.1.2 [13], Glide 6.5 [14, 15], GOLD 5.2 [16], MOE 2015.10 [17], PLANTS 1.2 [18], rDock [19]. DockBench 1.0 [3] was used to perform and analyze molecular docking benchmarks. DockBench default parameters have been set for all docking protocols. MOE 2015.10 was used for docking rescoring procedure, using the following scoring function: pKi, GBVI/WSA, Affinity dG [20].

2.6 Chemical similarity and docking analysis

In house bash or python scripts were used for determining Tanimoto’s similarity using OpenBabel [21] and for calculating root mean square deviations (RMSD) using OpenEye [22], respectively. Visual inspection was performed on MOE 2015.10 and Chimera UCSF [23]. ChEMBL database [24] was queried to obtain experimental affinities using a substructure search tool as implemented in MOE.

2.7 Molecular dynamics simulations

Ligand-Hsp90 complexes selected among docking poses were prepared with AmberTool14 [25] for Molecular Dynamics (MD) simulations as follows.

Each system was solvated with explicit waters (TIP3P model) resulting in a box with boundaries at least 11Å far from any atom of the complex. The simulation box was neutralized with Na+/Cl− ions to a final concentration of 0.1 M. Consequently, the prepared systems were simulated by using AMBER14 [26] Force Field [27] and periodic boundary conditions. General Amber Force Field (GAFF) [28] parameters were used for the ligands, along with RESP partial charges [29], which were obtained with Antechamber [25] by fitting electrostatic potential points calculated with Gaussian [30].

The system equilibration was performed through a stepwise procedure that begins with a conjugate-gradient minimization of 300 steps in order to reduce the steric clashes of the prepared system. The equilibration phase was performed through two consecutive steps, with different ensembles and atom positional restraints. In the first protocol, the MD simulation was performed in a NVE ensemble for 100 ps, with a force constant of 1 kcal mol⁻¹ Å⁻² applied to all protein atoms in order to allow the equilibration of the water molecules. Thereafter, a MD simulation of 500 ps in the NPT ensemble was performed by keeping the alpha-carbons of the protein restrained with the same force magnitude of the previous step. During this step, the
temperature was maintained at 310 K by a Langevin thermostat and the pressure at 1 atm by a Berendsen barostat. Subsequently, all MD simulations were conducted in the NVT ensemble, maintaining the temperature at 310 K.

In all MD simulations, the non-bonded long-range Coulomb interactions were handled by using the particle mesh Ewald summation method (PME) [31] with a cutoff distance of 9 Å and a switching distance of 7.5 Å. All the poses were simultaneously compared in a knockout tournament framework.

Each MD simulation was carried out for 10 ns during which a modified dynamic scoring function (DSF) [32] was computed. This scoring is defined as the cumulative sum of the ligand–protein interaction energy (IE): it includes electrostatic (IEele) and van der Waals (IEvdw) contributions.

The wIE are plotted against the simulation time and linearly fitted to the collected data to obtain the slope coefficient that provides an estimation of the strength of the interaction and the stability of the binding mode.

2.8 Electrostatic energy fingerprints

Electrostatic interactions in MAP4K4-ligand complexes were studied by calculating the Electrostatic Energy Fingerprints (EEF). Amber99 partial charges were computed for the proteins and PM3 partial charges were computed for the ligands using MOE. Per residue electrostatic energy interactions were computed thanks to a in-house SVL script used in MOE. Interactions of the residues within 10 Å from each ligand were plotted in a heat map. This graph, reporting on the X-axis the protein residues of the binding site and on the Y-axis the ligands, attributes a color to the strength of the interactions: in particular, electrostatic energy diminishes going from red to blue. Gnuplot4.5 [33] was used to draw the plots.

3. Results and discussion

D3R Grand Challenge 2015 was organized as a two-stage process applied to Hsp90 and MAP4K4 datasets. In both cases, stage 1 was subdivided in two tasks: the first consisting of a “pose prediction” phase, and the second of a “ranking prediction” phase. Stage 2 had the same aim of stage 1 “ranking prediction” phase, with, as an advantage, the disclosure of the crystallographic structures object of phase 1 “pose prediction” phase.

As anticipated in the Introduction, our computational work was mainly devoted to “pose prediction” following the mantra concept in SBDD that the identification/selection of the most accurate docking protocol is the key step in the prediction of the correct binding mode. For this purpose, we have compared the ability of different docking/scoring combinations in reproducing crystallographic poses, taking advantage of the DockBench software.
3.1 Hsp90

3.1.1 Stage 1: Pose prediction phase

The challenge of Hsp90 Stage 1—“Pose prediction” phase was to predict the coordinates of six protein–ligand complexes and to rank the affinities of 180 compounds, referred to as “ligand test set” in this paper. The workflow used for the pose prediction is reported in Fig. 1 (on the left), and it is divided into four tasks: Hsp90 complexes selection, selection of docking protocol, docking calculations and, finally, best pose selection.

Fig. 1 Workflow for Posing and Scoring predictions designed for the challenge on Hsp90. In blue panel is reported the procedure used in the docking stage divided in four main tasks as reported in the discussion section: Database selection, Docking Protocol Selection, Docking Calculation, and Pose Selection. The Scoring Prediction pipeline is schematized on the green panel. The Scoring Procedure consists in a first ligand preparation step, then two different prediction are sketched (Prediction S1-A and S1-B).
Hsp90 complexes selection

Hsp90 is a well-known target in medicinal chemistry which has been deeply investigated in the last two decades by structural biology. At the time of the challenge, we identified in the Protein Data Bank [11] 155 Hsp90-ligand complexes, as listed in Supplementary Information. Two further complexes provided by the organizers (PDB ID: 4YKR, 4YKY) were added to the structures collected from the PDB. Due to the large amount of structural information, we decided to reduce the number of the crystallographic structures focusing our attention only to those complexes in which the co-crystallized ligands were structurally similar to those provided to us by the organizers. The selection was carried out using a filter based on Tanimoto’s similarity (FP2 fingerprints): in particular similarity was evaluated for each of the 6 ligands to be docked against the 157 crystallographic ligands. We selected the Hsp90 crystallographic complexes in which the co-crystallized ligand showed a similarity index greater than 0.5, resulting in 13 structures: 3R4 M [34], 2YE4 [35], 3B27 [36], 2JJC [37], 3R4 N [34], 3B26 [36], 3OW6 [38], 4LWG [39], 2W14 [40], 2XDX [41], 3OWD [38], 4YKR [42], 4YKY [43] (referred to as their PDB ID). Co-crystallized ligands of these 13 complexes will be called “ligand training set” from here on out in this manuscript. Crystallographic structures and ligand training set were prepared for molecular docking study according to the pipeline reported in the Experimental section.

Selection of docking protocol

In the case of Hsp90 the selection of the docking protocol has been carried out taking into account the possible presence of water molecules as mediators of interactions between the ligand and the residues of the binding cavity. The criterion that has been used for the selection of the water molecules is based both on the assessment of their direct interaction with the ligand and the protein, and of the similarity of their B-factor with the average B-factor of the heavy atoms of the backbone of the protein. A list of the water molecules taken into account for each crystal structure is reported in Table SI1. The ligand training set was subjected to two benchmark studies as reported in the workflow (Fig. 1): the first one, in which each ligand of the training set has been self-docked using 17 docking/scoring combinations in the absence of water molecules, and the second one in which the same ligands have been self-docked using 13 docking/scoring combinations protocols taking into account the selected water molecules.

In DockBench, to judge the performances of the different docking protocols, 20 poses were generated for each ligand of the training set and the RMSD values between predicted and crystallographic poses were calculated.

In order to evaluate the performances of the docking protocols, the lowest (RMSD_{min}) and average (RMSD_{ave}) RMSD values over the 20 poses, as well as the highest number of conformations with a RMSD value lower than the corresponding X-ray resolution (R), N^{(RMSD<R)}, were compared for all the docking protocols. For the specific purpose of the D3R Grand Challenge 2015, we have exploited DockBench ability in suggesting the most accurate docking protocols, that are the protocols able to predict the pose closest to the experimental...
one. For this reason, we have focalized our attention on the docking protocols showing lowest $\text{RMSD}_{\text{min}}$ values. This resulted in different docking protocols in relation to different crystal structures, and sometimes in more than one successful docking protocol for the same crystal structure.

Interestingly, the results of DockBench indicated a significant improvement in reproducing the experimental crystallographic poses when the water molecules were included in the docking procedure, as reported in Fig. 2a. Indeed, including water molecules several protocols were able to reproduce the experimental coordinates with RMSD below 2.0 Å. Following these computational evidences, we decided to include the same crystallographic water molecules also during the docking simulations of the ligand test set.

Finally, for each ligand of the test set we have selected the crystallographic structure of Hsp90 in which the co-crystallized ligand was structurally more similar to the docked ligand.

---

**Fig. 2** Self-docking benchmark results obtained with DockBench. Two different benchmark are shown: in a a benchmark carried out including the most relevant water molecules, while in panel B the benchmark was performed on the same pool of complexes but removing all the crystallographic water molecules. For each panel two heat map are reported: the minimum RMSD values ($\text{RMSD}_{\text{min}}$) returned by the tested docking protocol ($y$-values) for the considered X-ray structures ($x$-values) and the Average RMSD values ($\text{RMSD}_{\text{ave}}$) for the 20 poses generated for each protocol considered. Values are color coded, blue spots identify the best obtained results.
Docking simulations and pose selection For each ligand of the test set we have chosen the crystallographic structures of Hsp90 whose co-crystallized ligand had higher Tanimoto similarity (calculated as FP2 fingerprints comparison) to it. In the case of compound 44 (a benzimidazol-2-one), we have decided to dock it to two structures: 3OWD, selected on the basis of highest similarity, and 4YKR, the benzimidazol-2-one derivative bound structure proposed by the challenge organizers. In the case of compound 73, Tanimoto index was not sufficient to discriminate structures, thus, besides evaluating chemical and structural similarity of the co-crystallized ligands, we have chosen three protein structures showing different conformations of loop 104–114 near the binding site.

Finally, we have selected the docking protocols with best RMSD\textsubscript{min} performance for those crystallographic structures. The final report of the selected crystal structures and the relative docking protocols for each ligand of the test set is summarized in Table 1.

**Table 1** List of test set ligands with relative docking protocols and protein crystallographic structures selected for docking in the pose prediction task

<table>
<thead>
<tr>
<th>Ligand</th>
<th>Protein</th>
<th>Docking Algorithm</th>
<th>Scoring Function</th>
<th>Pose</th>
<th>Slope (DSF)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hsp90_40</td>
<td>4YKR</td>
<td>Gold</td>
<td>Goldscore</td>
<td>1</td>
<td>-30.14</td>
</tr>
<tr>
<td>Hsp90_44</td>
<td>3OWD</td>
<td>Glide</td>
<td>SP</td>
<td>1</td>
<td>-12.94</td>
</tr>
<tr>
<td>Hsp90_44</td>
<td>3OWD</td>
<td>Gold</td>
<td>Goldscore</td>
<td>2</td>
<td>-19.74; -19.81</td>
</tr>
<tr>
<td>Hsp90_44</td>
<td>4YKR</td>
<td>Gold</td>
<td>Goldscore</td>
<td>3</td>
<td>-19.81; -22.57</td>
</tr>
<tr>
<td>Hsp90_44</td>
<td>4YKR</td>
<td>Gold</td>
<td>Goldscore</td>
<td>4</td>
<td>-17.48</td>
</tr>
<tr>
<td>Hsp90_73</td>
<td>3B27</td>
<td>Gold</td>
<td>PLP</td>
<td>1</td>
<td>-42.52</td>
</tr>
<tr>
<td>Hsp90_73</td>
<td>3B26</td>
<td>Gold</td>
<td>Goldscore</td>
<td>2</td>
<td>-46.41; -29.07</td>
</tr>
<tr>
<td>Hsp90_73</td>
<td>2WI4</td>
<td>rDock</td>
<td>Solv</td>
<td>3</td>
<td>-36.71; -36.02</td>
</tr>
<tr>
<td>Hsp90_73</td>
<td>2WI4</td>
<td>rDock</td>
<td>STD</td>
<td>4</td>
<td>-8.37</td>
</tr>
<tr>
<td>Hsp90_164</td>
<td>4YKY</td>
<td>Gold</td>
<td>ASP</td>
<td>1</td>
<td>-29.85</td>
</tr>
<tr>
<td>Hsp90_164</td>
<td>4YKY</td>
<td>Gold</td>
<td>Goldscore</td>
<td>2</td>
<td>-37.86; -35.23</td>
</tr>
<tr>
<td>Hsp90_164</td>
<td>4YKY</td>
<td>Gold</td>
<td>Goldscore</td>
<td>3</td>
<td>-48.70; -47.92</td>
</tr>
<tr>
<td>Hsp90_164</td>
<td>4YKY</td>
<td>Gold</td>
<td>Goldscore</td>
<td>4</td>
<td>-27.94</td>
</tr>
<tr>
<td>Hsp90_175</td>
<td>4YKY</td>
<td>Gold</td>
<td>ASP</td>
<td>1</td>
<td>-32.53; -30.17</td>
</tr>
<tr>
<td>Hsp90_175</td>
<td>4YKY</td>
<td>Gold</td>
<td>Goldscore</td>
<td>2</td>
<td>-30.25</td>
</tr>
<tr>
<td>Hsp90_179</td>
<td>3B27</td>
<td>Gold</td>
<td>PLP</td>
<td>1</td>
<td>-19.85</td>
</tr>
<tr>
<td>Hsp90_179</td>
<td>3B27</td>
<td>Gold</td>
<td>PLP</td>
<td>2</td>
<td>-14.63</td>
</tr>
</tbody>
</table>

For each ligand the number of poses picked after docking and submitted to MD is reported, together with the slope of the DSF computed along the MD trajectory and used as final score (when two slope values are reported, the refer to the first and the second turn of the knockout tournament, respectively). In bold are indicated the poses selected on the basis of DSF slope and finally submitted to the challenge.
After the preliminary validation step using the ligand training set, the Virtual Screening Tool of DockBench was used to perform the ligand test set docking simulations using the same set of parameters adopted in the validation step. A summary of information used in the docking simulations of the ligand test set is collected in Table 1.

After docking, we selected one or more poses resulting from each docking simulation, according to electrostatic and van der Waals interaction energy evaluation and visual inspection. Finally, we used Molecular Dynamics (MD) simulation as post-docking tool to select a unique pose for the challenge submission [29]. For each pose a 10 ns simulation was performed and the dynamic scoring function (DSF) was evaluated. This scoring is computed along the trajectory with the aim to obtain the slope coefficient as an estimation of the binding strength and of the stability of the complex.

![Fig. 3 Superposition of the predicted poses (light blue) on the experimental ones (tan). RMSD values were calculated on the heavy atoms.](image)

**Results** As anticipated, the proposed workflow was designed to produce a unique pose for each ligand of the test set. The superposition of the six predicted complexes on the corresponding crystallographic poses is reported on Fig. 3. The DockBench performance generally showed robust results (Table 2) with a mean RMSD of 0.86 Å considering only the heavy atoms of each docked ligand. Most notably, five complexes shown RMSD values under the 0.61 Å absolutely representative of crystallographic poses. Curiously, ligand Hsp90_44 showed a higher RMSD value, 2.69 Å, mainly ascribable to the 3-pyridinesulfonamide moiety. This substituent in the crystal structure points out to the bulk water and is characterized by high B-factor values while in our prediction it is differently oriented establishing a pi stacking interaction with the benzimidazol-2-one scaffold (as shown in Fig. 4). Despite the shift of the 3-pyridinesulfonamide moiety, the key interactions...
of this scaffold are conserved as well as the orientation of the N-substituted benzimidazol-2-one portion as confirmed by the good RMSD (0.62 Å) calculated considering only this portion of the molecule.

Table 2 Summary of the results of all scoring and docking prediction

<table>
<thead>
<tr>
<th>Receipt-ID</th>
<th>Pred. name</th>
<th>Target/Stage</th>
<th>Scoring Prediction</th>
<th>Pose prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Num. ligands</td>
<td>τ err (Kendall)</td>
</tr>
<tr>
<td>564e3304a7724</td>
<td>S1-A</td>
<td>Hsp90 Stage1</td>
<td>180</td>
<td>0.052</td>
</tr>
<tr>
<td>564e330569871</td>
<td>S1-B</td>
<td>Hsp90 Stage1</td>
<td>180</td>
<td>0.054</td>
</tr>
<tr>
<td>56afca8517dc5</td>
<td>S2-A</td>
<td>Hsp90 Stage2</td>
<td>180</td>
<td>0.05</td>
</tr>
<tr>
<td>56afca9a3644d</td>
<td>S2-B</td>
<td>Hsp90 Stage2</td>
<td>180</td>
<td>0.05</td>
</tr>
<tr>
<td>56afca7f6927b</td>
<td>S2-C</td>
<td>Hsp90 Stage2</td>
<td>180</td>
<td>0.056</td>
</tr>
<tr>
<td>5671ef9fdd7a3</td>
<td></td>
<td>MAP4K4 Stage1</td>
<td>18</td>
<td>0.15</td>
</tr>
<tr>
<td>56afc9e2ae8c8</td>
<td></td>
<td>MAP4K4 Stage2</td>
<td>18</td>
<td>0.203</td>
</tr>
</tbody>
</table>

The values reported correspond to those provided by the organizers. The Values of RMSD are indicated in angstrom (Å)

The final evaluation of Grand Challenge 2015 considers only 5 ligands; in the discussion we included also the ligand Hsp90_44, resulting in a mean RMSD for 6 ligand of 0.86 Å
Fig. 4 Comparison of the predicted pose (light blue) and the experimentally derived complexes. The crystallographic ligand is colored according the B-factor in a light-to-dark pink palette corresponding to low-to-high values. While the benzimidazol-2-one scaffold is in nicely reproduced 0.62 Å the 3-pyridinesulfonamide moiety is placed out from the binding pocket is not well predicted resulting in a RMSD of 2.69 Å for the whole molecule. The binding mode of the portion establishing the key interaction is not affected by the different orientation. This observation is in agreement with the higher B-factor values of the 3-pyridinesulfonamide moiety (dark pink).

3.1.2 Stage 1: ranking prediction step

The aim of Hsp90 Stage 1—“Ranking prediction” phase was to rank the affinities of 180 compounds, referred to as “ligand test set” in this paper. The workflow used for the ranking prediction is reported in Fig. 1 (on the right).

Scoring workflow As already anticipated in the Introduction, 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 ranking the “bioactive” binding mode. Aware of the current limitations of the scoring functions, however, we wanted to compare two ranking methodologies that represent on the one hand the most accurate ranking strategy available in our lab (S1-A) and the other the less expensive in terms of computational time (S1-B) (Fig. 1). This comparison was intriguing for us to establish the possible benefit-cost ratio of these two alternative strategies.

In the first pipeline (S1-A), we clustered the library of 180 compounds according to Tanimoto’s similarity exploiting the Fingerprint Database Clustering tool of MOE: briefly, Tanimoto’s similarity was computed for all the 180 compounds against all of them, and each cluster was composed by molecules which were similar to the same set of molecules. Each cluster was screened by structural similarity (evaluated on the basis of common scaffold search, guided by user’s chemical sensibility and experience) against the 13 ligands of the training set used in the previous benchmark. We selected the protein corresponding to the co-crystallized
ligand with highest similarity to each cluster. The PDB ID of the 13 protein–ligand complexes subjected to DockBench were: 3R4M, 2YE4, 3B27, 2JJC, 3R4 N, 3B26, 3OW6, 4LWG, 2WI4, 2DX, 3OWD, 4YKR, AND 4YKY. After merging some of the clusters according to structural similarity of compounds scaffolds (evaluated by user’s chemical sensitivity and experience), we identified 4 clusters (Table SI2) corresponding to 4 different protein–ligand complexes: 3OWD (2,3-dihydro-1H-benzimidazol-5-yl-methylsulfonamide scaffold), 4YKY (benzophenone scaffold), 4YKR (1,3-dihydro-2H-benzimidazol-2-one scaffold), and 3B27 (2-amino-1,3,5-triazine scaffold) as detailed in SI. Differently to the pose prediction challenge, here we selected the docking protocol for the four complexes using, in addition to DockBench results (RMSD$_{\text{min}}$ and RMSD$_{\text{ave}}$), also the Spearman’s and Kendall’s correlations to evaluate the ability of the protocol to rank the near native pose at the top positions of the ranking list. Briefly, each protocol showing RMSD$_{\text{min}}$ and RMSD$_{\text{ave}}$ below 1 Å and 4 Å (Fig. 2), respectively were then compared according Spearman’s and Kendall’s coefficients (score vs RMSD). The final selection is reported in Table 3. We performed the docking calculation using the Virtual Screening Platform implemented in DockBench using the same parameters adopted in the previous benchmark and the first pose (best score) for each ligand was selected. Finally, to have a homogeneous scoring method, different scoring functions were evaluated for the rescoring procedure. Briefly, we picked a subset of compounds from ChEMBL with known activity (true positive and true negative) for each cluster by a substructure search. Only for cluster2 and cluster3 we identified a sufficient number of ligands, 14 and 17 respectively (Table SI3), to have a raw indication of the classification ability of the tested scoring functions. For those clusters the Spearman and Kendall coefficient were calculated to identify the most performant scoring function (GBVI/WSA dG). The 180 compounds were finally ranked on the basis of the GBVI/WSA dG value of the selected pose.

Table 3. Combination of docking protocol, the PDB ID of protein conformation used for each cluster identify in the rank prediction stage 1 (Hsp90)

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Population</th>
<th>Protein</th>
<th>Protocol</th>
<th>$\rho$; $\tau$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>44</td>
<td>4YKR</td>
<td>Glide-sp</td>
<td>0.83; 0.63</td>
</tr>
<tr>
<td>2</td>
<td>17</td>
<td>3OWD</td>
<td>Glide-sp</td>
<td>0.68; 0.47</td>
</tr>
<tr>
<td>3</td>
<td>62</td>
<td>3B27</td>
<td>rDock-std</td>
<td>0.73; 0.51</td>
</tr>
<tr>
<td>4</td>
<td>57</td>
<td>4YKY</td>
<td>Gold-chemscore</td>
<td>0.75; 0.56</td>
</tr>
</tbody>
</table>

The size of the cluster is indicated for each protocol (population). The Spearman’s and Kendall’s coefficients are reported for the selected protocols (RMSD vs score)

In the “less than one hour approach” (S1-B), we selected Glide-sp, according to the metrics resulted by the benchmark without water molecules on the protein PDB ID 3OWD (RMSD$_{\text{min}}$: 0.67 Å), chosen on the basis of its wider binding pocket, suitable, at least in principle, to accommodate different classes of compounds. The
screening was performed using Glide-sp from the Dock-Benck Virtual Screening platform and for each ligand the best pose (lowest pseudo-energy) was selected. Glide score was used to rank the 180 compounds.

Scoring results As expected, both scoring strategies S1-A and S1-B showed their ineffectiveness in the ability to correctly rank ligands in terms of their binding affinities and, also, in discriminating between true positive and true negative active compounds. In fact, as reported in Table 2 their ranking performances measured by the Kendall correlation are 0.11 and 0.16 considering S1-A and S1-B ranking, respectively. These performances suggest that the apparently more accurate S1-A ranking strategy is not superior in terms of ranking accuracy respect the fast S1-B method.

3.1.3 Stage 2: ranking prediction step

The stage 2 of the D3R Grand Challenge 2015 was characterized by the release, from the organizers, of the Hsp90 crystallographic structures used as test set in the pose prediction phase of stage 1. As in the stage 1, also here it was compared the two previously ranking methodologies (S1-A and S1-B) with the aim to rank the same 180 ligands analyzed in the stage 1 but taking into account the additional available crystallographic information.

Scoring workflow The applied workflow in stage 2 retraced the pipeline described for stage 1, and reported in Fig. 1, with few exceptions. In fact, the stage 2 of the D3R Grand Challenge 2015 was characterized by the release, from the organizers, of the Hsp90 crystallographic structures used as test set in the stage 1. Consequently, we reperformed the docking benchmark study of stage 1 using Hsp90 crystallographic structures (PDB ID: 2XDX, 4YKW, 4YKY, 4YKQ, 2YE4, 4YKT, 3R4 N, 2JJC, 2W14, 3B26, 4YKZ, 3OW6, 3B27, 3OWD, 4YKR, 4YKU, 4YXK, 4LWG, 3R4M). As previously described, also in this case all docking simulations have been carried out including the more crucial water molecules (Table SI1). The new benchmark was also interesting to retrospectively analyze the ability of the docking/scoring combinations in reproducing the new crystallographic poses and, therefore, to evaluate the goodness of our protocol selection in the stage 1. The results of the new benchmark are reported in Figure SI1 (panel A). Interestingly, the protocols selected in the stage 1 showed low RMSDmin also in the self-docking exercise confirming, again, the goodness in the identification of the docking protocol.

Moreover, the ranking prediction (S2-A) also retraced the S1-A pipeline. Again, we clustered the 180 ligands according Tanimoto’s similarity to the ligands co-crystallized as included in the benchmark (in presence of the most relevant water). In this way we obtained 7 clusters as listed in SI (Table SI4). For each of them, we carried out the docking calculation selecting the protocol according the RMSDmin and RMSDave performances but also considering the ability in discriminating the near native conformation within the family of
conformations generated in the benchmark. To highlight this, we used the Spearman index correlating the RMSD versus the score. The resulting combination of cluster, protein and protocol is detailed in Table 4.

Table 4 Combination of docking protocol and PDB ID of protein conformation used for each cluster identify in the rank prediction stage 2 (Hsp90)

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Population</th>
<th>Protein</th>
<th>S2-A Protocol</th>
<th>S2-B Protocol</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>31</td>
<td>4YKU</td>
<td>rDock-std</td>
<td>Gold-goldscore</td>
</tr>
<tr>
<td>2</td>
<td>11</td>
<td>4LWG</td>
<td>Gold-chemscore</td>
<td>Plants-chemplp</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>3B26</td>
<td>Gold-asp</td>
<td>rDock-std</td>
</tr>
<tr>
<td>4</td>
<td>21</td>
<td>4YKZ</td>
<td>rDock-std</td>
<td>Gold-goldscore</td>
</tr>
<tr>
<td>5</td>
<td>63</td>
<td>4YKR</td>
<td>Gold-asp</td>
<td>Gold-goldscore</td>
</tr>
<tr>
<td>6</td>
<td>13</td>
<td>3R4N</td>
<td>Gold-plp</td>
<td>Gold-goldscore</td>
</tr>
<tr>
<td>7</td>
<td>34</td>
<td>4YKW</td>
<td>rDock-std</td>
<td>Gold-plp</td>
</tr>
</tbody>
</table>

The size of the cluster is indicated for each protocol (population). The S2-A and S2-B differ only for the protocol selected.

Then, we extracted the best scoring pose for each ligand according to the scoring method proper of the protocol. Finally, to be able to rank ligands conformations originated from different docking protocols, we sorted all the best conformations by using a rescoring procedure with MOE-pKi function.

As previously mentioned, the second ranking submission (S2-B) is strictly linked to the first. It was designed to highlight the effect in considering the Spearman’s correlation in the protocol selection.

In more detail, all steps of this pipeline were exactly the same of the prediction S2-A except in the selection of the protocol that in this case was merely based on the RMSD_{min} and RMSD_{ave} performances obtained in the benchmark. In Table 4 is reported which protocol was assigned for each clusters.

The third and last submission in the stage 2 (S2-C) was based on the submission S1-B in stage 1 and follow the same philosophy: simplest and fastest. The workflow adopted was exactly the same of stage 1. Briefly, we performed a new benchmark on the 19 complexes (13 complexes already known at stage 1 plus the new 6 unveil complexes) removing all the water molecules. Also in this case the benchmark outputs indicated Glide-sp, as the protocol more suitable in generate the near native conformations (Figure S11, panel B). We decided to use the same protein conformation used in S1-B (PDB ID: 30WD), which is characterized by a wider binding pocket able in principle to host different classes of compounds. The screening was carried out using the Virtual Screening tool of DockBenck selection for each ligand its more stable pose.

Scoring results Unexpectedly, the scoring performances of the stage 2 have been significantly different from those observed in phase 1. The three approaches appreciably differ in terms of ranking and classification.
capability (Table 2). The more articulated methods (S2-A and S2-B) outperformed the basic approach (S2-C); in the Kendall rank correlation the three predictions S2-A, S2-B, and, S2-C scored 0.24, 0.21, and, 0.12 respectively. Whereas the score of S2-C was expected due to the fact that it has been performed with the same methodology of in stage 1, the score of S2-A is doubled. The introduction of a more suitable protein conformation has improved the performance; however, the value is still far from a desirable value. Also considering the Spearman’s rank correlation, the S2-A and S2-B outperformed S2-C with a coefficient of 0.30, 0.35, and 0.18 respectively. From the performance comparison of S2-A and S2-B is interesting to note that the use of the Spearman’s correlation in the protocol selection has slightly improved the quality in the rank classification as partially expected.

3.2 MAP4K4

3.2.1 Stage 1: pose prediction step

The challenge on MAP4K4 Stage 1—“Pose prediction” step was to predict the coordinates of 30 protein–ligand complexes and to rank the affinity of 18 of these 30 compounds referred, also in these case, as “ligand test set”. The workflow used for the pose prediction is reported in Fig. 5 (on the left), and it is divided again into four tasks: MAP4K4 complexes selection, selection of docking protocol, docking calculations and, finally, best pose selection.
Fig. 5 Workflow for posing and scoring predictions designed for the challenge on MAP4K4. In blue panel is reported the procedure used in the docking stage divided in four main tasks as reported in the discussion section: Database selection, Docking Protocol Selection, Docking Calculation, and Pose Selection. The Scoring Prediction pipeline is schematized on the green panel. The Scoring Procedure consists in three tasks strictly correlated to the posing challenge: a first ligand preparation step, docking calculation and finally the rescoring and pose selection.

MAP4K4 complexes selection Similar to what was done in stage 1 for Hsp90, we retrieved all eight ligand-MAP4K4 complexes present in the PDB (PDB ID: 4OBO [44], 4OBP [44], 4OBQ [44], 4RVT [45], 4U43 [46], 4U44 [46], 4U45 [46], and 4ZK5 [47]) in which the co-crystallized ligand will be referred again as “ligand training set”. Crystallographic structures and ligand training set were prepared for molecular docking study according to the pipeline reported in the Experimental section.
Selection of docking protocol All the 8 known complexes were submitted to a self docking benchmark within DockBench using all the 17 different docking protocols available in the tool. Unlike what has been observed for Hsp90, in this case have not been highlighted water molecules that may play a crucial role in the recognition crystallized ligands. As reported in Fig. 6a, several protocols showed good results with RMSD$_{\text{min}}$ values below 2 Å. In particular, Gold and Plants software were able to reproduce the crystal pose in the majority of cases except when Gold was coupled with chemscore function. Following these preliminary information, we selected “Gold-goldscore” and “Plants-plp” as best docking/scoring combinations.

Fig. 6a Self-docking Benchmark results obtained with DockBench on 8 complexes containing MAP4K4. The minimum RMSD values (RMSD$_{\text{min}}$) returned by the tested docking protocol (y-values) for the considered X-ray structures (x-values) for the 20 poses generated for each protocol considered. Values are color coded, blue spots identify the best obtained results. b Electrostatic Energy Fingerprints representing per-residue electrostatic contribution to interaction energy. This term was calculated for the eight training set complexes subjected to the benchmark. The interaction strength is coded in the heatmap using a red to blue palette going from a highly positive to a deeply negative potential. The calculation was performed for the most relevant residues for the binding. The blue bars corresponding to E106 and C108 highlight the relevance of these residues.
Unfortunately, in this case the chemical variability of the 30 ligands of the test set didn’t give us the opportunity to cluster them according to chemical similarity to the ligands of the training set, as in the case of Hsp90. Therefore, we adopted a different strategy to select MAP4K4 structures for docking: in particular, we took into consideration the interaction network of co-crystallized ligands in the PDB complexes, and selected those structures that conserved the same pattern for the docked poses of the test ligands. We used EEF to estimate the residues mainly involved in electrostatic interactions with the ligands.

**Docking calculation and best pose selection** The EEF of the MAP4K4 complexes suggested E106 and C108 as key residues in ligand binding; in fact, those residues are involved in strong electrostatic interactions in almost all ligand of the training set (Fig. 6b).

From this, we decided to pick the structure with lowest crystallographic resolution (PDB ID 4OBO) presenting the P-loop in a “closed” conformation. Among all docking/scoring combinations, Gold-goldscore was selected as docking protocol due to its good performance in reproducing the 4OBO ligand pose, as indicated by the corresponding RMSD\textsubscript{min} value in Fig. 6. The Virtual Screening Tool of DockBench was used to dock the 30 ligands of the test set. Using this strategy, it was possible to select a pose showing interactions with E106 or C108 for the following ligands: MAP01, MAP02, MAP03, MAP04, MAP08, MAP09, MAP14, MAP15, MAP16, MAP18, MAP19, MAP20, MAP21, MAP23, MAP26, MAP27, MAP28, MAP32 (Figure SI2).

For the remaining ligands of the test set, alternative selection strategies have been used in the selection of both MAP4K4 crystallographic structures and docking/scoring protocols. The first important alternative was to change the protein structure in which the P-loop was in an “open” conformation and the crystallographic structure coded as 4U44 was selected as the best compromise between its crystallographic resolution and its DockBench performance.

Moreover, Plants-plp combination was selected as docking/scoring protocol adopted for the 12 remaining ligands (see Fig. 6) and acceptable poses interacting with either E106 or C108 were selected for MAP05, MAP06, MAP07, MAP11, MAP22, MAP25, MAP29, MAP30, MAP31as shown in Figure SI3. At the end, were only three exceptions: MAP12, MAP13 and MAP17. Since those ligands are voluminous, for these three ligands we chose the MAP4K4 crystallographic structure coded as 4ZK5, which performed well in the benchmark and whose co-crystallized ligand is the bulkiest among the training set (Figure SI4). In this specific case, we carried out the docking simulation using Gold-goldscore combination. Unfortunately, even with these changes, we were not able to find ligand poses directly interacting with E106 and C108 and, consequently, we decided to select the best poses by visual inspection.

**Results** The superposition of the 30 predicted complexes on the corresponding X-ray crystal structures is reported in Fig. 7. In general, the proposed workflow has shown encouraging results: the pose of several
ligands were appropriately predicted but, understandably, there are a certain number of exceptions. In particular, 14 ligands were predicted with a RMSD below 2 Å and notably 11 of them below 1.5 Å. These values fall below the resolution of the crystal structures, which range from 1.59 to 3.04 Å. 4 ligands were in the range between 2 and 3 Å, whereas 12 showed a RMSD bigger than 3 Å. However, four ligands were poorly predicted (with an RMSD values >8 Å). In particular, the poses of the three ligands containing the dehydro-oxepin ring were completely wrong. The poor predictions are mainly due to the erroneous pose selection performed by visual inspection. In fact, a retrospective analysis of the docking result revealed the presence of a native like poses in the ensemble of the generated conformations. Not surprisingly, a subset of small ligands with molecular weight lower than 300 Da (MAP04, MAP20, MAP22, MAP26, MAP29, MAP30, MAP31) resulted in inaccurate poses confirming the difficulties of docking protocols with fragments in particular when docked in wide binding side and in absence of a clear shape complementarity between the ligand and the docking site. In addition, the experimental structure of four of them revealed the presence of molecules of water stabilizing their conformation (MAP04, MAP20, MAP22, MAP29).

Fig. 7 Superposition of the predicted poses (light blue) on the experimental ones (tan). RMSD values were calculated on the heavy atoms.
3.2.2 Stage 1 ranking prediction step

The aim of MAP4K4 Stage 1—“Ranking prediction” phase was to rank the affinities of 18 of the 30 compounds docked in the previous phase. The workflow used for the ranking prediction is reported in Fig. 5 (on the right).

**Scoring workflow** The selected poses for MAP01, MAP02, MAP03, MAP04, MAP05, MAP06, MAP07, MAP08, MAP09, MAP11, MAP12, MAP13, MAP14, MAP15, MAP16, MAP17, MAP18, and MAP19 were rescored with MOE using GBVI/WSA method, in order to have a homogeneous scoring method.

**Scoring results** Also in this case and as expected, the scoring strategy showed its ineffectiveness in the ability to correctly rank ligands in terms of their binding affinities (Table 2). Pearson and Kendall coefficients values (0.46 and 0.32, respectively) show a modest positive correlation between affinities and GBVI/WSA scores.

3.2.3 Stage 2 ranking prediction step

Also in this case, the stage 2 of the D3R Grand Challenge 2015 was characterized by the release, from the organizers, of the MAP4K4 crystallographic structures used as test set in the pose prediction phase of stage 1. As in the ranking prediction phase of stage 1, the aim of this stage was the ranking of the same 18 ligands but taking into account the additional available crystallographic information.

**Scoring workflow** With the release of the new 30 MAP4K4 crystallographic structures, we re-performed to DockBench analysis (see Figure SI5). Also in this case, for each of the 18 compounds which were to be analyzed we selected the pose corresponding to the best value of RMSD<sub>min</sub> obtained in the benchmark (see Table 5). Finally, the complexes were rescored, and sorted, using MOE dock_pKi scoring function.

**Table 5** Combination of docking protocol, the PDB ID of protein conformation used for each cluster identify in the rank prediction stage 2 (Hsp90)

<table>
<thead>
<tr>
<th>Complex</th>
<th>Protocol</th>
<th>RMSD&lt;sub&gt;min&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>lig-MAP01</td>
<td>Plants-chemplp</td>
<td>0.49</td>
</tr>
<tr>
<td>lig-MAP02</td>
<td>Gold-asp</td>
<td>0.31</td>
</tr>
<tr>
<td>lig-MAP03</td>
<td>Gold-goldscore</td>
<td>0.22</td>
</tr>
<tr>
<td>lig-MAP04</td>
<td>Gold-chemscore</td>
<td>0.28</td>
</tr>
<tr>
<td>lig-MAP05</td>
<td>Plants-plp95</td>
<td>0.26</td>
</tr>
<tr>
<td>lig-MAP06</td>
<td>Gold-plp</td>
<td>0.19</td>
</tr>
<tr>
<td>lig-MAP07</td>
<td>Gold-plp</td>
<td>0.27</td>
</tr>
<tr>
<td>lig-MAP08</td>
<td>Gold-chemscore</td>
<td>0.44</td>
</tr>
<tr>
<td>lig-MAP09</td>
<td>Gold-asp</td>
<td>0.19</td>
</tr>
</tbody>
</table>

**Scoring stage results** Also in this case, as expected, the ranking performance in the second stage was even less accurate than that obtained in the first stage. As reported in Table 2, Spearman’s rank coefficient (0.01)
showed absence of correlation between affinities and dock_pKi scores, and, even worse, the Kendall coefficient (-0.02) showed a tendency to negative correlation.

4. Conclusions and consideration

Our sincere feeling is that D3R Grand Challenge represented an important moment of scientific and methodological reflection regarding the real robustness of docking/scoring methodologies currently available to our scientific community. Molecular docking is certainly one of the most popular and used tools in computational medicinal chemistry and beyond. For this reason, we believe that our community must pay particular attention to point out what are the intrinsic limitations of this tool and to appropriately describe the best practice for its correct use.

In this contest, we could evaluate the predictive ability of a docking selection tool recently developed in our laboratory and called DockBench. Considering the peculiarity of the DockBench tool in facilitating the prediction of the ligand poses, we decided to concentrate our efforts in determining the best docking method able to reproduce the most accurate pose geometries.

The take home message learned from the GC2015 is that an accurate selection of both the docking protocol and protein conformation may lead in remarkable improvement of the prediction. In addition, the differences emerged in the accuracy between the two targets reveal two interesting points. First, when more data is already available as in the case of Hsp90 of which a notable number of complexes are available in the PDB, lead to better results if the similarity of between the ligand is taken into account. In particular, is not always straightforward the definition of similarity in this context and the selection of which kind of similarity can be the more appropriated (e.g. fingerprint similarity, shape similarity, substructure matching, etc.). The second point is that the role of the water molecule that improved the quality of the ligand-Hsp90 prediction. The significance of this two points are convincing us to introduce these aspects in our software also considering that the automation of these tasks into the docking pipeline would reduce the time needed to the user.

Even if the overall performance of DockBench is encouraging, from this assessment have emerged still delicate issues which limit the performance of docking/scoring algorithms and, consequently, their positive impact in the design of new drugs. Some of these are briefly summarized below:

(a) with the increasing number of docking programs (docking/scoring combinations), it becomes progressively more complex and risky to determine a priori which of these will be more accurate in reproducing a realistic poses of a ligand in its binding cavity;

(b) With the increasing number of crystal structures available in the PDB for a single protein, it becomes increasingly hazardous to determine a priori which crystallographic structure will be more appropriate to use to obtain a realistic pose of a ligand in its binding cavity;
(c) Nowadays, it is clear the crucial role of the water molecules, eventually present in the binding cavity, in determining the performance of the docking algorithms;

(d) Scoring functions are very often useless in realistically ranking a set of ligands.

As this D3R Grand Challenge has demonstrated, each docking run can be considered a singularity in a mathematical sense, or rather, a point in which a function is undefined. In fact, considering the degree of theoretical simplification of the problem we are dealing with docking and the large number of variables that define the problem itself, it is extremely difficult to determine a priori the degree of accuracy of the solution of our problem (realistic pose).

To paraphrase Albert Einstein, our take-home message may be summarized as follows: “Docking should be made as simple as possible, but not simpler.”
References

12. C C G (I Molecular Operating Environment (MOE).
17. OpenEye Scientific Software Inc. (2016) OEChem. Santa Fe, NM, USA


21. Molecular Networks GmbH CORINA. Germany


39. Li J, Shi F, Xiong B, He J Crystal Structure of the human Hsp90-alpha N-domain bound to the hsp90 inhibitor FJ4. To be published


42. Kang Y, Stuckey JA Structure of Heat Shock Protein 90 Bound to CS302. To Be Published

43. Kang Y, Stuckey JA Structure of Heat Shock Protein 90 Bound to CS319. To Be Published


Combining self- and cross-docking as benchmark tools: the performance of DockBench in the D3R Grand Challenge 2

Veronica Salmaso, Mattia Sturlese, Alberto Cuzzolin, Stefano Moro


Abstract

Molecular docking is a powerful tool in the field of computer-aided molecular design. In particular, it is the technique of choice for the prediction of a ligand pose within its target binding site. A multitude of docking methods is available nowadays, whose performance may vary depending on the data set. Therefore, some non-trivial choices should be made before starting a docking simulation. In the same framework, the selection of the target structure to use could be challenging, since the number of available experimental structures is increasing. Both issues have been explored within this work. The pose prediction of a pool of 36 compounds provided by D3R Grand Challenge 2 organizers was preceded by a pipeline to choose the best protein/docking-method couple for each blind ligand. An integrated benchmark approach including ligand shape comparison and cross-docking evaluations was implemented inside our DockBench software. The results are encouraging and show that bringing attention to the choice of the docking simulation fundamental components improves the results of the binding mode predictions.

1. Introduction

Computer-Aided Drug Design (CADD) has been extensively applied in the drug discovery process, with concrete successful examples in the market [1]. The prediction of a ligand binding mode within the targeted protein is of fundamental importance in hit identification and hit-to-lead optimization. Molecular docking is the technique of choice for the prediction of a ligand position and conformation (ligand pose) within the protein binding site and has been used since its first application in the 1980s [2].

The primary requirement to perform a docking study is the availability of the target structure, so the protein data bank [3], with its pool of more than 130,000 structures, is certainly a golden goose for structure-based drug design (SBDD). The choice of the best protein structure may condition the success of the docking simulation, with holo binding sites giving the best performances [4]. The number of protein experimental structures is rapidly increasing, and more and more often a variety of three-dimensional structures are available for the same target. This puts the modeler in front of the dilemma of the protein choice in docking simulations.
In addition, various docking methods are available nowadays, with a list of more than 50 software in the Click2Drug directory [5]. The accuracy of docking algorithms and especially of scoring functions is sensitive to the dataset of proteins and compounds.

Thus, a plethora of variables should be considered before performing a docking simulation and, citing Isaac Asimov, “If knowledge can create problems, it is not through ignorance that we can solve them”, it is valuable to consider the whole pool of variables to assess the performance of the docking procedure that better suits the dataset of interest.

In this framework, docking challenges are precious because offer the possibility to realize limits and strong points of this computational technique, and to analyze the performance of computer-aided pipelines. D3R Grand Challenge 2 consisted of a two-stage process, comprising the pose and ranking prediction of a pool of 36 and 102 compounds, respectively, on the Farnesoid X receptor (FXR) target. FXR is a nuclear receptor implicated in bile acids, lipids, and glucose homeostasis; this makes it a drug target against cholestasis and lipids and glucose dysregulation [6].

Taking advantage of our experience with the previous Grand Challenge 2015 [7, 8], we have strengthened our previous pipeline improving the strategy to identify the best performing combination of docking protocol and protein structure for each ligand to be predicted. The selection of the docking protocol was assessed through a docking benchmark, which enabled to compare the performances of 16 docking-algorithm/scoring function couples. Moreover, a major effort was put on the choice of the protein structure, and a combined approach made up of ligands shape similarity evaluation to individuate congeneric series and cross-docking assessment was employed. The DockBench [9] platform was exploited and the whole procedure was set up to be completely automatic, with the aim to integrate the new-developed tools into the DockBench software.

The results of this pipeline are encouraging if compared to a simpler procedure consisting in the use of a single apoprotein and a single docking protocol for the whole pool of compounds. The detailed description of the adopted procedure will be described through this work.

2. Experimental section

2.1 Overview of the work

The main focus of this work was the pose prediction challenge, and in this framework a combined strategy was employed, comprising a choice of the protein and docking protocol on the basis of ligand similarity and on the results of a benchmark exercise.
Throughout this work, a pool of crystallographic structures (proteins and ligands) were used for the benchmark and are referred to as “training-set”, while the ligands object of the challenge are named “blind-challenge ligands”.

Two predictions were made, and the pipeline can be summarized as follows.

**Prediction 1**

**Pose prediction phase**

1. A training set of structures of the target was retrieved from the protein data bank (PDB) [3].

2. The training set complexes were clustered according to the co-crystallized ligands shape similarity.

3. A three-phase benchmark strategy was conducted to evaluate the “protein structure - docking protocol” performances:
   - **CLUSTER CROSS-DOCKING**: a cross-docking benchmark was executed among the complexes of each cluster, to individuate the “protein structure-docking protocol” couple able to better reproduce the crystallographic conformations of the ligands (similarly shaped) belonging to the cluster.
   - **TOTAL CROSS-DOCKING**: a cross-docking exercise was executed among the whole pool of PDB structures, to individuate the protein and docking-protocol able to averagely better reproduce the crystallographic conformation of all the training-set ligands (disregarding shape similarity).
   - **SELF-DOCKING**: as regards the clusters populated with just one PDB structure, a self-docking benchmark was performed to evaluate which docking protocol succeeded in better replicating the crystallographic binding mode.

4. The blind-challenge ligands were compared in terms of shape similarity to the representative of each cluster, previously selected in the benchmark phase. In this way, the blind challenge ligands were distributed over the pre-defined clusters, each one associated with the protein structure-docking protocol couple selected in the benchmark phase.

5. Blind-challenge ligands with no significant similarity to any of the training-set ligand were identified as outliers. They were associated with the protein structure-docking protocol couple selected for its highest performance in the total cross-docking exercise.

6. Each ligand was docked to the respectively selected protein with the selected docking protocol.
7. The top 5 scoring poses were selected and submitted.

**Scoring phase**

1. The highest scoring pose for each ligand was retained and submitted to a common rescoring phase.
2. The complexes obtained by docking simulation were prepared and submitted to molecular dynamics (MD) simulations.
3. MM-GBSA [10] profiles were computed along the trajectories. The MM-GBSA average value was used as common re-scoring and ranking method for all the ligands.

**Prediction 2**

**Pose prediction phase**

1. A docking simulation was performed for the blind-challenge ligands using the apoprotein crystallographic structure provided by the organizers, and employing the docking protocol that showed better performance in the total cross-docking job.
2. The top 5 scoring poses were selected and submitted.

**Scoring phase**

1. Docking score was used as a ranking method.

**2.2 Hardware**

Docking studies were carried out on a 200 cores CPU cluster based on Ubuntu operating system (distribution 14.04, 64 bit) under the Network File System (NFS) service. MD simulations were performed using Acemd [11] on a 20 NVIDIA GTX graphics cards GPU cluster.

**2.3 Preparation of training set ligand–protein complexes**

The protein preparation tool of MOE [12] was used to fix crystal structures problems, such as prediction of coordinates of missing atoms of partially solved residues. Co-crystallized solvent molecules and impurities (such as co-solvents) were removed, and only protein and ligand coordinates belonging to chain A of the crystal structures were retained. The Protonate-3D tool of MOE [13] was used to assign protonation states (assuming pH 7.4) of protein and the respective co-crystallized ligand.

**2.4 Ligands preparation**

Blind-challenge ligands were prepared using the LigPrep Tool of Schrödinger [14], retaining specified chiralities and without generating tautomers. Strong acids were deprotonated and strong basis protonated using the Wash Tool of MOE. MMFF94 Force Field has employed for ligands minimization.
2.5 Ligand similarity

RDKit [15] Shape Tanimoto Distance was employed to evaluate shape similarity among the training set ligands, as reported below.

OMEGA [16] of the OpenEye suite was used to generate 5 conformations for each blind-challenge ligand, which were submitted to the ROCS tool [17] for shape comparison to the representative of each training-set cluster.

2.6 Molecular docking

DockBench 1.0.5 [9] was used for the self-docking simulation and analysis, while the cross-docking computation and evaluation were carried out with an in-house tool that will be implemented in DockBench.

The following software packages were used to perform molecular docking calculations: AutoDock 4.2.5.1 [18], Glide 6.5 [19, 20], GOLD 5.2 [21], MOE 2015.1001 [12], PLANTS 1.2 [22, 23], rDock [24]. Either for the training set self-docking and cross-docking or for the blind-challenge ligands docking, the DockBench default parameters were used for docking simulations.

2.7 Molecular dynamics simulations

The ligand–protein complexes were prepared for MD simulations with AmberTools14 [25], assigning Gasteiger charges [26] and General Amber Force Field (GAFF) [27] parameters to the ligands and Amber14 partial charges and parameters to the proteins. Each system was solvated with explicit waters (TIP3P model) resulting in a tetragonal box with boundaries at least 11 Å far from any atom of the complex. Each system was neutralized adding Na⁺/Cl⁻ ions to a final concentration of 0.1 M. Each system was subjected to 300 steps of conjugate-gradient minimization and to 100 ps NVE and 500 ps NPT equilibration, applying harmonic positional constraints (1 kcal mol⁻¹ Å⁻²) on protein and ligands atoms. The pressure was maintained to 1 atm by Berendsen barostat and the temperature to 310 K by a Langevin thermostat.

Subsequently, three 2 ns unconstrained MD simulations in the NVT ensemble were conducted for each complex.

All MD simulations were carried out with the ACEMD engine, with a time-step of 4fs, by handling the nonbonded long-range Coulomb interactions with the particle mesh Ewald summation method (PME) [28, 29] with a cutoff distance of 9 Å and a switching distance of 7.5 Å.

2.8 MMGBSA calculations

AmberTools14 was used to perform MM-GBSA calculations, using a GB model developed by Onufriev-Bashford-Case [30] (igb = 5), a salt concentration of 0.1 M and calculating the surface area with the LCPO
model [31]. 50 snapshots (every 40 ps) were collected for each trajectory, and MM-GBSA was computed along each simulation. Finally, the average value among the three replicas of the same complex was computed and used as ranking score.

3. Results and discussion

D3R Grand Challenge 2016 consisted of a two-stage process: stage 1 was devoted to the prediction of the binding mode of 36 blind compounds (FXR1-FXR36) and to the ranking of a pool of 102 blind compounds (FXR1-FXR102), containing the 36 ligands previously mentioned. Stage 2 asked for a new ranking prediction of the same pool of 102 compounds, once the FXR1-FXR36 crystallographic structures had been unveiled by the D3R Grand Challenge organizers.

The major effort of our work was devoted to the pose prediction task, using the docking protocol-protein structure couple able to better reproduce the crystallographic binding mode of ligands similarly shaped to those to be predicted. Moreover, the strong point of this work is that it is completely automated: from the choice of the docking protocol and protein structure to use to the selection of the final docking poses, the pipeline is completely performed without user’s intervention.

Fig. 1 Workflow of the pipeline employed for pose prediction phase of Prediction 1
The results section is organized into two paragraphs, devoted to two separate predictions and corresponding submissions to the GrandChallenge. Prediction 1 involved the multi-step benchmark procedure described in the workflow (Fig. 1), and the results of each phase are described hereinafter, while prediction 2 employed a simpler strategy.

3.1 Prediction 1

3.1.1 Pose prediction

Training-set construction

The starting point of this work was the provision of crystal-structures of FXR to use in the docking benchmark studies. 26 complexes of human FXR with small organic ligands were retrieved from the PDB site: 1OSH [32], 3BEJ [33], 3DCT [34], 3DCU [34], 3FLI [35], 3FXV [36], 3GD2 [37], 3HCS [38], 3HC6 [38], 3L1B [39], 3OKI [40], 3OKH [40], 3OLF [41], 3OMK [41], 3OMM [41], 3OOF [41], 3OOK [41], 3P88 [42], 3P89 [42], 3RUT [43], 3RUU [43], 3RVF [43], 4OIV [44], 4QE6 [45], 4QE8 [45], 4WVD [46].

PDB structures were filtered on the basis of the degree of completeness of the protein structure. Proteins lacking the coordinates of more than 15 residues were removed, causing the exclusion of 1OSH, 3L1B, and 4OIV structures.

In addition, structure 3OKH was removed by the collection because has two co-crystallized ligands, making it difficult to use for the docking benchmark. Also structures 4WVD was not considered because the co-crystallized high-molecular weight compound was difficult to use for docking.

Training-set ligands clusterization

Training-set proteins were clustered on the basis of the shape similarity among their co-crystallized ligands. The aim of the ligand-shape-based clustering is to divide proteins according to the footprint left by ligands on the binding site. This information will be later employed to associate each blind-challenge ligand to its best-hosting protein.

All FXR-ligand complexes were superposed by protein alignment, and the co-superposed ligands were used for shape comparison. RDKit Shape Tanimoto Distance was computed between each couple of compounds, resulting in a distance matrix. The matrix was subjected to scikit-learn [47] DBSCAN Clustering algorithm [48], using a cutoff of 0.45: structures with a distance value lower than the cutoff fell in the same cluster. The ligands, and consequently the relative protein structures, were clustered in 6 groups, as reported in Table 1.
Table 1 List of PDB training-set structures subdivided into clusters according to 3D shape similarity

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Training-set structures - PDB ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3DCT, 3DCU, 3FXV, 3GD2, 3HC5, 3HC6, 3P88, 3P89, 3RUT, 3RUU, 3RVF</td>
</tr>
<tr>
<td>2</td>
<td>4QE8</td>
</tr>
<tr>
<td>3</td>
<td>3FLI</td>
</tr>
<tr>
<td>4</td>
<td>4QE6</td>
</tr>
<tr>
<td>5</td>
<td>3BEJ</td>
</tr>
<tr>
<td>6</td>
<td>3OKI, 3OLF, 3OMK, 3OMM, 3OOF, 300K</td>
</tr>
</tbody>
</table>

Cluster 2–5 are single-populated, while cluster 1 and 6 are characterized by more ligands having a similar chemical structure, as can be appreciated by Fig. 2: ligands of cluster 1 share an isoxazole-4-yl-methoxymultiaryl-carboxylic acid scaffold and cluster 6 is characterized by a (2-phenyl-benzimidazol-1-yl)-2-cyclohexyl-ethanamide moiety.

![Fig. 2](Image)

**Docking-benchmark**

The docking-benchmark was set up with a triple strategy, described below. In particular, the training-set clusters populated by more than one structure (clusters 1 and 6) were subjected to an intra-cluster cross-docking exercise. The structure corresponding to mono-populate clusters (structures 4QE8, 3FLI, 4QE6 and 3BEJ respectively of clusters 2–5) were used for a self-docking benchmark. Finally, the whole pool of training-set structures was used for a total cross-docking exercise.
Cross-docking

Given a pool of crystallographic structures of the same protein in complex with different ligands, cross-docking consists in docking each co-crystallized ligand to all the protein structures. The aim of this operation is to evaluate the performance of different proteins to host various ligands. The evaluation is made by computing the geometrical deviation of the predicted binding mode as compared to the binding mode on the crystallographic complex. The protein able to better host the higher number of ligands with a conformation similar to the crystallographic one could be a good candidate to use in a virtual screening simulation.

A cross-docking platform has been developed to automatically perform the cross-docking simulation and analysis among a group of PDB structures. Along with guiding the choice of the best protein structure, the tool aims to individuate the best docking/scoring combination for the subsequent virtual screening simulation. For this reason, the tool enables to automatically repeat the cross-docking simulation with different docking/scoring protocols.

After superposing the crystallographic complexes, the ligands were extracted, merged in a unique database and docked to each protein structure with 16 docking/scoring combination protocols, employing DockBench 1.0.5 default parameters: AutoDock-Generic Algorithm (GA), AutoDock-Lamarckian Genetic Algorithm (LGA), AutoDock-Local Search (LS), Glide-Standard Precision (SP), GOLD-ASP, GOLD-Chemscore, GOLD-Goldscore, GOLD-PLP, MOE-Affinity dG, MOE-GBVI/WSA, MOE-London dG, Plants-ChemPLP, Plants-PLP, Plants-PLP95, rDock with (SOLV) or without (STD) the desolvation potential.

For each of the 16 docking protocols and N PDB structures, a database of N-1 ligands was docked to each of the N protein structures (the self-ligand was excluded for the cross-docking computation to follow a strict cross-docking procedure). 20 poses (or up to 20 poses in the case of Glide) were generated for each of the N(N-1) ligand–protein couples and the RMSD values between the five top-scoring poses and the crystallographic conformation of the ligand were computed. The evaluation of each protein-docking protocol couple was performed considering the mean RMSD of the five top-scoring poses of each of the N-1 ligands docked on the same protein with the same protocol (i.e. average among S(N-1) RMSD values); this value was called Top5RMSDave. The Top5RMSDave was computed for each protein-docking protocol couple resulting in a 16N-items matrix. The matrix was rendered in the form of a heat map where each row represents a docking/scoring couple and each column represents a different PDB structure. The Top5RMSDave is rendered by a colorimetric scale going from blue to red for values from 0 to 20 Å. The protein-docking protocol couple was chosen as the one which minimized the Top5RMSDave value.
Cluster cross-docking

Clusters 1 and 6 were populated by more than one structure, so they were subjected to an intra-cluster cross-docking evaluation. This means that the results of the cross-docking simulation were organized considering the structures belonging to clusters 1 and 6 as two separate pools, and the Top5RMSDave values were computed just for the structures belonging to the same cluster. The aim of this was to choose the protein able to better host the ligands sharing a similar shape to the co-crystallized one.

The results of the Cluster cross-docking phase are reported in Fig. 3.

![Fig. 3 Cluster cross-docking benchmark results. Structures belonging to cluster 1 (grouped together by a gray square) and cluster 6 (grouped together by a purple square) were subjected to two independent intra-cluster cross-docking runs. The mean RMSD for the five top-scoring poses of all the ligands docked to the same protein (Top5RMSDave) are reported in the heat-map. The Top5RMSDave for each protein (x-values) and docking protocol (y-values) is represented by a colorimetric scale, going from blue to red from 0 to higher RMSD values. A white circle highlights the selected protein-docking protocol couple for each cluster.]

3GD2 protein structure with GOLD-PLP docking protocol was chosen for cluster 1 and 3OMK with GOLD-PLP for cluster 6 because they respectively obtained the lowest Top5RMSDave value within their cluster.

Self-docking

Clusters 2–5 are populated by a single PDB structure, so the Cluster Cross-Docking strategy could not be adopted for them. Instead, a self-docking benchmark was performed with these structures, employing DockBench 1.0.5 and the same docking protocols mentioned above for the cross-docking phase. 20 poses (or up to 20 poses in the case of Glide) were generated for each ligand and the lowest RMSD among the poses (RMSDmin) was used to evaluate the docking performance. The RMSDmin plot obtained by DockBench is
reported in Fig. 4, where the RMSDmin is reported for each protein (columns)—docking protocol (rows) couple and rendered by a colorimetric scale going from blue to red for values from 0 to 20 Å.

The following docking protocols were selected: Plants-PLP95 for protein 4QE8, rDock-SOLV for 3FLI, GOLD-Goldscore for 4QE6 and GOLD-ASP for 3BEJ.

**Fig. 4** Self-docking benchmark results. The minimum RMSD values (RMSDmin) returned by each docking protocol (y-values) for each training-set PDB structure (x-values) are represented by a colorimetric scale, going from blue to red from 0 to higher RMSD value. A white circle highlights the selected docking protocol for structure 4QE8 (cluster 2, red squares), 3FLI (cluster 3, blue squared), 4QE6 (cluster 4, green squared), 3BEJ (cluster 5, yellow squared).

**Fig. 5** Total cross-docking benchmark results. The mean RMSD for the five top-scoring poses of all the ligands docked to the same protein (Top5RMSDave) are reported in the heat-map. The Top5RMSDave for each protein (x-values) and docking protocol (y-values) is represented by a colorimetric scale, going from blue to red from 0 to higher RMSD values. A white circle highlights the selected protein-docking protocol couple.
Total cross-docking

In addition, a cross-docking simulation was performed considering the total pool of available PDB structures (21 complexes), in order to evaluate the protein that could better host differently shaped compounds.

This simulation resulted in the choice of 3GD2 protein along with GOLD-Goldscore protocol, as can be seen in the Top5RMSDave heat map reported in Fig. 5.

Blind-challenge ligands clusterization

After the selection of one PDB structure per cluster, blind-challenge ligands were screened against the representatives of each cluster using the ROCS tool. In particular, TanimotoComboShapeSimilarity was computed between 5 conformations of each blind-challenge ligand and the representatives of each cluster. A 0.8 TanimotoComboShapeSimilarity cutoff was used to distribute blind-challenge ligands to the clusters identified by the training-set ligands.

Ligands with similarity lower than the cutoff to any of the training-set clusters representatives were collected in the so-called “outliers” cluster, composed of 29 compounds.

The so-defined blind-challenge ligands clusters were associated with the protein-docking protocol couple selected in the docking benchmark phase for each cluster. The outliers cluster was associated with the protein-docking protocol couple identified as a winner in the total cross-docking exercise, because of the higher efficacy of this couple in docking ligands with a different shape to a conformation close to the experimental one.

Table 2 summarizes the results of blind-challenge ligands clusterization, reporting the protein and the docking protocol associated with each ligand.

Cluster 6 is populated by 57 ligands, with 47 of them sharing the 1,2-disubstituted benzimidazole scaffold with the 3OMK ligand. Cluster 1 hosts two ligands presenting the isoxazole-4-yl-methoxy-multiaryl-carboxylic acid typical of the training-set cluster. As regards the other clusters, the compounds do not have a common scaffold with the cluster representative, even if they were associated to it by 3D shape similarity.

20 poses of each blind-challenge ligand were obtained by docking to the associated protein with the selected docking protocol and the 5 top scoring poses of compounds FXR1-36 were automatically selected and submitted to the D3R Grand Challenge 2.
<table>
<thead>
<tr>
<th>Cluster ID</th>
<th>Selected protein PDB ID</th>
<th>Selected docking/scoring protocol</th>
<th>Benchmark Strategy</th>
<th>Blind-challenge compounds</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3GD2</td>
<td>GOLD-PLP</td>
<td>Cluster cross-docking</td>
<td>FXR33*, FXR65</td>
</tr>
<tr>
<td>2</td>
<td>4QE8</td>
<td>Plants-PLP95</td>
<td>self-docking</td>
<td>FXR23, FXR101</td>
</tr>
<tr>
<td>3</td>
<td>3FLI</td>
<td>rDock-SOLV</td>
<td>self-docking</td>
<td>FXR3, FXR5</td>
</tr>
<tr>
<td>4</td>
<td>4QE6</td>
<td>GOLD-Goldscore</td>
<td>self-docking</td>
<td>FXR34</td>
</tr>
<tr>
<td>5</td>
<td>3BEJ</td>
<td>GOLD-ASP</td>
<td>self-docking</td>
<td>FXR16, FXR79, FXR92, FXR94, FXR97</td>
</tr>
<tr>
<td>6</td>
<td>3OMK</td>
<td>GOLD-PLP</td>
<td>Cluster cross-docking</td>
<td>FXR2, FXR4, FXR6, FXR7, FXR8, FXR9, FXR13, FXR14, FXR17, FXR18, FXR19, FXR20, FXR21, FXR22, FXR24, FXR25, FXR26, FXR27, FXR28, FXR29, FXR30, FXR31, FXR32, FXR35, FXR36, FXR37, FXR39, FXR40, FXR42, FXR46, FXR47, FXR48, FXR49, FXR50, FXR51, FXR52, FXR53, FXR54, FXR55, FXR56, FXR57, FXR58, FXR59, FXR60, FXR61, FXR62, FXR63, FXR64, FXR66, FXR67, FXR68, FXR69, FXR70, FXR71, FXR72, FXR91, FXR93, FXR95, FXR96, FXR98, FXR100</td>
</tr>
<tr>
<td>outliers</td>
<td>3GD2</td>
<td>GOLD-Goldscore</td>
<td>Total cross-docking</td>
<td>FXR1, FXR10, FXR11, FXR12, FXR15, FXR38, FXR41, FXR43, FXR44, FXR45, FXR73, FXR74, FXR75, FXR76, FXR77, FXR78, FXR80, FXR81, FXR82, FXR83, FXR84, FXR85, FXR86, FXR87, FXR88, FXR89, FXR90, FXR99, FXR102</td>
</tr>
</tbody>
</table>

The protein structure and docking/scoring protocol used to dock the members of each cluster are indicated. The “Benchmark Strategy” column shows the benchmark method employed for the choice of protein structure and docking/scoring method. Compound FXR1-FXR36 are indicated in bold to highlight that pose prediction was required for them.

*Compound FXR33 was removed from pose prediction evaluation.

**Poses evaluation**

Poses evaluation was performed in terms of mean RMSD values between the predicted poses and the crystal structures unveiled by the Grand Challenge organizers. In particular, three RMSD values were employed: the mean RMSD of the first top scoring poses of each ligand (RMSDpose1), the mean RMSD of the whole pool of submitted poses (5 for each compound) (RMSDave) and the mean RMSD of the lowest RMSD poses of each ligand (RMSDbest). The mean values over the pool of 35 ligands (FXR1 to FXR36,
excluding compound FXR33 because of crystal artifacts) results in 3.92, 3.81 and 3.25 Å (Table 3), respectively.

The superposition of each ligand pose to the relative crystal structure unveiled by the Grand Challenge organizers is shown in Fig. 6 (pose 1) and in Figs. SI1, SI2, SI3, SI4 (poses 2–5).

**Table 3** Evaluation of the pose prediction results in terms of mean RMSD of the first top scoring poses of each ligand (RMSDpose1), mean RMSD of the whole pool of submitted poses (5 for each compound) (RMSDave) and mean RMSD of the lowest RMSD poses of each ligand (RMSDbest), for prediction 1 and 2

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Submission ID</th>
<th>RMSDpose1 (Å)</th>
<th>RMSDave (Å)</th>
<th>RMSDbest (Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>gff3a</td>
<td>3.92 Å</td>
<td>3.81 Å</td>
<td>3.25 Å</td>
</tr>
<tr>
<td>2</td>
<td>knz3v</td>
<td>7.67 Å</td>
<td>6.09 Å</td>
<td>4.84 Å</td>
</tr>
</tbody>
</table>

Results are taken from D3R Challenge web site: https://drugdesigndata.org/about/grand-challenge-2

Fig. 6 Superposition of the first top scoring predicted poses (pose1) of compounds FXR1–FXR36, prediction 1 (light blue sticks), on the experimental ones (red sticks). Compound FXR33 was excluded from the comparison because of crystal artifacts. Compounds FXR1, FXR2, FXR3, FXR4, FXR10, FXR11, FXR13, FXR15, FXR16, FXR18, FXR21, FXR22, FXR23, FXR25, FXR26, FXR28, FXR29, FXR32, FXR34, and FXR35 were superposed to chain A, compound FXR7 to chain AA, compounds FXR12 and FXR14 to chain AB, compounds FXR5, FXR8, FXR17, FXR19, FXR20, FXR24, FXR27, FXR30, FXR31 and FXR36 to chain C, compound FXR6 to chain CA, compound FXR9 to chain E, as in the results provided by the GrandChallenge organizers. RMSD values are reported for each pose. Each compound name is underscored by a red, blue, green, yellow, purple, brown line, meaning it belongs to cluster 2–6, “outliers”, respectively.
Cluster 6 played a major role in lowering the whole RMSD values. In fact, as can be appreciated by Fig. 7a, the mean RMSD values relative to cluster 6 are lower than the values of the other clusters: the mean RMSDpose1, RMSDave, and RMSDbest of the 25 ligands of cluster 6 are respectively 2.66, 2.48 and 1.95 Å, while the mean values of the remaining 10 compounds (organized in clusters 2, 3, 4, 5 and outliers) are 7.06, 7.14 and 6.50 Å, respectively. Focusing on cluster 6, 20 ligands (i.e. FXR6, FXR7, FXR8, FXR9, FXR13, FXR14, FXR19, FXR21, FXR22, FXR24, FXR25, FXR26, FXR27, FXR28, FXR29, FXR30, FXR31, FXR32, FXR35, FXR36) present at least one pose with RMSD lower or near 2 Å, as summarized in Fig. 7a. It is valuable to notice that these ligands constitute the set of compounds with common 1,2-disubstituted benzimidazole scaffold, typical of the respective training set cluster. Compound FXR20 is an exception, since, even sharing the before mentioned structure, was not accurately predicted. Among these 20 compounds, the best pose coincides with the top scoring one just for compounds FXR13, FXR19, FXR21. However, in most of the cases, there are no big RMSD differences between the top scoring pose and the lowest RMSD one, except for compounds FXR9, FXR22, and FXR32. So, given our completely automatic procedure, the selection of just one pose for each ligand would have resulted in a worse scenario. In some cases, the automatic procedure gave discouraging results. Here compound FXR23 is reported as an example, with all poses showing RMSD values higher than 18.0 Å. All the 5 selected poses fall out of the binding site of the receptor. However, considering the whole pool of 20 poses predicted by PLANTS-plp95, the 11th -scoring pose is located within the protein binding site (Fig. S15). Thus, it is clear that selecting poses just on the basis of the score may be not a good strategy. Honestly, pose shown in Fig. S15 is still far from the experimental one, with an RMSD of 6.5 Å. Problems related to protein conformations could be excluded since the conformation of the binding site of the unveiled crystallographic structure and the one used for docking (4QE8) do not differ a lot (Cα-RMSD 1.7 Å), so the bad prediction is due to docking bad sampling.

3.1.2 Ligand scoring

Scoring protocol

As regards Stage 1 scoring phase, the first top scoring pose of each blind-challenge (FXR1-FXR102) ligand was submitted to MD simulation. Each MD system was prepared as reported in the Molecular Dynamics Simulations paragraph of the experimental section and three replicas of 2 ns MD simulations were performed. The average MM-GBSA value over the three simulations was used as rescoring method.

As regards Stage 2, the same rescoring method was adopted, but docking poses of compounds FXR1-FXR36 were substituted by the unveiled crystal structures. The crystallographic complexes were prepared and submitted to the same protocol of MD simulation and MM-GBSA evaluation described before.
Ranking evaluation

Kendall’s and Spearman’s coefficients were used for ranking evaluation. Both values were near 0 and slightly negative (Table 4) as regards Stage 1, meaning a disagreement between our prediction and the experimental binding data. Stage 2 turned out to be slightly improved in the ranking prediction, with a feeble positive value both for Kendall’s and Spearman’s coefficients, which are still far from the unity value meaning good correlation.

Fig. 7 Representation of the RMSDs of the top score pose (RMSDpose1, blue histograms), the average RMSDs over the 5 poses (RMSDave, red histograms) and the minimum RMSDs (RMSDbest, green histograms), decomposed for each ligand. Values relative to prediction 1 and 2 are shown in a and b, respectively. The graph on the top of both panels represents the mean RMSDpose1, RMSDave and RMSDbest values over the compounds belonging to cluster 2 (red squares), 3 (blue squared), 4 (green squared), 5 (yellow square), 6 (purple squared) and “outliers” (brown squared)
### Table 4: Evaluation of the ranking results in terms of Kendall’s and Spearman’s coefficients, as provided by Grand Challenge 2 organizers

<table>
<thead>
<tr>
<th>Stage</th>
<th>Submission ID</th>
<th>Kendall’s Tau</th>
<th>Kendall’s Tau Error</th>
<th>Spearman’s Rho</th>
<th>Spearman’s Rho Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-prediction 1</td>
<td>cs2lm</td>
<td>-0.0896</td>
<td>0.0595</td>
<td>-0.146</td>
<td>0.0894</td>
</tr>
<tr>
<td>2-prediction 1</td>
<td>vxvhq</td>
<td>0.185</td>
<td>0.0651</td>
<td>0.274</td>
<td>0.0936</td>
</tr>
<tr>
<td>1-prediction 2</td>
<td>jr0oc</td>
<td>-0.394</td>
<td>0.0574</td>
<td>-0.549</td>
<td>0.0762</td>
</tr>
</tbody>
</table>

Moreover, Kendall’s and Spearman’s coefficients were computed for the subset of 35 compounds (FXR1-FXR36, excluding compound FXR33) whose docking pose was substituted by the crystallographic one during Stage 2. They were respectively −0.0655 and −0.0944 as regards Stage 1, while 0.173 and 0.269 for Stage 2. Thus, a shift from negative to positive coefficients can be noted when considering crystallographic poses instead of docking poses. This observation suggests that poorly predicted poses were part of the problem in Stage 1 ranking prediction. Nevertheless, the correlation coefficients values are still low after Stage 2, meaning that the employed MM-GBSA score is not correlated to the experimental binding affinity of the ligands. However, the prediction of compounds affinity still remains a very delicate issue, as demonstrated by Kendall’s coefficient of 0.46 as best GrandChallenge value (Submission ID: f2wjs).

### 3.2 Prediction 2

#### 3.2.1 Pose prediction

**Pose prediction protocol**

Along with the above-described procedure, a simpler pipeline was adopted to predict the binding mode of the blind challenge ligands, with the aim to compare these results with that of the previous procedure. The crystallographic structure of the apo form of the Farnesoid receptor was used for docking simulation. The results of the total cross-docking exercise were exploited and the GOLD-Goldscore protocol was chosen since it gave the best performance in terms of lowest Top5RMSDave value.

20 poses of each blind-challenge ligand were obtained by docking and the 5 top scoring poses of compounds FXR1-36 were submitted to the D3R Grand Challenge 2.

**Poses evaluation**

In the case of prediction 2, the mean RMSD of the first top scoring poses of each ligand (RMSDpose1), the mean RMSD of the whole pool of submitted poses (5 for each compound) (RMSDave) and the mean RMSD
of the lowest RMSD poses of each ligand (RMSDbest) are 7.67, 6.09 and 4.84 Å (Table 3), respectively. These values are higher than those of prediction 1, in particular, RMSDpose1 is nearly twofold the previous value.

Figure 7b shows that the high deviation from X-ray structures is spread over the whole pool of predicted poses, with compounds FXR5, FXR12, and FXR17 constituting few singular exceptions.

The results decomposition into the previously defined clusters (Fig. 7b) shows that the mean RMSD values of each cluster are very similar. The 8.11 Å RMSDpose1, 6.11 Å RMSDave and 4.53 Å RMSDbest values for cluster 6 are in line with the values of the other clusters and are more than double the mean values of cluster 6 of prediction 1.

The superposition of each ligand pose to the relative crystal structure unveiled by the Grand Challenge organizers is shown in Figs. SI6, SI7, SI8, SI9, SI10 (poses 1–5).

3.2.2 Ligand scoring

Scoring protocol

The first top-scoring poses were considered for ranking evaluation, and Goldscore Fitness score was used as a scoring method.

Ranking evaluation

Kendall’s and Spearman’s coefficients are both negative (Table 4), meaning a wrong correlation among the experimental binding data and the ranking based on the docking scoring function.

4. Conclusions

D3R Grand Challenge gave us the possibility to test an in-house totally automatic procedure to predict ligands binding modes to their protein target. An overall high mean RMSD value of the predicted poses from the X-ray structures is compensated by the comparison of the results between the two protocols we have employed. In fact, the application of a combined procedure which took into account ligand shape similarity and results of a cross-docking-benchmark improved the results, giving a reduction of the average RMSD (RMSDpose1) of the first top-scoring poses of nearly a half. Evaluating the ability of a crystallographic protein structure to host compounds that share a similar chemical shape constituted a strategy that ameliorated the prediction of the binding mode for a subset of 20 compounds on a pool of 35. This strategy seems to help more when ligands share a common scaffold with the compound co-crystallized with the protein used for docking. In fact, this is the case of the 20 before-mentioned compounds, which have a 1,2-disubstituted benzimidazole scaffold in common with the 3OMK compound. For this reason, the entire pipeline presented in this work will be implemented in a new version of DockBench, adding different chemical similarity methods in addition to shape-similarity in the clustering phase.
Moreover, the cross-docking work is highly demanding, but the integration of the described process into automatic pipeline lays the foundations for the application of the same protocol to virtual screening campaigns.

Given the high total mean RMSD values, there are great margins of improvement. Poor results may be obtained relying just on the docking score for pose selection; computing the interaction network of the predicted bound state and comparing this interaction fingerprint with that of a true positive ligand could be more valuable. In addition, the role of key water molecules should be taken into account in the different passages of the proposed pipeline.
References


Sulfonamido-derivatives of unsubstituted carbazole as BACE1 Inhibitors

Simone Bertini, Elisa Ghilardi, Valentina Asso, Filippo Minutolo, Simona Rapposelli, Maria Digiacomo, Giuseppe Saccomanni, Veronica Salmaso, Mattia Sturlese, Stefano Moro, Marco Macchia, Clementina Manera


Abstract

A novel series of variously substituted N-[3-(9H-carbazol-9-yl)-2-hydroxypropyl]-arylsulfonamides has been synthesized and assayed for β-Secretase (BACE1) inhibitory activity. BACE1 is a widely recognized drug target for the prevention and treatment of Alzheimer’s Disease (AD). The introduction of benzyl substituents on the nitrogen atom of the arylsulfonamide moiety has so far led to the best results, with three derivatives showing IC50 values ranging from 1.6 to 1.9 μM. Therefore, a significant improvement over the previously reported series of N-carboxamides (displaying IC50’s ≥ 2.5 μM) has been achieved, thus suggesting an active role of the sulfonamido-portion in the inhibition process. Preliminary molecular modeling studies have been carried out to rationalize the observed structure-activity relationships.

Alzheimer’s disease (AD) is the most common form of dementia, which occurs predominantly in older people (over 65 years of age). It is a progressive and irreversible neurodegenerative disorder, which compromises cognitive functions (memory, thinking, reasoning) and behavioral skills in such a way as to interfere with daily life and with the fulfillment of the simplest tasks. The main neuropathological features of AD are extracellular senile plaques, essentially made of amyloid β peptide (Aβ) [1], and intracellular neurofibrillary tangles, caused by the aggregation of phosphorylated tau proteins [2].

The Aβ peptide is generated through proteolytic processing of the amyloid precursor protein (APP) first by β-secretase (BACE1) followed by γ-secretase. In particular, the cleavage operated by BACE1 produces the secreted amino-terminal part of APP (sAPPβ) and the membrane-bound carboxy-terminal fragment, which is 99 amino acids in length (C99). γ-Secretase subsequently cleaves the C99 fragment, releasing the Aβ peptide, which aggregates to form toxic amyloid plaques in the brain [3]. It has been demonstrated that BACE1 knockout (BACE1 −/−) mice are unable to generate C99 and Aβ, are viable [4], and drastically ameliorate the pathology when crossed with APP transgenic mice (a mouse model of AD) [5]. Therefore, BACE1 is an attractive drug target for lowering brain levels of amyloid beta and, consequently, for the treatment or prevention of AD.

Many BACE1-inhibitors studied to date are peptidomimetics and incorporate the hydroxyethylamine (HEA) moiety [6], which is known to well mimic the transition state of aspartyl proteases (like BACE1) substrates.
These compounds have in general high molecular weights and suffer from poor blood brain barrier permeability [8]. So, in recent years, more efforts have been dedicated to the development of non-peptidomimetic BACE1-inhibitors, with the aim of obtaining smaller active molecules and, therefore, more drug-like agents for the treatment of AD [9-11]. Among those, only a few compounds have entered in advanced clinical phases [11,12].

We previously reported a series of α-naphthylaminoalcohol derivatives of unsubstituted carbazole showing a BACE1 inhibitory activity in the low μM range [13]. Furthermore, we recently reported that N-carboxamido-derivatives are still active against this enzyme (Fig. 1) [14].

![Fig. 1 General structure of already reported N-[3-(9H-carbazol-9-yl)-2-hydroxypropyl]-arylcaboxamides and novel variously substituted N-[3-(9H-carbazol-9-yl)-2-hydroxypropyl]-arylsulfonamides (1-24, see also Table 1).](https://example.com/fig1.png)

In this work, we describe the synthesis and the evaluation of BACE1 inhibitory activity of a further series of derivatives, in which the unsubstituted carbazole-methyl carbinol portion was kept constant and the N-atom was sulfonylated with different groups (1–24, Fig. 1). A focused screening in the literature of this type of compounds was primarily carried out, and six derivatives were found commercially available (1–3, 5, 11 and 22, see Table 1).

These sulfonamido-derivatives were synthesized as shown in Scheme 1. Commercially available carbazole 25 was alkylated with epichlorohydrin (2.5 eq., added dropwise at 0 °C) in the presence of KOH (1.2 eq.) in DMF, affording epoxide 26. Reaction of this intermediate with the appropriate amine (2.0 eq.) in EtOH afforded aminoalcohols 27–38. These derivatives were then treated with variously substituted aryl-sulfonylchlorides (1.1 eq.) in the presence of PS-DIEA (1.2 eq.) and DMAP (catalytic amount) in CH₂Cl₂, obtaining final compounds 1–24.

![Scheme 1 Reagents and conditions: a) epichlorohydrin, KOH, DMF, 0 °C, 5 h, 50%; b) R¹-NH₂, EtOH, 65 °C, overnight, 40–79%; c) ArSO₂Cl, PS-DIEA, DMAP, CH₂Cl₂, r. t., overnight, 13–86%.](https://example.com/scheme1.png)
The inhibitory activity of the newly synthesized compounds towards BACE1 was determined by a previously reported fluorescence-based assay [15] and the results are shown in Table 1.

The introduction of a simple phenyl substituent on the sulfonamide nitrogen (R₁), together with the presence of an unsubstituted or substituted aryl sulfonamide (compounds 1–4, IC₅₀ ranging from 2.4 to 3.0 μM), leads to a BACE1-inhibitory activity comparable to that of some of the most active compounds included in the previous series of N-carboxamides (IC₅₀ = 2.5 μM) [14]. When the sulfonamide aromatic ring is unsubstituted (R² = H) and the phenyl on the sulfonamide nitrogen is substituted in position 4, the activity worsens slightly (compounds 7 and 8), significantly (compound 5) or is completely lost (compounds 6 and 9). The presence of

### Table 1 Structures and BACE1 inhibitory activities of novel variously substituted arylsulfonamides 1–24.

<table>
<thead>
<tr>
<th>Compd</th>
<th>R¹</th>
<th>R²</th>
<th>IC₅₀ (μM)</th>
<th>logBBpred⁹</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Ph</td>
<td>H</td>
<td>3.0</td>
<td>-0.58</td>
</tr>
<tr>
<td>2</td>
<td>Ph</td>
<td>CH₃</td>
<td>2.6</td>
<td>-0.57</td>
</tr>
<tr>
<td>3</td>
<td>Ph</td>
<td>Cl</td>
<td>2.9</td>
<td>-0.61</td>
</tr>
<tr>
<td>4</td>
<td>Ph</td>
<td>OCH₃</td>
<td>2.4</td>
<td>-0.64</td>
</tr>
<tr>
<td>5</td>
<td>4-CH₃-Ph</td>
<td>H</td>
<td>7.1</td>
<td>-0.57</td>
</tr>
<tr>
<td>6</td>
<td>4-Cl-Ph</td>
<td>H</td>
<td>&gt; 10</td>
<td>-0.61</td>
</tr>
<tr>
<td>7</td>
<td>4-CF₃-Ph</td>
<td>H</td>
<td>3.6</td>
<td>-0.60</td>
</tr>
<tr>
<td>8</td>
<td>4-NO₂-Ph</td>
<td>H</td>
<td>3.8</td>
<td>-0.79</td>
</tr>
<tr>
<td>9</td>
<td>4-F-Ph</td>
<td>H</td>
<td>&gt; 10</td>
<td>-0.63</td>
</tr>
<tr>
<td>10</td>
<td>Bn</td>
<td>H</td>
<td>1.9</td>
<td>-0.61</td>
</tr>
<tr>
<td>11</td>
<td>Bn</td>
<td>CH₃</td>
<td>2.7</td>
<td>-0.61</td>
</tr>
<tr>
<td>12</td>
<td>Bn</td>
<td>Cl</td>
<td>1.7</td>
<td>-0.65</td>
</tr>
<tr>
<td>13</td>
<td>Bn</td>
<td>OCH₃</td>
<td>1.6</td>
<td>-0.67</td>
</tr>
<tr>
<td>14</td>
<td>4-CH₃-Bn</td>
<td>H</td>
<td>&gt; 10</td>
<td>-0.61</td>
</tr>
<tr>
<td>15</td>
<td>4-Cl-Bn</td>
<td>H</td>
<td>2.7</td>
<td>-0.65</td>
</tr>
<tr>
<td>16</td>
<td>4-OCH₃-Bn</td>
<td>H</td>
<td>5.7</td>
<td>-0.68</td>
</tr>
<tr>
<td>17</td>
<td>Phenethyl</td>
<td>H</td>
<td>2.8</td>
<td>-0.60</td>
</tr>
<tr>
<td>18</td>
<td>Phenethyl</td>
<td>CH₃</td>
<td>2.5</td>
<td>-0.59</td>
</tr>
<tr>
<td>19</td>
<td>Phenethyl</td>
<td>Cl</td>
<td>3.8</td>
<td>-0.63</td>
</tr>
<tr>
<td>20</td>
<td>Phenethyl</td>
<td>OCH₃</td>
<td>3.0</td>
<td>-0.66</td>
</tr>
<tr>
<td>21</td>
<td>Cyclohexyl</td>
<td>H</td>
<td>4.1</td>
<td>-0.61</td>
</tr>
<tr>
<td>22</td>
<td>Cyclohexyl</td>
<td>CH₃</td>
<td>3.9</td>
<td>-0.60</td>
</tr>
<tr>
<td>23</td>
<td>Cyclohexyl</td>
<td>Cl</td>
<td>6.5</td>
<td>-0.65</td>
</tr>
<tr>
<td>24</td>
<td>Cyclohexyl</td>
<td>OCH₃</td>
<td>3.2</td>
<td>-0.67</td>
</tr>
</tbody>
</table>

⁹ IC₅₀ measurements were performed as reported in Ref. 15. Data represent mean values for at least three separate experiments. Standard errors are not shown for the sake of clarity and were never higher than 15% of the means.

⁹ Predicted blood-brain barrier permeation (logBBpred = log[Brain]/[Blood]) [28].
a benzyl substituent on the sulfonamide nitrogen causes an appreciable increase in the inhibitory activity, except in one case (compound 11); compounds 10, 12 and 13 proved to be the most potent inhibitors of this series, with IC_{50} values ranging from 1.6 to 1.9 μM. The para-substitution of the benzyl group (leaving the aryl sulfonamide unsubstituted) generally causes a decrement in the activity (compounds 14–16). If a phenethyl group is introduced on the sulfonamide nitrogen, regardless the presence of substituents in position 4 of the aryl sulfonamide moiety (compounds 17–20), the activity is preserved or slightly decreased respect to the most active compounds of the previous series of N-carboxamides (IC_{50} = 2.5 μM) [14]. The activity generally decreases when R^1 is a completely aliphatic group, such as a cyclohexyl (compounds 21–24). Regarding to the substituent on the sulfonamide aromatic ring, when R^2 = H and R^1 = phenyl, 4-substituted phenyl, 4-substituted benzyl, phenethyl or cyclohexyl, the BACE1 inhibitory activity worsens slightly (compounds 1, 7–8, 15–17 and 21) or is lost (compounds 6, 9 and 14). When R^2 = Me, the activity is preserved (compounds 2, 11 and 18) or slightly worsened (compounds 22). When the phenyl ring of the sulfonamide is substituted with an alogen atom (Cl in this case) or a methoxy group, the activity is almost preserved (compounds 3 and 4), improved (compounds 12 and 13) or worsened (compounds 19, 23, 20 and 24). In general, the BACE1 inhibitory activity of this kind of molecules is mostly influenced by the substituent on the sulfonamide nitrogen (R^1) and the N-benzyl-substituted compounds are the most active (except 11) regardless the substitution on the 4-position of the sulfonamide phenyl ring (R^2).

A molecular docking study of the sulfonamide derivatives in BACE1 active site was conducted to give an interpretation of the structure-activity relationship at a molecular level. First, BACE1 holo crystal structures were retrieved from the Protein Data Bank, resulting in 302 ligand-BACE1 complexes. The preliminary operations of our work were devoted to the identification of the best protein structure and docking protocol to use in the subsequent docking calculations.

According to a previously validated protocol [16], crystal structures were filtered according to chemical similarity of their co-crystallized ligand to compound 13, chosen as representative of the series of inhibitors herein described because of its highest potency (IC_{50} = 1.6 μM). MACCS Tanimoto similarity was computed exploiting RDKit [17] functionalities, and the 6 structures characterized by highest similarity values were selected (PDB ID: 2WF2 [18], 2WF3 [18], 2WF4 [19], 2VNN [20], 2VKM [21], 4FCO, with the addition of structure 1W51 [22] used in a previous study [14]). The selected structures were subjected to a docking benchmark study to evaluate the performance of different docking protocols and scoring functions in the self-docking exercise. The co-crystallized ligands were prepared by adding hydrogens using the Protonate3D tool of MOE [23], while the proteins were prepared by exploiting the Protein-preparation tool and the Protonate3D of the same software suite.
DockBench tool [24] was employed to automatically perform the docking benchmark, and, after the analysis of the benchmark results, the protein 2WF4 with the Gold-goldscore [25] protocol were chosen because of the high performance in DockBench Protocol Score.

The structures of the sulfonamide derivatives (stereoisomer \( R \) and \( S \) of each compound) were constructed by using the MOE builder function, the starting conformation was initially generated exploiting Corina [26] and then minimized using PM3 theory. The docking calculations were performed for each compound, limiting the conformational search within a 20 Å radius sphere centered on the center of mass of the co-crystallized ligand in the corresponding complex.

Fig. 2 Schematic representation of the principal interactions resulting from the docking study of compound 13.

Both \( R \) and \( S \) stereoisomers find a good accommodation within the active site of BACE1, with the (2-hydroxypropyl)sulfonamide portion protected behind the flap region and the benzylic and carbazolic moieties pointing toward two hydrophobic clefts positioned on left and right. The interaction established by compound 13 are depicted in Fig. 2 using the ligand interaction diagram as implemented in the Schrodinger suite [27]. In detail, the predicted binding mode of compound 13 is reported in Fig. 3, panel A. The benzylic moiety attached to the sulfonamide nitrogen is inserted into a hydrophobic pocket defined by Leu91, Ile179, Trp176, Ile171, and Phe169. Also, the carbazole portion leans on a hydrophobic portion of the protein, characterized by Tyr259, Ile287, and Val393. The sulfonamide function acts as a hydrogen bond acceptor with Gln134 positioned in the flap region. The hydroxyl groups of both \( R \) and \( S \) stereoisomers are involved in a hydrogen bond with one of the two catalytic aspartates: in particular, the \( R \)-enantiomer interacts with
Asp93, while the Senantiomer with Asp289. Moreover, both enantiomers are stabilized by a further hydrogen bond between the methoxy group in R² and Thr293. The pose of compound 13 (Fig. 3, panel B) is similar to that displayed by the crystallographic binding mode of the co-crystallized ligand (PDB ID: 2WF4; ligand ID: ZY4) within the protein conformation used in our docking calculations. The (2-hydroxypropyl)sulfonamide of compound 13 resembles the gem-diol group of the (2,2-dihydroxypropyl)amide moiety of the crystallographic compound: here the amide carbonyl group makes a hydrogen bond with Gln134 and the two hydroxyls are engaged as donors in two hydrogen bonds with Asp93.

![Fig. 3](image)

**Fig. 3** (A) Docking results of compound 13 in BACE1. Both (R) and (S)-stereoisomer are respectively shown in cyan and pink. (B) The obtained docking result is compared to the crystallographic binding mode of ligand ZY4 within the BACE1 conformation selected for docking studies (PDB ID: 2WF4).

This makes plausible the binding of both the stereoisomers of compound 13 to the BACE1 binding site. A binding mode similar to that of compound 13 was obtained by docking simulations for almost all the other sulfonamide derivatives, as reported in Video-S1 (Supplementary Material).
All the BACE1 inhibitors are meant to have central nervous system (CNS) activity, so they are expected to cross the blood-brain barrier (BBB). Thus, the logBB was computed with the Stardrop software [28] for all compounds, to estimate their capability to distribute from blood to the CNS. The logBB predicted values, being higher than −0.8 in all cases, fall under the -1 limit for passing the blood-brain barrier, so they seem to satisfy the CNS permeability expectation. Compound 8, which is characterized by the presence of a nitrophenyl substituent, shows the highest value of this series (logBB = −0.79), due to the relatively high polarity of the NO₂ group, thus making its putative BBB permeation less promising than those of the other analogues.

In conclusion, we have synthesized a series of BACE1 inhibitors possessing a N-[3-(9H-carbazol-9-yl)-2-hydroxypropyl]-arylsulfonamido structure. Among the 24 derivatives, 21 active analogues were found, with three highly active compounds (IC₅₀ values ranging from 1.6 to 1.9 μM). The docking study showed that both enantiomers of the most active compound of this series (13) find a good accommodation within the active site of BACE1; a similar binding mode was obtained by docking simulations of almost all the other sulfonamide derivatives, as reported in Video-S1 (Supplementary Material). Moreover, the predicted logBB values of all compounds (ranging from −0.57 to −0.79) indicate satisfactory BBB permeabilities.
References


26. Molecular Networks GmbH CORINA; Germany


Synthesis, structure-activity relationships and biological evaluation of 7-phenyl-pyrroloquinolinone 3-amide derivatives as potent antimitotic agents

Davide Carta, Roberta Bortolozzi, Mattia Sturlese, Veronica Salmaso, Ernest Hamel, Giuseppe Basso, Laura Calderan, Luigi Quintieri, Stefano Moro, Giampietro Viola, Maria Grazia Ferlin


Abstract

A small library of 7-pyrrolo[3,2-f]quinolinones was obtained by introducing benzoyl, sulfonyl and carbamoyl side chains at the 3-N position, and their cytotoxicity against a panel of leukemic and solid tumor cell lines was evaluated. Most of them showed high antiproliferative activity with GI50s ranging from micro-to sub-nanomolar values, and these values correlated well with the inhibitory activities of the compounds against tubulin polymerization. Based on a recently proposed colchicine bind site inhibitors (CBSIs) pharmacophore, the interactions of the novel 7-PPyQs at the colchicine domain were rationalized. The most active compounds (4a and 4b) did not induce significant cell death in normal human lymphocytes, suggesting that the compounds may be selective against cancer cells. In particular, 4a was a potent inducer of apoptosis in both the HeLa and Jurkat cell lines. On the other hand, the sulfonyl derivative 4b exhibited a lower potency in comparison with 4a. With both compounds, induction of apoptosis was associated with dissipation of the mitochondrial transmembrane potential and production of reactive oxygen species, suggesting that cells treated with the compounds followed the intrinsic pathway of apoptosis.

1. Introduction

Drugs interfering with microtubules (MTs) represent a class of compounds of great interest in the area of anticancer therapy. MTs are an essential component of the cellular cytoskeleton, as they regulate and participate in a variety of cellular functions that include motility, morphology, intracellular transport, signal transduction, and cell division [1]. MTs are composed of α/β tubulin heterodimers that have polymerized into cylindrical structures, and, therefore, both natural and synthetic agents able to interfere with tubulin polymerization or depolymerization, thereby altering MT dynamics, continue to attract considerable attention in the field of chemotherapeutic research [2]. Not only are there clinically used natural and semisynthetic antimitotics (such as, paclitaxel, other taxanes, eribulin, ixabepilone and vinca alkaloids), but there is also a large number of structurally dissimilar small molecules with high affinity for the colchicine site on tubulin. These compounds are able to inhibit the proliferation of a wide variety of human cancer cells [3].
Moreover, these agents can also affect the tumor endothelial vasculature required for the growth of tumor mass. These types of tubulin inhibitors might provide new therapeutic approaches to treat cancers and overcome limitations of existing tubulin interactive drugs [4]. In the last decade, we have been developing phenylpyrroloquinolinone (PPyQ) derivatives that show interesting in vitro and in vivo antitumor activity. Both 2-PPyQs and 7-PPyQs act as tubulin polymerization inhibitors by binding at the colchicine site in β-tubulin [5, 6]. Although less cytotoxic, the 2-PPyQ compounds were also found to exhibit interesting in vitro and in vivo antiangiogenic properties [7]. The more cytotoxic 7-PPyQ derivatives showed very remarkable in vitro biological properties and good antitumor activity in vivo. In particular, some 7-PPyQs, characterized by alkyl substitutions at the pyrrole nitrogen, showed increased cytotoxicity with nanomolar GI50 values, and these compounds overcame the resistance observed with the clinically used agents vincristine and taxol [8, 9]. In the latter series, the 3N-cyclopropyl methyl 7-PPyQ derivative MG 2477 (Fig. 1, 10), was taken as lead compound due to its very strong cytotoxicity (nanomolar range GI50) and its potent interaction with tubulin. Its activities as an inhibitor of tubulin polymerization and of colchicine binding to tubulin were similar to those of the reference compound combretastatin A-4: 0.90 μM assembly IC50 and 83% inhibition of colchicine binding for compound 10 versus values of 1.1 and 99%, respectively, for CA-4 [10, 11]. Compound 10 was also demonstrated to induce autophagy in the A549 cell line [12]. Very recently, in an effort to produce additional highly active compounds, numerous related analogues were designed, synthesized and studied, resulting in the discovery of a potent 3N-acyl derivative MG 2603 (Fig. 1, 11) showing low nanomolar GI50 values. Compound 11, too, showed an anti-tubulin mechanism profile similar to that of 10, but it was also able to inhibit a number of kinases involved in tumor progression. Moreover, 11 showed reduced toxicity in non tumor cell lines and synergized with conventional chemotherapeutic agents in inhibiting leukemia cell proliferation [13].

Fig. 1 Structure-activity relationships of 7-phenyl-pyrrolo[3,2-f]quinolinones.

Published: Eur J Med Chem. 2017 Feb 15;127:643–660
Driven by the SARs we have collected during the development of the PPyQs [14] (Fig. 1) and remembering the recent results with the 3N-acyl 7-PPyQ derivative, the aim of the present work was the design, synthesis and evaluation of novel analogues substituted at the pyrrole N with acyl, sulfonyl and carbamoyl side chains. In designing the novel derivatives, we preserved the structural elements crucial for the best antiproliferative activity, such as the [3,2-f] geometry of the pyrroloquinoline core, the un-substituted phenyl ring at the 7 and the carbonyl group at the 9 position, without any other substitutions except for the 3 position (Fig. 1). Biological investigations included cellular cytotoxicity, tubulin inhibition assays and an apoptosis assay, together with docking simulations in the colchicine site of β-tubulin. This allowed us to obtain more knowledge on the key substitutions at the pyrrole N for effective interactions at the colchicine site.

2. Results and discussion

2.1 Chemistry

Scheme 1 shows the route to 7-PPyQs bearing a side chain bound to the pyrrole N via a carbonyl group according to the previously reported general synthesis to 7-PPyQs [7]. The starting commercial 5-nitroindole was reacted with various acyl and sulfonyl chlorides in order to obtain directly the 3N-substituted indole derivatives 1a-d. Compound 1e was prepared by means of a one-pot procedure consisting first of activation of 5-nitroindole with p-nitrophenylchloroformate to give the reactive 3-p-nitrophenylcarbamate intermediate and then reaction with cyclopropylamine (19% yield). The next reduction step of 5-nitro- to 5-aminoindole intermediates was accomplished by a chemical procedure with SnCl₂·2H₂O, 37% HCl in methanol at reflux to give indole derivatives 2a-d. In the case of 1e, the reduction did not produce any corresponding amino compound. While, by a catalytic procedure with H₂ and C/Pd 10% in EtOAc/EtOH at atmospheric pressure, indoline compounds 5a, b, d and e were obtained. Note that by the above chemical method aminoindole derivative 2a was obtained in poor yields due to the formation of a mixture of various chloro-derivatives (not shown). Therefore the synthesis to 4a did not proceed beyond the c step. Aminoindole derivatives 2b-d were then condensed with ethyl benzoyl acetate to provide the enamine intermediates 3b-d to be then thermally cyclized into the final 7-PPyQs 4b-d. In the same way, indolines 5a, b, d and e gave enamine derivatives 6a, b, d and e with benzoyl acetate. However, when these were submitted to thermal cyclization in diphenyl ether, 6a, b and d gave a mixture of isomeric compounds angular 7a, b, d and linear 8a, b, d, whereas derivative 6e did not react. It is worth emphasizing that, by the catalytic procedure at the reaction conditions used here, we never observed the formation either of aminoisindolines or the subsequent cyclization to linear tricyclic compounds.
Scheme 1  a) Benzyol chloride, methansulfonyl chloride, p-methylbenzensulfonyl chloride, p-trifluoromethylbenzensulfonylchloride, NaH (60%), anhydrous DMF, rt, 2 h, 92%; p-nitrophenylchloroformate and cyclopropylamine, THF, 3 h, 87%; b) SnCl$_2$·2H$_2$O, HCl 37%, methanol, reflux, 36 h, 53%; c) ethylbenzoyl acetate, absolute ethanol, cat CH$_3$COOH, drierite, reflux, 36 h, 60%; d) diphenyl ether, reflux, 15 min, 79%; e) H$_2$, Pd/C 10%, EtOAc, atmospheric pressure, 50 °C, 24 h, 95%.

Scheme 2 shows an alternative method to obtain the final compounds 4a and 4e, which could not be prepared by the route shown in Scheme 1. The previously described 7-PPyQ 9 [8], available in our laboratory, was submitted to an acylation reaction with benzoyl chloride. After a laborious purification procedure, 7-PPyQ 4a was obtained in a 37% yield. Compound 4e was obtained by the same one pot procedure described above, consisting of the reaction of 7-PPyQ 9 to give the intermediate 3-p-nitrophencarbamate, which was not isolated, followed by reaction with cyclopropylamine (30% yield).

Scheme 2  a) Benzyol chloride, NaH (60%), anhydrous DMF, rt, 3 h, 37%; c) p-nitrophenylchloro formate; NaH (60%), cyclopropyl amine, THF, rt, 3 h, 30%.
2.2. Biological evaluation

2.2.1. In vitro antiproliferative activities and SAR analysis

On the basis of previous biological activity data on 7-PPyQs and docking simulations of 11 into the colchicine site of tubulin [13], the new compounds were designed to obtain additional SAR information by modifying the nature and size of substituents at the 3 position of the 7-PPyQ tricycle. Evaluation of antiproliferative activities of 4a-e, 7a,b, 7d, and 8a,b was performed with the 3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide (MTT) assay against a panel of 11 human tumor cell lines (CCRF-CEM, HL-60, RS4; 11, Jurkat, SEM, MV4; 11, THP-1, HeLa, A549, HT-29, MCF-7). GI\textsubscript{50} values, the concentrations that inhibit cell growth by 50%, are presented in Table 1. Most of the novel 7-PPyQs possessed antiproliferative activity, inhibiting cell growth with nanomolar to micromolar GI\textsubscript{50} values, except for the linear 1,2-dihydro 8b (GI\textsubscript{50} > 10000 nM) having a methanesulfonyl group at the 3 position. In contrast, its angular isomer 7b showed GI\textsubscript{50} values in a high nanomolar range, demonstrating that for partially hydrogenated pyrroloquinoliones the [3,2-f] geometry is also preferred for cytotoxic activity as it was for fully aromatic compounds. Comparable behavior was also observed for the benzamidic derivatives 7a and 8a, although not as dramatic: the linear 8a showed GI\textsubscript{50} values in the micromolar, while the angular compound 7b had IC\textsubscript{50} values in the sub-micromolar range. Moreover, from the data presented in Table 1, it is evident that the angular, fully aromatic 7-PPyQs 4a,b,d were more cytotoxic than the corresponding hydrogenated analogues 7a,b,d. This was most remarkable for the pair 4a and 7a, with the former having GI\textsubscript{50} values in the 0.1–10 nM range and the latter GI\textsubscript{50} values in the 250–2650 nM range. Of particular note were the low nanomolar and sub-nanomolar GI\textsubscript{50} values obtained with the series 4a-4e, in both the leukemic and solid tumor cell lines. Overall, the most active of the new compounds was the benzamidic derivative 4a, with subnanomolar concentrations in five of the eleven cell lines, with slightly lower GI\textsubscript{50} values in the solid tumor lines (GI\textsubscript{50}s 0.2, 0.1 and 0.2 nM in the HeLa, HT-29 and MCF-7 cells, respectively). Previously, we had observed that 7-PPyQs were more cytotoxic against leukemic cells. Thus, the preferential activity of 4a against solid tumor cell lines is worth emphasizing. The same relative activity against the solid tumor cell lines was also observed for sulfamidic derivatives 4b-d, but not for the ureidic derivative 4e. We also note that there did not appear to be a significant steric hindrance factor among the compounds evaluated here, with similar antiproliferative activities observed among the series 4a-4e and the previously evaluated compound 11.

We conclude that the 7-PPyQ derivatives 4a-e, chemically modified at 3 position with substitutions such as carbonyl, sulfonyl and carbamoyl groups, maintained strong cytotoxicity against tumor cell lines. These substitutions were made to further explore the SARs, and they have confirmed what was observed previously with carbonyl compound 11. Substitutions with oxygenated groups, although of diverse nature and volume,
onfer very high antiproliferative activity on 7-PPyQs. Due to their broad spectrum of potent activity, 4a and 4b were selected for further biological investigations on mechanism of action.

### Table 1 In vitro cell growth inhibitory effects of compounds 4a-e, 7a,b,d, 8a-b, and 11

<table>
<thead>
<tr>
<th>cmp</th>
<th>CCRF-CEM</th>
<th>HL-60</th>
<th>RS 4; 11</th>
<th>Jurkat</th>
<th>SEM</th>
<th>MV 4; 11</th>
<th>THP-1</th>
<th>HeLa</th>
<th>A549</th>
<th>HT-29</th>
<th>MCF-7</th>
</tr>
</thead>
<tbody>
<tr>
<td>4a</td>
<td>12 ± 0.2</td>
<td>0.3 ± 0.1</td>
<td>16 ± 0.1</td>
<td>0.9 ± 0.1</td>
<td>2 ± 0.9</td>
<td>5 ± 0.4</td>
<td>2 ± 0.6</td>
<td>0 ± 0.04</td>
<td>10 ± 6</td>
<td>0.2 ± 0.08</td>
<td>0.1 ± 0.1</td>
</tr>
<tr>
<td>4b</td>
<td>42 ± 7</td>
<td>215 ± 7</td>
<td>18 ± 9</td>
<td>34 ± 4</td>
<td>23 ± 6</td>
<td>49 ± 16</td>
<td>92 ± 31</td>
<td>36 ± 9</td>
<td>7 ± 2</td>
<td>152 ± 95</td>
<td></td>
</tr>
<tr>
<td>4c</td>
<td>32 ± 1</td>
<td>335 ± 1</td>
<td>48 ± 1</td>
<td>55 ± 4</td>
<td>33 ± 6</td>
<td>65 ± 31</td>
<td>222 ± 52</td>
<td>26 ± 5</td>
<td>5 ± 2</td>
<td>195 ± 95</td>
<td></td>
</tr>
<tr>
<td>4d</td>
<td>32 ± 1</td>
<td>39 ± 1</td>
<td>147 ± 1</td>
<td>51 ± 3</td>
<td>43 ± 2</td>
<td>7 ± 11</td>
<td>86 ± 42</td>
<td>42 ± 22</td>
<td>63 ± 23</td>
<td>5 ± 3</td>
<td></td>
</tr>
<tr>
<td>4e</td>
<td>92 ± 14</td>
<td>155 ± 3</td>
<td>27 ± 14</td>
<td>115 ± 11</td>
<td>85 ± 6</td>
<td>66 ± 21</td>
<td>22 ± 17</td>
<td>171 ± 25</td>
<td>55 ± 22</td>
<td>157 ± 112</td>
<td></td>
</tr>
<tr>
<td>7a</td>
<td>1045 ± 212</td>
<td>2640 ± 309</td>
<td>256 ± 133</td>
<td>1345 ± 302</td>
<td>918 ± 211</td>
<td>1543 ± 223</td>
<td>2166 ± 614</td>
<td>331 ± 55</td>
<td>918 ± 268</td>
<td>1586 ± 145</td>
<td>1112 ± 178</td>
</tr>
<tr>
<td>7b</td>
<td>235 ± 56</td>
<td>291 ± 13</td>
<td>35 ± 4</td>
<td>72 ± 26</td>
<td>31 ± 1</td>
<td>36 ± 0.2</td>
<td>71 ± 65</td>
<td>71 ± 54</td>
<td>451 ± 59</td>
<td>623 ± 96</td>
<td></td>
</tr>
<tr>
<td>7d</td>
<td>435 ± 0.85</td>
<td>321 ± 35</td>
<td>211 ± 32</td>
<td>336 ± 21</td>
<td>356 ± 25</td>
<td>846 ± 63</td>
<td>373 ± 25</td>
<td>541 ± 29</td>
<td>255 ± 62</td>
<td>475 ± 47</td>
<td></td>
</tr>
<tr>
<td>8a</td>
<td>818 ± 154</td>
<td>3740 ± 233</td>
<td>132 ± 51</td>
<td>231 ± 407</td>
<td>1656 ± 185</td>
<td>1978 ± 326</td>
<td>2323 ± 945</td>
<td>1447 ± 416</td>
<td>3562 ± 624</td>
<td>5436 ± 852</td>
<td>6323 ± 684</td>
</tr>
<tr>
<td>8b</td>
<td>&gt;10000</td>
<td>&gt;10000</td>
<td>&gt;10000</td>
<td>&gt;10000</td>
<td>&gt;10000</td>
<td>&gt;10000</td>
<td>&gt;10000</td>
<td>&gt;10000</td>
<td>&gt;10000</td>
<td>&gt;10000</td>
<td>&gt;10000</td>
</tr>
<tr>
<td>11b</td>
<td>17 ± 0.6</td>
<td>4 ± 0.1</td>
<td>0.3 ± 0.3</td>
<td>19 ± 0.4</td>
<td>74 ± 0.4</td>
<td>1.6 ± 0.8</td>
<td>9 ± 25</td>
<td>1 ± 6</td>
<td>5 ± 0.5</td>
<td>1 ± 0.5</td>
<td></td>
</tr>
</tbody>
</table>

*a* IC₅₀ = compound concentration required to inhibit tumor cell proliferation by 50%. Data are expressed as the mean ± SE from the dose-response curves of at least three independent experiments.

*b* Data taken from ref. [13].

#### 2.2.2. Evaluation of cytotoxicity in human non-cancer cells

To obtain a preliminary indication of the cytotoxic potential of these derivatives in normal human cells, the two most active compounds (4a and 4b) were evaluated in vitro against peripheral blood lymphocytes (PBL) from healthy donors (Table 2). Compound 4a showed a Gₛ₀ of 28 μM in quiescent lymphocytes, while in the presence of the mitogenic stimulus phytohemagglutinin (PHA), the Gₛ₀ decreased to about 15 μM. Notably, this value was almost 1000–2000 times higher than that observed against the lymphoblastic cell lines CCRF-CEM and Jurkat. These results indicate that 4a has a significant effect in rapidly proliferating cells but not in quiescent cells, as previously observed for other antimitotic derivatives developed by our group [13]. Compound 4b was completely inactive in both quiescent and proliferating lymphocytes.
Table 2 Cytotoxicity of 4a-b for human peripheral blood lymphocytes (PBL).

<table>
<thead>
<tr>
<th></th>
<th>IC50 (µM)(^a)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4a</td>
<td>28.0 ± 2.3</td>
</tr>
<tr>
<td>4b</td>
<td>&gt; 100</td>
</tr>
</tbody>
</table>

Values are the mean ± SEM from three separate experiments.

\(^a\) Compound concentration required to reduce cell growth inhibition by 50%.

\(^b\) PBL not stimulated with PHA.

\(^c\) PBL stimulated with PHA.

2.2.3. Inhibition of tubulin polymerization and colchicine binding

To evaluate the tubulin interaction properties of compounds 4a-e, we investigated their effects on inhibition of tubulin polymerization and the binding of \([^3H]\)colchicine to tubulin (Table 3) [15, 16]. For comparison, CA-4 and 3c were examined in contemporaneous experiments as references compounds. Among the test compounds, 4a strongly inhibited tubulin assembly assay with an IC\(_{50}\) of 0.89 µM, a value that was lower than that obtained for the reference compound CA-4 (IC\(_{50}\) = 1.2 µM). Compounds 4b and 4c showed an IC\(_{50}\) similar to that of CA-4 while 4d and 4e were less effective than the reference compound (IC\(_{50}\) = 2.2–2.4 µM). These results correlate well with the growth inhibitory effects exhibited by the test compounds, indicating that their antiproliferative activity derives from an interaction with tubulin.

Table 3 Inhibition of tubulin polymerization and colchicine binding by compounds 4a-e and CA-4.

<table>
<thead>
<tr>
<th>Compound</th>
<th>Tubulin assembly(^a)</th>
<th>Colchicine binding(^b)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IC(_{50})±S.D. (µM)</td>
<td>% inhibition ±S.D.</td>
</tr>
<tr>
<td>4a</td>
<td>0.89 ± 0.04</td>
<td>70 ± 2</td>
</tr>
<tr>
<td>4b</td>
<td>1.2 ± 0.01</td>
<td>42 ± 4</td>
</tr>
<tr>
<td>4c</td>
<td>1.1 ± 0.04</td>
<td>37 ± 5</td>
</tr>
<tr>
<td>4d</td>
<td>2.4 ± 0.2</td>
<td>29 ± 3</td>
</tr>
<tr>
<td>4e</td>
<td>2.2 ± 0.3</td>
<td>18 ± 4</td>
</tr>
<tr>
<td>CA-4</td>
<td>1.2 ± 0.1</td>
<td>98 ± 0.7</td>
</tr>
</tbody>
</table>

\(^a\) Inhibition of tubulin polymerization. Tubulin was at 10 µM.

\(^b\) Inhibition of \([^3H]\)colchicine binding. Tubulin and colchicine were at 1 and 5 µM concentrations, respectively.

In the colchicine studies, compound 4a was the most active inhibitor of the binding of \([^3H]\)colchicine to its domain on tubulin, with 70% inhibition occurring with this derivative at 5 µM. Nevertheless, 4a was less potent than CA-4 in this assay. In these experiments CA-4 inhibited colchicine binding by 98% at 5 µM. The
other investigated compounds (4b-e) showed weaker inhibitory activity, with less than 50% inhibition of colchicine binding to tubulin.

2.2.4. Computational studies

Docking studies were carried out to investigate the binding mode of the novel inhibitors with the aim of interpreting experimental affinity data. A relevant number of experimentally derived complex structures of colchicine binding site inhibitors (CBSI) were recently deposited in the Protein Data Bank (PDB) [17]. Interestingly, the superposition of the different crystal structures reveals a significant variability in the sidechains of the residues belonging to the colchicine site depending on the chemical nature of the ligand, as shown in SI_Fig. 1 (see Supplementary Material). Moreover, the resolution of the crystal structures spanned a broad range (2.19–3.75 Å). As a consequence of this heterogeneity, we carried out a benchmark study, using the DockBench tool [18], to identify the most accurate docking model among 14 different ones and to determine which protein conformation was most appropriate to model our analogues. The benchmark study on the self-docking procedure was performed on 14 tubulin–CBSI complexes from the PDB as listed in table SI_Table 1. The benchmark results, summarized in SI_Fig. 2, revealed that several protocols showed a good ability to reproduce the experimental complex geometries for most of the experimental structures. Among them, GOLD software coupled to the PLP/goldscore/chemscore scoring function, gave the best results. In particular, GOLD protocols returned accurate predictions for the complete dataset. As a consequence of the overall good performance of the benchmark, we focused our attention on the identification of the most suitable crystallographic protein structure for the docking simulation of the PPyQ class of compounds. This step was crucial because of the variety of different sidechain orientations for certain residues in the colchicine site such as βGlu200, βCys239, βLeu248, and βLeu255, as shown in SI_Fig. 1. We have addressed this critical issue by comparing the shape similarity and the pharmacophoric determinants conservation between the ligands present in the complexes (summarized in SI_Table 1) and our representative compound, 4a. Plinabulin (PDB ID: 5C8Y) showed the highest shape similarity according Tversky coefficient (>0.7) and, more notably, the plinabulin key moieties for the tubulin interaction are nicely conserved as shown in Fig. 2, Panel A. Not surprisingly, all analogues showed a common binding mode similar to that of plinabulin, as shown by SI_Video 1. As depicted in Fig. 2 (Panel B), the diketopiperazinic core of plinabulin is mimicked by the pyrroloquinolinone core, maintaining the key hydrogen bond interaction with the backbone of βVal236 as anticipated by the shape-based superposition (plinabulin and 4a, Fig. 2, Panel A). In addition to the conserved hydrogen bond, the pyrroloquinolinone scaffold guarantees strong hydrophobic interactions with βLeu253, βAla314 and βLeu368. The phenyl ring in position 7 reproduced the same scheme of interaction to the benzylidene moiety of plinabulin through hydrophobic interactions with residues βPhe167, βTyr200, and βLeu250. The substituents at the N-pyrrole were placed in the pocket formed by
residues: βLys350, βThr351, βAla314, βAla352 and, for the more bulky substituents, also αThr179. Notably, this binding mode is compatible with a competitive mechanism of action at the colchicine site. The 1,2-dihydro derivatives showed minor differences in their orientation. The main difference is the reduction in the interaction strength with βVal236, in particular for the linear isomers (8a and 8b) (SI_Video 1).

Fig. 2 Panel A. Superposition of 4a (green sticks, grey surface) and plinabulin (grey sticks) derived from the shape similarity calculations. Panel B. The energetically most favorable pose of 4a (in green) obtained by molecular docking simulation using the protein conformation of the plinabulin complex (PDB ID: 5C8Y). The ribbon as well as the residue atoms of the colchicine binding site are colored according the subunit to which they belong: white for β-tubulin and magenta for α-tubulin. Hydrogen atoms are not shown. Panel C. Per-residue analysis of the protein-ligand interaction for compound 4a (green) and plinabulin (red). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

To evaluate if molecular docking was also able to explain the differences in the inhibition of colchicine binding by compounds 4a-e, we performed a per-residue analysis along the series, in which the contribution for each residue belonging to the binding site is computed for each synthesized compound (Fig. 2, Panel C). In particular, we measured the electrostatic interaction energy and score, taking into account hydrophobic interactions. In SI_Video 1 is reported the per-residue analysis for all the analogues in Table 1, including the reference compound 11. The resulting heatmap suggests a very similar pattern of interaction for all the compounds. In this context, the narrow differences in colchicine binding and tubulin assembly inhibition are difficult to rationalize. The higher inhibition in the colchicine binding of 4a among the 4a-4e derivatives could
be ascribed to the orientation of the benzamidic moiety that is directed between the βLys350 and αThr179 residues, while in the 4b-4e analogues the N substituent assumes a slightly different orientation. The differences in the cell growth inhibition data are difficult to rationalize only with tubulin docking, possibly because of the involvement of other targets [13]. A clearer interpretation can be made for compound 8b, since its inhibition of tumor cell growth was negligible, and this analogue showed poor scores in molecular docking primarily because of a poor interaction with βVal236, as shown in the per residue analysis (SI_Video 1).

2.2.5. Compounds 4a and 4b induce mitotic arrest of the cell cycle

To investigate whether compounds 4a and 4b affected cell cycle progression, we evaluated by flow cytometry the effect of different concentrations of compounds after a 24 h of treatment of HeLa and Jurkat cells. As shown in Fig. 3, compound 4a caused a significant G2/M arrest in a concentration-dependent manner in both cell lines, with a rise in G2/M cells occurring at a concentration of 50 nM, while at the highest concentration (100 nM) more than 50% of the cells were arrested in G2/M. In the HeLa cells, the G2/M block was accompanied by a significant reduction of both G1 and S phase cells, suggesting that cell proliferation is impaired. For compound 4b, we observed a similar behavior but less marked as compared with 4a, in good agreement with the respective IC_{50} values found in the tubulin polymerization assay.

![Graphs showing cell cycle distribution for compounds 4a and 4b](image-url)

**Fig. 3** Percentage of cells in each phase of the cell cycle in HeLa (Panel A) and Jurkat cells (Panel B) treated with compounds 4a or 4b at the indicated concentrations for 24 h. Cells were fixed and labeled with PI and analyzed by flow cytometry as described in the Experimental Section. Data are presented as mean of two independent experiments with similar results.
2.2.6. Compounds 4a and 4b induce apoptosis through the mitochondrial death pathway

To evaluate the mode of cell death induced by 4a and 4b, we performed the annexin-V/propidium iodide (PI) assay by flow cytometry. Staining with annexin-V and with PI allows discrimination between live cells (annexin-V-/PI-), early apoptotic cells (annexin-V+/PI-), late apoptotic cells (annexin-V+/PI+) and necrotic cells (annexin-V-/PI+). The experiments were carried out both in Hela and Jurkat cells. As shown in Fig. 4; Fig. 5, compounds 4a and 4b induce a significant time- and concentration-dependent increase of apoptotic cells in both cell lines. In particular, after 24 h we found a significant accumulation of early apoptotic (annexin-V+/PI-) cells starting from the lower concentrations and an increase of late apoptotic cells (annexin-V+/PI+) after 48 h treatments, indicating that the compounds trigger cells to a massive apoptotic cell death. In good agreement with the respective GI50 values (Table 1), compound 4a is the more potent inducer of apoptosis both in Hela and Jurkat cells at the 50 nM concentration.

Fig. 4 Flow cytometric analysis of apoptotic cells after treatment of HeLa cells with 4a or 4b at the indicated concentrations after incubation for 24 or 48 h. The cells were harvested and labeled with annexin-V-FITC and PI and analyzed by flow cytometry. Data are presented as mean ± SEM of three independent experiments.

Fig. 5 Flow cytometric analysis of apoptotic cells after treatment of Jurkat cells with 4a or 4b at the indicated concentrations after incubation for 24 or 48 h. The cells were harvested and labeled with annexin-V-FITC and PI and analyzed by flow cytometry. Data are presented as mean ± SEM of three independent experiments.
Loss of mitochondrial transmembrane potential ($\Delta \psi_{mt}$) and release of apoptogenic factor has been described as an early event in the apoptotic process [19, 20]. $\Delta \psi_{mt}$ was evaluated by flow cytometry using the fluorescence of the dye JC-1. In normal conditions (high $\Delta \psi_{mt}$), JC-1 displays a red fluorescence (590 nm), while mitochondrial depolarization is indicated by a shift to green fluorescence (525 nm).

In both Hela and Jurkat cells, treatment with 4a and 4b induced a marked increase in the percentage of cells with low $\Delta \psi_{mt}$ (Fig. 6), and this occurred in a time- and concentration-dependent fashion. Depolarization of mitochondrial potential leads to the induction of the intrinsic pathway of apoptosis and is associated with the appearance of annexin-V positivity in the treated cells, as shown above and indicating the cells are in an early apoptotic stage. In fact, the disruption of $\Delta \psi_{mt}$ and the intrinsic activation of apoptosis are characteristic of antimitotic drugs and have been observed with both microtubule stabilizing and destabilizing agents in different cell types. It is also well known that mitochondrial potential impairment and the resulting damage to mitochondrial function induce generation of reactive oxygen species (ROS) [21, 22]. Superoxide anion is produced by mitochondria due to a shift from the normal 4-electron reduction of $O_2$ to a 1-electron reduction when cytochrome c is released from mitochondria upon apoptosis [23, 24].

Using dichlorodihydrofluorescein diacetate ($H_2$-DCFDA), which is oxidized to the fluorescent compound dichlorofluorescein (DCF) upon ROS induction [23], we measured ROS production after treatment with compounds 4a and 4b. As shown in Fig. 7 (Panels B and D), the two compounds induced the production of large amounts of ROS in comparison with control cells. This was observed in both the Jurkat and HeLa cells, in good agreement with the dissipation of $\Delta \psi_{mt}$ described above.
Fig. 7 Assessment of ROS production after treatment of HeLa (Panel A) or Jurkat (Panel B) cells with the indicated compounds. Cells were treated with the indicated concentration of compound for 24 or 48 h and then stained with H$_2$-DCFDA for the evaluation of ROS levels. Cells were then analyzed by flow cytometry as described in the Experimental Section. Data are presented as mean ± SEM of three independent experiments.

2.2.7. Metabolic stability of 4a in human liver microsomes

Liver microsomal oxidation and hydrolysis represent major routes of drug metabolism in mammals, including humans [25]. In vitro studies were therefore carried out to get preliminary information on the stability of compound 4a to oxidative and hydrolytic metabolism by human liver microsomes. As shown in Fig. 8 (panel A), compound 4a (10 μM) was relatively stable in human liver microsomes (1 mg/mL) with more than 60% compound remaining after 60 min incubation at 37 °C. Interestingly, compound 4a disappearance was not influenced by the presence of NADPH (Fig. 8, panel A), a cofactor for both cytochrome P450- and flavin monooxygenase-mediated oxidations [25], and was accompanied by formation of a fluorescent metabolite whose retention time corresponded exactly to that of authentic compound 9 (panel B). Collectively, these findings indicate that compound 4a is partially susceptible to microsomal enzyme hydrolysis and that this catabolism produce compound 9 which retain a significant antiproliferative activity as previously demonstrated [8].
Fig. 8 Assessment of metabolic stability of 4a in human liver microsomes. (A) 4a (10 μM) was incubated in the presence of human liver microsomes (1.0 mg/mL; HLMs; □□□), HLMs plus NADPH (1 mM; □□□), or buffer only (0.1 M KH₂PO₄, pH 7.4; ↔️), for 0, 15, 30 or 60 min at 37 °C. The data are expressed as percent of parent compound (4a) remaining at each time compared with time 0 min, and represent the mean ± SD of n = 3 or 4 independent determinations. (B) Representative stacked HPLC-fluorescence traces of supernatants from mixtures containing 4a (10 μM) and HLMs (1.0 mg/mL), incubated for 0, 15, 30 or 60 min at 37 °C. M, 4a metabolite.

3. Conclusion

With the aim to further explore the SARs of the 7-PPyQ class of compounds, we examined the effects of various oxygenated functionalities at the 3 N position. A small series of novel derivatives was synthesized, and their antiproliferative activities and with their mechanism of action were investigated. The chemical series included both angular and linear compounds and fully aromatic and partially hydrogenated derivatives. Most compounds had significant antiproliferative activity, inhibiting cell growth with nanomolar to micromolar GI₅₀ values, thus confirming that, in general, oxygenated substitutions at the 3 position improve cytotoxicity. In the series, the [3,2-f] angular geometry once again was required for obtaining potent cytotoxic compounds, and this [3,2-f] configuration was present both in the fully aromatic benzyol 4a and the methanesulfonyl 4b. Compound 4a was the most active of the new agents, even having sub-nanomolar GI₅₀s in some of the cell lines studied. Thus, compound 4a was even more cytotoxic than the previously described compound 11. Moreover, 4a had a significant effect only in rapidly proliferating cells but not in quiescent and proliferating lymphocytes. Investigations on the mechanism of action of 4a confirmed that it was a strong inhibitor of tubulin polymerization, as were previously described 7-PPyQ derivatives. Both in the Hela and Jurkat cell lines, 4a was more effective than 4b in blocking the cell cycle at the G₂M phase, inducing apoptosis through the mitochondrial death pathway with production of ROS. In agreement with the experimental results obtained in this work, docking simulations suggested that the synthesized inhibitors had high affinity for the colchicine site of tubulin, with interactions with the binding site most similar to those observed with the known inhibitor plinabulin. The 1,2-dihydro derivative 8b was slightly penalized by affecting the geometry of the hydrogen bond with βVal236.
Additionally, *in vitro* metabolic stability studies indicated that a human liver microsomes esterases can catalyze the cleavage of the amide bond of 4a, leading to formation of an active metabolite, namely compound 9. These findings may be valuable for future *in vivo* studies.

4. Experimental section

4.1. Chemistry

Melting points were determined on a Buchi M – 560 capillary melting point apparatus and are uncorrected. Infrared spectra were recorded on a PerkinElmer 1760 FTIR spectrometer with KBr pressed disks and on a Varian ATR FTIR; all values are expressed in cm⁻¹. UV–vis spectra were recorded on a Thermo Helyos α spectrometer. ¹H NMR spectra were determined on Bruker 300 and 400 MHz spectrometers, with the solvents indicated; chemical shifts are reported in δ (ppm) downfield from tetramethylsilane as internal reference. Coupling constants are given in Hertz. In the case of multiplets, chemical shifts were measured starting from the approximate center. Integrals were satisfactorily in line with those expected on the basis of compound structure. Elemental analyses were performed in the Microanalytical Laboratory, Department of Pharmaceutical Sciences, University of Padova, on a PerkinElmer C, H, N elemental analyzer model 240B, and analyses indicated by the symbols of the elements were within ±0.4% of the theoretical values. Analytical data are presented in detail for each final compound in the Supporting Information. Mass spectra were obtained on a Mat 112 Varian Mat Bremen (70 eV) mass spectrometer and on an Applied Biosystems Mariner System 5220 LC/MS (nozzle potential 140 eV). Column flash chromatography was performed on Merck silica gel (250–400 mesh ASTM); chemical reactions were monitored by analytical thin-layer chromatography (TLC) on Merck silica gel 60 F- 254 glass plates. Solutions were concentrated on a rotary evaporator under reduced pressure. Starting materials were purchased from Sigma-Aldrich and Alfa Aesar, and solvents were from Carlo Erba, Fluka and Lab-Scan. DMSO was obtained anhydrous by distillation under vacuum and stored on molecular sieves.

The purity of new tested compounds was checked by HPLC using the instrument HPLC VARIAN ProStar model 210, with detector DAD VARIAN ProStar 335. The analysis was performed with a flow of 1 mL/min, a C-18 column of dimensions 250 mm × 4.6 mm, a particle size of 5 μm, and a loop of 10 μL. The detector was set at 300 nm. The mobile phase consisted of phase A (Milli-Q H₂O, 18.0 MO, TFA 0.05%) and phase B (95% MeCN, 5% phase A). Gradient elution was performed as reported: 0 min, % B = 10; 0–20 min, % B = 90; 25 min, % B = 90; 26 min, % B = 10; 31 min, % B = 10.

4.1.1. General procedure for the synthesis of 1N-substituted nitroindoles (1a–e)

As a typical procedure, the synthesis of 1-benzoyl-5-nitro-1H-indole 1a is described in detail. Into a two-necked 50 mL round-bottomed flask, 0.666 g (27.7 mmol, 3 eq.) of NaH, 60% dispersion in mineral oil, was
placed and washed with toluene (3 × 10 mL). With stirring, a solution of commercial 5-nitroindole, 1.50 g (9.25 mmol, 1 eq.) in 5 mL of anhydrous DMF, was dropped into the flask, and the initial yellow color changed to red with the formation of H$_2$ gas. After 30 min at room temperature, a solution of benzoyl chloride, 3.21 mL (27.7 mmol, d = 1.21 g/mL, 3 eq.) in 3 mL dry DMF, was added, and the reaction mixture was stirred for 2 h. The reaction was monitored by TLC analysis (elucent toluene/n-Hex/EtOAc, 1:1:1). At the end of the reaction, 25 mL of water was added, and the solvent was evaporated under reduced pressure, leaving a residue, which was extracted with EtOAc (3 × 50 mL). The organic phase, washed with water and dried over anhydrous Na$_2$SO$_4$, was concentrated under vacuum giving a crude yellow solid (2.446 g). This crude product was purified with a silica gel chromatographic column 3 × 35 cm, 230–400 mesh, eluent toluene/n-Hex/EtOAc, 1:1:1, yielding 2.345 g of a pure yellow solid.

4.1.1.1. 1-Benzoyl-5-nitro-1H-indole (1a)

Yield: 94.9%. Rf = 0.83 (elucent toluene/n-Hex/EtOAc, 1:1:1); $^1$H NMR (400 MHz, DMSO-d$_6$) $\delta$ = 8.67 (d, $J$ = 2.25 Hz, 1H, H-4), 8.43 (d, $J$ = 9.12 Hz, 1H, H-7), 7.74 (m, 1H, H-4’), 8.28 (dd, $J$ = 9.12 Hz and 2.37 Hz, 1H, H-6), 7.82 (m, 2H, H-2’ and -H-6’), 7.64 (m, 2H, H-3’ and H-5’), 7.62 (d, $J$ = 3.64 Hz, 1H, H-2), 7.00 ppm (dd, $J$ = 3.78 Hz and $J$ = 0.60 Hz, 1H, H-3); $^{13}$C NMR (75 MHz, DMSO-d$_6$) $\delta$ = 109.74 (C-3), 113.19 (C-6), 117.05 (C-7), 120.55 (C-4), 129.69 (C-2’ and C-6’), 130.27 (C-3’ and C-5’), 131.56 (C-4’), 132.39 (C-2), 133.55 (C-3a), 133.94 (C-7a), 139.43 (C-1’), 144.66 (C-5), 169.24 ppm (C=O); HRMS (ESI-MS, 140 eV): m/z [M + H$^+$$] calculated for C$_{13}$H$_{11}$N$_2$O$_3$, 267.0779; found, 267.0787.

4.1.1.2. 1-Methanesulfonyl-5-nitro-1H-indole (1b)

Compound 1b was prepared as for compound 1a by reacting 0.888 g of NaH 60% (37.03 mmol, 3 eq.) and 2 g (12.34 mmol) of 5-nitroindole dissolved in 5 mL of DMF and 2.86 mL of methanesulfonyl chloride (37.03 mmol, d = 1.48 g/mL, 3 eq.). Reaction time: 2 h (TLC, eluent EtOAc/n-Hex, 2:1). 2.564 g of a solid bright yellow solid was obtained, and this was used in the next synthetic step without further purification. Yield: 91.9%. Rf = 0.54 (elucent EtOAc/n-Hex, 2:1); $^1$H NMR (400 MHz, DMSO-d$_6$) $\delta$ = 8.67 (d, $J$ = 2.20 Hz, 1H, H-4), 8.25 (dd, $J$ = 9.14 Hz and $J$ = 2.30 Hz, 1H, H-6), 8.05 (d, $J$ = 9.16 Hz, 1H, H-7), 7.85 (d, $J$ = 3.68 Hz, 1H, H-2), 7.08 (dd, $J$ = 7.08 Hz and $J$ = 0.72 Hz, 1H, H-3), 3.60 ppm (s, 3H, SO$_2$CH$_3$); $^{13}$C NMR (101 MHz, DMSO-d$_6$) $\delta$ = 42.07 (SO$_2$CH$_3$), 109.22 (C-3), 114.08 (C-7), 119.81 (C-6), 130.33 (C-2), 130.58 (C-3a), 137.71 (C-7a), 143.96 ppm (C-5); HRMS (ESI-MS, 140 eV): m/z [M + H$^+$$] calculated for C$_9$H$_8$N$_2$O$_5$S, 241.0283; found, 241.0236.

4.1.1.3. 1-(4-Methylbenzenesulfonyl)-5-nitro-1H-indole (1c)

Compound 1c was prepared as for compound 1a by reacting 0.444 g of NaH 60% (18.51 mmol, 3 eq.) and 1 g (6.17 mmol) of 5-nitroindole dissolved in 5 mL of DMF and 5.53 g of $p$-toluenesulfonyl chloride (18.51 mmol, 3 eq.). Reaction time: 2 h (TLC, eluent n-Hex/EtOAc, 9:1). 1.786 g of a solid bright orange solid was obtained
which was used in the next synthetic step without any additional further purification. Yield: 91.6%. \text{RF} = 0.76
\text{(eluent EtOAc/n-Hex, 9:1); \text{H NMR} (400 MHz, DMSO-d_6) \delta = 8.59 \text{ (d, } J = 1.88 \text{ Hz, 1H, H-4), 8.21
\text{(dd, } J = 9.18 \text{ Hz and } J = 2.26 \text{ Hz, 1H, H-6), 8.15 \text{(dt, } J = 9.18 \text{ Hz and } J = 0.98 \text{ Hz, 1H, H-7), 8.07 \text{(d, } J = 3.72 \text{ Hz, } H-2, 7.94 \text{(m, } \text{AA’BB’}, \text{ J = 8.44 Hz, } J = 2.11 \text{ Hz and } J = 1.83 \text{ Hz, 2H, H-2’ and H-6’), 7.43 \text{(m, } \text{AA’BB’}, \text{ J = 8.58 Hz and J = 0.62 Hz, 2H, H-3’ and H-5’), 7.07 \text{(dd, } J = 3.70 \text{ Hz and } J = 0.66 \text{ Hz, 1H, H-3, 2.23 ppm
\text{(s, 3H, -CH}_3); ^{13}\text{C NMR (75 MHz, DMSO-d_6) } \delta = 21.91 (-\text{CH}_3), 110.93 \text{(C-3), 114.54 \text{(C-7), 118.85 \text{(C-4), 120.55
\text{(C-6), 127.77 \text{(C-2’ and C-6’), 130.90 \text{(C-2), 131.34 \text{(C-3a), 131.37 \text{(C-3’ and C-5’), 134.59 \text{(C-4’), 137.81 \text{(C-7a), 144.61 \text{(C-5), 147.09 ppm (C-1’); HRMS (ESI-MS, 140 eV): } m/z [M + H^+] \text{ calculated for C}_{13}\text{H}_{13}\text{N}_3\text{O}_4\text{S}, 317.0596; found, 317.0585.}

4.1.1.4. 5-Nitro-1-[4-(trifluoromethyl)benzenesulfonyl]-1H-indole (1d)

Compound 1d was prepared as described for compound 1a by reacting 0.670 g of NaH 60% (27.75 mmol, 3 eq.) and 1.5 g (9.25 mmol) of 5-nitroindole dissolved in 5 mL of DMF and 3.39 g of 4-(trifluoromethyl)benzenesulfonyl chloride (13.87 mmol, 1.5 eq.). Reaction time: 1 h (TLC, eluent n-Hex/EtOAc, 9:1). 2.845 g of a solid bright yellowish powdery solid was obtained, and this was used in the next synthetic step without further purification. Yield: 82.7%. \text{RF} = 0.79 \text{(eluent EtOAc/n-Hex/toluene, 1:1:1); \text{H NMR} (400 MHz, DMSO-d_6) \delta = 8.60 \text{ (d, } J = 2.12 \text{ Hz, 1H, H-4), 8.29 \text{(m, } \text{AA’BB’}, \text{ J = 8.28 Hz, 2H, H-2’ and H-6’), 8.22 \text{(dd, } J = 9.20 \text{ Hz and } J = 2.16 \text{ Hz, 1H, H-6), 8.18 \text{(d, } J = 9.16 \text{ Hz, 1H, H-7), 8.13 \text{(d, } J = 3.72 \text{ Hz, 1H, H-2), 8.01
\text{(m, } \text{AA’BB’}, \text{ J = 8.44 Hz, 2H, H-3’ and H-5’), 7.13 ppm \text{(d, } J = 3.72 \text{ Hz, 1H, H-1); ^{13}\text{C NMR (75 MHz, DMSO-d_6) } \delta = 111.29 \text{(C-3), 114.17 \text{(C-7), 118.59 \text{(C-4), 120.49 \text{(C-6), 123.41 \text{(q, } J = 273.30 \text{ Hz, -CF}_3), 127.8250
\text{(q, } J = 3.65 \text{ Hz, C-3’ and C-5’), 128.46 \text{(C-2’ and C-6’), 130.44 \text{(C-2), 131.12 \text{(C-3a), 134.84 \text{(q, } J = 32.69 \text{ Hz, C-4’), 137.42 \text{(C-7a), 140.53 \text{(C-1’), 144.46 ppm (C-5); HRMS (ESI-MS, 140 eV): } m/z [M + H^+] \text{ calculated for C}_{13}\text{H}_{13}\text{F}_3\text{N}_3\text{O}_4\text{S}, 371.0313; found, 371.0309.}

4.1.1.5. N-cyclopropyl-5-nitro-1H-indole-1-carboxamide (1e)

To a stirred slurry of NaH (0.630 g of a 60% mineral oil dispersion, 26.27 mmol, 3 eq.) in THF (25 mL) at 0 °C, under N_2, was cautiously added 5-nitroindole (1.42 g, 8.75 mmol, 1 eq.) previously dissolved in THF (5 mL). The reaction mixture was stirred at 0 °C for 60 min, then transferred, via cannula, to a solution of 4-nitrophenyl chloroformate (2.11 g, 10.5 mmol, 1.2 eq.) in THF (8.5 mL). The resultant reaction mixture was stirred at ambient temperature for 15 h (TLC, eluent n-Hex/EtOAc, 2:1), prior to removal of the solvent by concentration in vacuo. The residue obtained was suspended in EtOAc (100 mL), then filtered and washed with EtOAc and Et_2O to give 2.515 g (87.6%) of a pale yellow solid. The resulting activated carbamate 5-nitro-1-[4-nitrophenoxycarbonyl]indole was immediately used as follows: a 2.0 M solution of cyclopropylamine (0.382 mL, 5.48 mmol, d = 0.824 g/mL, 8 eq.) in THF (2.75 mL) was added to a solution of 5-nitro-1-[4-nitrophenoxycarbonyl]indole (0.245 g, 0.686 mmol, 1 eq.) in THF (5 mL). The resultant reaction mixture was...
stirred at ambient temperature for 4 h and monitored by TLC (eluent n-Hex/EtOAc, 2:1), prior to removal of the solvent by concentration to dryness in vacuo. The residue obtained was partitioned between EtOAc (100 mL) and H$_2$O (100 mL). The layers were separated, and the aqueous phase was extracted with EtOAc (2 × 30 mL). The combined organic extracts were washed with sat’d NaHCO$_3$ (100 mL), brine and finally dried over NaSO$_4$ and concentrated in vacuo to give a yellow solid which was suspended in Et$_2$O (35 mL), filtered and washed with Et$_2$O (2 × 20 mL) to give 0.066 g of a pale yellow solid. Yield: 39.1%. Rf = 0.26 (eluent n-Hex/EtOAc, 2:1); $^1$H NMR (400 MHz, DMSO-$d_6$) $\delta$ = 8.59 (d, $J$ = 2.25 Hz, 1H, H-4), 8.50 (d, $J$ = 2.31 Hz, 1H, NH), 8.39 (d, $J$ = 9.18 Hz, 1H, H-7), 8.15 (dd, $J$ = 9.21 Hz and $J$ = 2.40 Hz, 1H, H-6), 8.02 (d, $J$ = 3.89 Hz, 1H, H-2), 6.92 (dd, $J$ = 3.69 Hz and $J$ = 0.55, 1H, H-3), 2.80 (sex $J$ = 3.12 Hz, 1H, NH-CH), 0.8–0.6 ppm (m, 4H, -CH$_2$CH$_2$-); $^{13}$C NMR (75 MHz, DMSO-$d_6$) $\delta$ = 6.75 (-CH$_2$CH$_2$-), 22.75 (NCH-CH$_2$CH$_2$-), 104.32 (C-3), 112.16 (C-7), 116.74 (C-6), 117.62 (C-4), 127.30 (C-2), 129.55 (C-3a), 135.38 (C-7a), 140.95 (C-5), 160.11 ppm (C═O); HRMS (ESI-MS, 140 eV): $m/z$ [M + H$^+$] calculated for C$_{15}$H$_{12}$N$_2$O$_3$$^+$, 246.0879; found, 246.0871.

4.1.2. General procedure for the synthesis of 1N-substituted aminoindoles 2a–d

As a typical procedure, the synthesis of 1-benzoyl-5-amo-1H-indole 2a is described in detail. Into a two-necked 50 mL round-bottomed flask, 3.241 g of 1-benzoyl-5-nitro-1H-indole (1a) (12.17 mmol, 1 eq.), 10.986 g of SnCl$_2$-2H$_2$O (48.68 mmol, 4 eq.), 2 mL of HCl 37% and 30 mL of methanol were added. The reaction mixture was refluxed for 3 h, and the reaction progress was monitored by TLC (n-Hex/EtOAc, 1:1). At the end, the solvent was evaporated, the residue was taken up with aqueous NaOH 20% (20 mL), and the resulting suspension was extracted with diethyl ether (4 × 50 mL). The combined extracts, washed with brine and treated with anhydrous Na$_2$SO$_4$, were evaporated to dryness on a rotary evaporator to yield 1.536 g of a semisolid yellow product, made up of three different reaction products. In order to obtain the desired pure 1-benzoyl-5-amo-1H-indole, the raw powder was purified in a silica gel chromatographic column 3 × 28 cm, 230–400 mesh, eluent n-Hex/EtOAc, 1:1, yielding 0.267 g of a pure yellow solid.

4.1.2.1. 1-Benzoyl-5-amo-1H-indole (2a)

Yield: 19.5%. Rf = 0.33 (eluent n-Hex/EtOAc, 1:1); $^1$H NMR (400 MHz, DMSO-$d_6$) $\delta$ = 7.98 (d, $J$ = 8.73 Hz, 1H, H-7), 7.70 (m, 2H, H-2’ and H-6’), 7.66 (m, 1H, H-4’), 7.59 (m, 2H, H-3’ and H-5’), 7.16 (d, $J$ = 3.72 Hz, 1H, H-3), 6.75 (d, $J$ = 2.07 Hz, 1H, H-4), 6.65 (dd, $J$ = 8.74 Hz and $J$ = 2.20 Hz, 1H, H-6), 6.51 (d, $J$ = 3.42 Hz, 1H, H-2), 5.05 ppm (s, 2H, NH$_2$); $^{13}$C NMR (75 MHz, DMSO-$d_6$) $\delta$ = 101.99 (C-3), 113.14 (C-7), 115.82 (C-6), 118.34 (C-4), 128.44 (C-2’ and C-6’), 129.04 (C-3’ and C-5’), 129.47 (C-4’), 131.83 (C-1’), 133.84 (C-2), 134.72 (C-3a), 137.46 (C-7a), 144.04 (C-5), 172.01 ppm (C═O); HRMS (ESI-MS, 140 eV): $m/z$ [M + H$^+$] calculated for C$_{15}$H$_{12}$N$_2$O$_3$$^+$, 237.1028; found, 237.1031.
4.1.2.2. 1-Methanesulfonyl-5-amino-1H-indole (2b)

Compound 2b was prepared as described for compound 2a by reacting 1 g (4.16 mmol, 1 eq.) of the appropriate S-nitroindole derivative 1b and 4.69 g of SnCl2·2H2O (20.80 mmol, 5 eq.), obtaining 0.963 g of a slightly brown solid. Yield: 80.6%; Rf = 0.37 (eluent n-Hex/EtOAc, 1:1); 1H NMR (400 MHz, DMSO-d6) δ = 7.48 (dt, J = 8.76 Hz and J = 0.66 Hz, 1H, H-7), 7.34 (d, J = 3.96 Hz, 1H, H-2), 6.75 (dd, J = 2.20 Hz and J = 0.50 Hz, 1H, H-4), 6.67 (dd, J = 8.76 Hz and J = 2.24 Hz, 1H, H-6), 6.59 (dd, J = 3.64 Hz and J = 0.75 Hz, 1H, H-3), 4.95 (bs, 2H, NH2), 3.24 ppm (s, 3H, SO2CH3); 13C NMR (101 MHz, DMSO-d6) δ = 105.92 (C-4), 110.02 (C-3), 114.18 (C-7), 114.51 (C-6), 124.55 (q, J = 245.01 Hz, CF3), 127.00 (C-3′ and C-5′), 127.50 (C-2), 127.81 (C-3a), 130.45 (C-2′ and C-6′), 133.83 (q, J = 33.01 Hz, C-4′), 134.69 (C-7a), 143.98 (C-5), 145.87 ppm (C-5′); HRMS (ESI-MS, 140 eV): m/z [M + H+] calculated for C11H11N2O2S+, 211.0541; found, 211.0542.

4.1.2.3. 1-[4-(Trifluoromethyl)benzenesulfonyl]-5-amino-1H-indole (2d)

Compound 2d was prepared as described for compound 2a by reacting 1.25 g (3.37 mmol, 1 eq.) of the appropriate S-nitroindole derivative 1d and 3.80 g of SnCl2·2H2O (16.87 mmol, 5 eq.), obtaining 1.160 g of a slightly brown solid. Yield: 99.9%; Rf = 0.30 (eluent n-Hex/EtOAc/toluene, 1:1:1); 1H NMR (400 MHz, DMSO-d6) δ: 7.77 (m, J = 8.40 Hz, J = 1.98 and J = 1.76 Hz, 2H, H-3′ and H-5′), 7.67 (d, J = 8.76 Hz, 1H, H-7), 7.58 (d, J = 3.60 Hz, 1H, H-2), 7.30 (m, J = 8.12 Hz, 2H, H-2′ and H-6′), 6.77 (d, J = 2.04 Hz, 1H, H-4), 6.73 (dd, J = 8.72 Hz and J = 2.09 Hz, 1H, H-6), 6.61 (dd, J = 3.64 Hz and J = 0.60 Hz, 1H, H-3), 4.98 ppm (bs, 2H, -NH2); 13C NMR (101 MHz, DMSO-d6) δ: 105.92 (C-4), 110.02 (C-3), 114.18 (C-7), 114.51 (C-6), 124.55 (q, J = 245.01 Hz, CF3), 127.00 (C-3′ and C-5′), 127.50 (C-2), 127.81 (C-3a), 130.45 (C-2′ and C-6′), 133.83 (q, J = 33.01 Hz, C-4′), 134.69 (C-7a), 143.98 (C-5), 145.87 ppm (C-5′); HRMS (ESI-MS, 140 eV): m/z [M + H+] calculated for C13H12F3N2O2S+, 341.0572; found, 341.0569.

4.1.3. General procedure for the synthesis of 1N-substituted aminooindoles 2c, 5a, 5b, 5d, 5e

As a typical procedure, the synthesis of 1-(4-methylbenzenesulfonyl)-5-amino-1H-indole 2c is described in detail. Into a three-necked flask of 500 mL, previously dried in an oven, about 0.300 g of C/Pd 10% and approximately 60 mL of EtOAc were placed. After connecting the flask to an elastomer balloon containing H2 gas, the mixture was stirred at room temperature for 1 h in order to saturate the suspension of C/Pd with H2. Then, 1.9 g (6.00 mmol) of the appropriate 5-nitroindole derivative 1c in 15 mL of EtOAc was added dropwise to the suspension, and the mixture was stirred under H2 at atmospheric pressure and heated by means of an oil bath at 50–60 °C for 15 h, monitoring the progress of the reaction by TLC analysis (EtOAc/n-Hex, 9:1). At the end of the reaction, the mixture was filtered, and the solution was concentrated to dryness on a rotary evaporator to give 1.680 g of semisoloid dark purple sticky product.
4.1.3.1. 1-(4-Methylbenzenesulfonyl)-5-amino-1H-indole (2c)

Yield: 97.8%. Rf = 0.76 (eluent n-Hex/EtOAc, 9:1); 1H NMR (400 MHz, DMSO-d6) δ = 7.77 (m, AA’BB’, J = 8.40 Hz, J = 1.98 Hz and J = 1.76 Hz, 2H, H-2’ and H-6’), 7.67 (d, J = 8.76 Hz, 1H, H-7), 7.58 (d, J = 3.60 Hz, 1H, H-2), 7.30 (m, AA’BB’, J = 8.12 Hz, 2H, H-3’ and H-5’), 6.77 (d, J = 2.04 Hz, 1H, H-4), 6.73 (dd, J = 8.72 Hz and J = 2.09 Hz, 1H, H-6), 6.61 (dd, J = 3.64 Hz and J = 0.60 Hz, 1H, H-3), 4.95 (bs, 2H, -NH2), 2.26 ppm (s, 3H, -CH3); 13C NMR (101 MHz, DMSO-d6) δ = 21.40 (C-H), 105.92 (C-4), 110.02 (C-3), 114.18 (C-7), 114.51 (C-6), 127.00 (C-2’ and C-6’), 127.50 (C-2), 127.81 (C-3a), 130.45 (C-3’ and C-5’), 132.27 (C-4’), 134.69 (C-7a), 143.98 (C-5), 145.87 ppm (C-1’); HRMS (ESI-MS, 140 eV): m/z [M + H+] calculated for C15H15N2O2S', 287.0854; found, 287.0851.

4.1.3.2. 1-Benzoyl-2,3 dihydro-5-amino-1H-indole (5a)

Yield: 98.8%. Rf = 0.54 (eluent n-Hex/EtOAc, 5:4); 1H NMR (400 MHz, DMSO-d6) δ = 7.94 (d, J = 8.40, 1H, H-6), 7.49 (m, 6H, H-7, H-2’, H-3’, H-4’, H-5’ and H-6’), 6.48 (d, J = 1.89, 1H, H-4), 4.98 (bs, 2H, -NH2), 3.87 (t, J = 8.15 Hz, 2H, H2-2), 2.93 ppm (t, J = 8.14, 2H, H2-3); 13C NMR (101 MHz, DMSO-d6) δ = 29.27 (N-CH2CH2), 51.24 (N-CH2CH2), 110.98 (C-4), 113.94 (C-6), 115.98 (C-7), 128.24 (C-2’ and C-6’), 128.99 (C-3’ and C-5’), 129.32 (C-4’), 132.37 (C-3a), 134.23 (C-7a), 135.54 (C-1’), 143.82 (C-5), 166.83 ppm (C=O); HRMS (ESI-MS, 140 eV): m/z [M + H+] calculated for C15H15N2O+, 239.1184; found, 239.1179.

4.1.3.3. 1-Methanesulfonyl-2,3 dihydro-5-amino-1H-indole (5b)

Yield: 94.8%. Rf = 0.15 (eluent n-Hex/EtOAc, 2:1); 1H NMR (400 MHz, DMSO-d6) δ = 6.95 (d, J = 8.48 Hz, 1H, H-7), 6.50 (m, J = 2.37 Hz, 1H, H-4), 6.39 (dd, J = 8.50 Hz and J = 2.34 Hz, 1H, H-6), 4.91 (bs, 2H, -NH2), 3.82 (t, J = 8.24 Hz, 2H, N-CH2CH2), 2.96 (t, J = 8.20 Hz, 2H, N-CH2CH2), 2.81 ppm (s, 3H, SO2CH3); 13C NMR (101 MHz, DMSO-d6) δ = 28.30 (N-CH2CH2), 33.48 (SO2CH3), 50.39 (N-CH2CH2), 111.26 (C-4), 112.92 (C-6), 115.57 (C-7), 131.88 (C-3a), 133.66 (C-7a), 146.17 ppm (C-5); HRMS (ESI-MS, 140 eV): m/z [M + H+] calculated for C8H14NO2S', 213.0698; found, 213.0663.

4.1.3.4. 1-[4-(Trifluoromethyl)benzenesulfonyl]-2,3 dihydro-5-amino-1H-indole (5d)

Yield: 98.2%. Rf = 0.51 (eluent n-Hex/EtOAc/toluene, 1:1:1); 1H NMR (400 MHz, DMSO-d6) δ = 7.90 (m, J = 8.98 Hz, 2H, H-3’ and H-5’), 7.40 (m, J = 8.98 Hz, 2H, H-2’ and H-6’), 7.19 (d, J = 8.52 Hz, 1H, H-7), 6.41 (dd, J = 8.50 Hz and J = 2.26 Hz, 1H, H-6), 6.33 (d, J = 2.16 Hz, 1H, H-4), 4.99 (s, 2H, -NH2), 3.85 (t, J = 8.11 Hz, 2H, H2-3), 2.56 ppm (t, J = 8.10 Hz, 2H, H2-2); 13C NMR (101 MHz, DMSO-d6) δ = 28.17 (C-2), 50.69 (C-3), 111.00 (C-4), 113.13 (C-6), 116.93 (C-7), 126.27 (q, J = 3.80 Hz, C-3’ and C-5’), 127.35 (q, J = 269.56 Hz, CF3), 128.57 (C-2’ and C-6’), 132.44 (C-3a), 133.35 (q, J = 32.32 Hz, C-4’), 134.44 (C-7a), 140.59 (C-1’), 146.93 ppm (C-5); HRMS (ESI-MS, 140 eV): m/z[M + H+] calculated for C15H14F3N2O2S', 341.0572; found, 341.0569.
4.1.3.5. N-cyclopropyl-2,3-dihydro-5-amino-1H-indole-1-carboxamide (5e)

Yield: 91.2%. Rf = 0.23 (eluent n-He/EtOAc, 2:1); 1H NMR (400 MHz, DMSO-d6) δ = 7.56 (d, J = 8.43 Hz, 1H, H-7), 6.46 (m, 2H, H-4 and NH), 6.35 (dd, J = 8.44 Hz and J = 2.32 Hz, 1H, H-6), 4.63 (s, 2H, -NH2), 3.77 (t, J = 8.58 Hz, 2H, H-2), 2.99 (t, J = 8.53 Hz, 2H, H-3), 2.60 (sex, J = 3.46 Hz, 1H, NH-CH-CH2CH2-), 0.6–0.4 ppm (m, 4H, -CH2CH2-); 13C NMR (101 MHz, DMSO-d6) δ = 6.42 (-CH2CH2-), 23.47 (C-3), 27.89 (C-2), 46.89 (NH-CH-CH2CH2-), 111.27 (C-4), 112.39 (C-6), 115.04 ppm (C-7), 130.32 (C-3a), 133.83 (C-7a), 143.82 (C-5), 162.35 ppm (C=O); HRMS (ESI-MS, 140 eV): m/z[M + H+] calculated for C13H12N3O+, 218.1293; found, 218.1245.

4.1.4. General procedure for the synthesis of acrylate derivatives 3b–d and 6a, 6b, 6d and 6e

As a typical procedure, the synthesis of (E,Z)-Ethyl 3-(1-(methanesulfonyl)-1H-indol-5-ylamino)-3-phenylacrylate 3b is described in detail. In a 100 mL round-bottomed flask, 1.4 g (6.66 mmol, 1 eq.) of 1-methanesulfonyl-5-amino-1H-indole 2a in 25 mL of absolute ethanol was condensed with 1.73 g (17 mmol, 1 eq.) of commercial ethyl benzoylacetate and 0.5 mL of glacial acetic acid in the presence of 100 mg of Drierite (anhydrous CaSO4). The mixture was refluxed for about 24 h, the reaction being monitored by TLC analysis (eluent n-He/EtOAc, 2:1). Even though the reaction was not complete after 24 h, the mixture was cooled and filtered to remove the Drierite; the resulting solution was evaporated to dryness under vacuum and the residue (2.420 g) purified by silica gel chromatography (3 × 35 cm, 230–400 mesh, eluent n-He/EtOAc, 2:1) to yield 1.54 g of a deep yellow powdery solid.

4.1.4.1. (E,Z)-Ethyl 3-(1-(methanesulfonyl)-1H-indol-5-ylamino)-3-phenylacrylate (3b)

Yield: 60.2%. Rf = 0.65 (eluent n-He/EtOAc, 2:1); 1H NMR (400 MHz, DMSO-d6) δ = 10.25 (s, 1H, NH), 7.58 (d, J = 8.80 Hz, 1H, H-7), 7.49 (d, J = 3.68 Hz, 1H, H-2), 7.36 (m, 5H, 2′,3′,4′-5′,6′-H), 7.02 (d, J = 2.16 Hz, 1H, H-4), 6.84 (dd, J = 8.82 Hz and 2.18 Hz, 1H, H-6), 6.62 (dd, J = 3.68 Hz and J = 0.72 Hz, 1H, H-3), 4.94 (s, 1H, C=C-H), 4.15 (q, J = 7.09 Hz, 2H, OCH2CH3), 3.38 (s, 3H, SO2CH3), 1.24 ppm (t, J = 7.10 Hz, 3H, OCH2CH3); 13C NMR (101 MHz, DMSO-d6) δ = 14.76 (OCH2CH3), 41.35 (SO2CH3), 59.21 (OCH2CH3), 90.70 (C=C-H), 109.09 (C-3), 113.26 (C-7), 115.26 (C-4), 120.62 (C-6), 127.80 (C-2), 128.45 (C-2′ and C-6′), 128.94 (C-3′ and C-5′), 130.00 (C-4′), 130.77 (C-3a), 131.07 (C-1′), 135.80 (C-7a), 136.22 (C-5), 159.49 (C=C-H), 169.45 ppm (COOCH2CH3); HRMS (ESI-MS, 140 eV): m/z[M + H+] calculated for C20H12N2O4S+, 385.1222; found, 385.1213.

4.1.4.2. (E,Z)-Ethyl 3-(1-(p-toluenesulfonyl)-1H-indol-5-ylamino)-3-phenylacrylate (3c)

Compound 3c was prepared as described for compound 3b by reacting 3.177 g (11.10 mmol) of the appropriate 5-aminoindole derivative 2c, obtaining after column chromatography 2.336 g of a brownish sticky semisolid product. Yield: 45.7%. Rf = 0.62 (eluent n-He/EtOAc, 2:1); 1H NMR (400 MHz, DMSO-d6) δ = 10.18 (s, 1H, NH), 7.79 (m, AA′BB′, J = 8.32 Hz, 2H, H-2′ and H-6′), 7.69 (d, J = 3.64 Hz, 1H, H-2), 7.65
Compound 3c was prepared as described for compound 3b by reacting 1.160 g (3.41 mmol) of the appropriate 5-aminooindole derivative 2d, obtaining after column chromatography 1.020 g of a yellow powdery solid. Yield: 58.1%. Rf = 0.78 (eluent n-Hex/EtOAc, 8:2); 1H NMR (400 MHz, DMSO-d6) δ = 10.18 (s, 1H, NH), 7.79 (m, AA′BB′, J = 8.32 Hz, 2H, H-2′ and H-6′), 7.69 (d, J = 3.64 Hz, 1H, H-2), 7.65 (d, J = 8.84 Hz, 1H, H-7), 7.35 (m, AA′BB′, J = 8.61 Hz, 2H, H-3′ and H-5′), 7.31 (m, 5H, H-2″, H-3″, H-4″, H-5″ and H-6″), 6.78 (dd, J = 8.61 Hz and J = 1.82 Hz, 1H, H-6), 6.61 (d, J = 3.68 Hz, 1H, H-3), 4.92 (s, 1H, CH=C-H), 4.12 (q, J = 7.14 Hz, 2H, -OCH2CH3), 2.31 (s, 3H, -CH3), 1.22 ppm (t, J = 7.17 Hz, 3H, -OCH2CH3); 13C NMR (101 MHz, DMSO-d6) δ = 14.85 (-OCH2CH3), 51.54 (-OCH2CH3), 90.99 (C=C-H), 109.72 (C-3), 113.65 (C-7), 115.30 (C-4), 120.87 (C-6), 127.13 (C-2′ and C-6′), 128.24 (C-2), 128.53 (C-2″ and C-6″), 129.03 (C-3″ and C-5″), 130.11 (C-4″), 130.65 (C-3′ and C-5′), 130.91 (C-3a), 131.26 (C-1″), 134.53 (C-4″), 135.85 (C-7a), 136.69 (C-5), 145.95 (C-1′), 159.39 (C=C-H), 169.48 ppm (COOCH2CH3); HRMS (ESI-MS, 140 eV): m/z [M + H+] calculated for C28H28N2O5S2, 461.1535; found, 461.1539.

4.1.4.3. (E,Z)-Ethyl-3-(1-(4-(trifluoromethyl)benzenesulfonyl)-1H-indol-5-ylamino)-3-phenylacrylate (3d)

Compound 5a was prepared as described for compound 3b by reacting 1.96 g (8.22 mmol) of the appropriate 5-aminooindole derivative 5a, obtaining after column chromatography 0.940 g of a brown viscous oil. Yield: 27.7%. Rf = 0.70 (eluent n-Hex/EtOAc/toluene, 1:1:1); 1H NMR (400 MHz, DMSO-d6) δ = 10.14 (s, 1H, NH), 7.95 (m, 2H, H-2′ and H-6′), 7.68 (m, 1H, H-4′), 7.60–7.32 (m, 8H, H-2″, -3″, -5″, -6″ and H-3′, -5′, and H-6, H-7), 6.72 (d, J = 1.32 Hz, 1H, H-4), 4.90 (s, 1H, CH=C-H), 4.14 (q, J = 7.08 Hz, 2H, -OCH2CH3), 3.92 (t, J = 8.23 Hz, 2H, H-3), 2.90 (t, J = 8.32 Hz, 2H, H-2), 1.20 ppm (t, J = 7.09 Hz, 3H, -CH2CH3); 13C NMR (101 MHz, DMSO-d6) δ = 15.24 (-OCH2CH3), 29.23 (C-3), 52.26 (C-2), 58.92 (-OCH2CH3), 91.07 (C=C-H), 114.23 (C-7), 116.32 (C-4), 119.28 (C-6), 127.11 (C-2′ and C-6′), 128.52 (C-2″ and C-6″), 129.91 (C-3′ and C-5′), 130.23 (C-3″ and C-5″), 130.91 (C-3a), 131.18 (C-1″), 133.84 (C-4″) 134.82 (C-4′), 135.75 (C-7a), 136.82 (C-5), 146.23 (C-1′), 158.83 (C=C-H), 165.24 (NC=O), 171.91 ppm (COOCH2CH3); HRMS (ESI-MS, 140 eV): m/z [M + H+] calculated for C26H26N2O5S, 413.1865; found, 413.1859.
4.1.4.5. (E,Z)-Ethyl 3-(1-(methanesulfonyl)-2,3-dihydro-1H-indol-5-ylamino)-3-phenylacrylate (6b)

Compound 6b was prepared as described for compound 3b by reacting 0.460 g (2.16 mmol) of the appropriate S-aminoindole derivative 5b, obtaining after column chromatography 0.540 g of a brown sticky oil. Yield: 64.8%. Rf = 0.37 (elucent n-Hex/EtOAc, 2:1); 1H NMR (400 MHz, DMSO-d6) δ = 10.11 (s, 1H, NH), 7.38 (m, 1H, H-4′), 7.34 (m, 4H, H-2′, -3′, -5′ and 6′), 6.96 (d, J = 8.56 Hz, 1H, H-7), 6.72 (d, J = 2.20 Hz, 1H, H-4), 6.52 (dd, J = 8.72 Hz and J = 2.22 Hz, 1H, H-6), 4.90 (s, 1H, C＝C＝H), 4.31 (q, J = 7.09 Hz, 2H, OCH2CH3), 3.86 (t, J = 8.46 Hz, 2H, H-2), 2.93 (t, J = 8.44 Hz, 2H, H-3), 2.90 (s, 3H, SO2CH3), 1.23 ppm (t, J = 7.08 Hz, 3H, OCH2CH3); 13C NMR (101 MHz, DMSO-d6) δ = 14.74 (OCH2CH3), 27.68 (C-3), 34.38 (SO2CH3), 50.31 (C-2), 59.21 (OCH2CH3), 90.63 (C＝C＝H), 113.64 (C-7), 120.18 (C-4), 122.20 (C-6), 128.38 (C-2′ and C-6′), 129.92 (C-3′ and C-5′), 130.03 (C-4′), 133.07 (C-3a), 135.75 (C-7a), 136.46 (C-1′), 138.14 (C-5), 159.19 (C＝C＝H), 169.39 ppm (COOCH2CH3); HRMS (ESI-MS, 140 eV): m/z [M + H+] calculated for C20H23N2O4S+, 387.1379; found, 387.1381.

4.1.4.6. (E,Z)-Ethyl 3-(1-(4-(trifluoromethyl)benzenesulfonyl)-2,3-dihydro-1H-indol-5-ylamino)-3-phenylacrylate (6d)

Compound 6d was prepared as described for compound 3b by reacting 1.560 g (4.55 mmol) of the appropriate S-aminoindole derivative 5d, obtaining after column chromatography 1.29 g of a bright yellow powdery solid. Yield: 54.8%. Rf = 0.75 (elucent n-Hex/EtOAc, 8:2); 1H NMR (400 MHz, DMSO-d6) δ = 10.15 (s, 1H, NH), 7.92 (m, J = 8.98 Hz, 2H, H-3″ and H-5″), 7.45 (m, J = 8.98 Hz, 2H, H-2″ and H-6″), 7.38 (m, 1H, H-4′), 7.34 (m, 4H, H-2′, -3′, -5′ and -6′), 6.96 (d, J = 8.56 Hz, 1H, H-7), 6.72 (d, J = 2.20 Hz, 1H, H-4), 6.52 (dd, J = 8.72 Hz and J = 2.22 Hz, 1H, H-6), 4.96 (s, 1H, C＝C＝H), 4.24 (q, J = 7.09 Hz, 2H, OCH2CH3), 3.96 (t, J = 8.46 Hz, 2H, H-2), 2.86 (t, J = 8.44 Hz, 2H, H-3), 1.15 ppm (t, J = 7.08 Hz, 3H, OCH2CH3); 13C-NMR (101 MHz, DMSO-d6) δ = 14.74 (OCH2CH3), 27.68 (C-3), 50.31 (C-2), 59.21 (OCH2CH3), 90.63 (C＝C＝H), 113.64 (C-7), 120.18 (C-4), 122.20 (C-6), 126.47 (q, J = 3.75 Hz, C-3″ and C-5″), 127.98 (q, J = 265.67 Hz, CF), 128.38 (2′- and 6′-C), 128.81 (C-2″ and C-6″), 129.92 (3′- and 5′-C), 130.03 (C-4′), 133.07 (C-3a), 134.72 (q, J = 32.81 Hz, C-4″), 135.75 (C-7a), 136.46 (C-1′), 138.14 (C-5), 159.19 (C＝C＝H), 169.39 ppm (COOCH2CH3); HRMS (ESI-MS, 140 eV): m/z [M + H+] calculated for C28H22F4N2O4S+, 517.1409; found, 517.1401.

4.1.4.7. (E,Z)-Ethyl 3-(1-(N-cyclopropyl-1-carboxamide)-2,3-dihydro-1H-indol-5-ylamino)-3-phenylacrylate (6e)

Compound 6e was prepared as described for compound 3b by reacting 0.128 g (0.589 mmol) of the appropriate S-aminoindole derivative 5e, obtaining after column chromatography 0.139 g of a dark brown tarry oil. Yield: 60.4%. Rf = 0.56 (elucent CHCl3/MeOH, 95:5); 1H NMR (400 MHz, DMSO-d6) δ = 10.08 (s, 1H, NH), 7.52 (d, J = 8.55 Hz, 1H, H-7), 7.33 (m, 5H, H-2′, -3′, -4′, -5′ and -6′), 6.62 (m, 2H, H-4 and NH), 6.41 (dd, J = 8.55 Hz and J = 2.16 Hz, 1H, H-6), 4.83 (s, 1H, C＝CH), 4.12 (q, J = 7.08 Hz, 2H, -OCH2CH3), 3.74
As a typical procedure, the synthesis of 3-methanesulfonyl-7-phenyl-6H-pyrrolo[3,2-f]quinolin-9-one 4b is described in detail. In a two-necked round-bottomed flask, 20 mL of diphenyl ether was heated to boiling. To this 0.383 g (4.2 mmol) of the appropriate phenylacrylate derivative 3b was added portionwise, and the resulting mixture was refluxed for 15 min. After cooling to room temperature, 25 mL of diethyl ether was added, and the mixture was left for 12 h. The precipitate was collected by filtration and washed many times with diethyl ether. The product (0.437 g) was additionally purified by silica gel column chromatography (2.5 × 30 cm, 230–400 mesh, eluent CHCl₃/MeOH, 95:5) obtaining 0.303 g of a slightly brown solid.

4.1.5.1. 3-Methanesulfonyl-7-phenyl-6H-pyrrolo[3,2-f]quinolin-9-one (4b)

Yield: 89.9%; Rf = 0.16 (blue fluorescent spot, eluent CHCl₃/MeOH, 95:5); mp: 321.5 °C (decomposition); UV-Vis (H₂O/MeOH, 99:1): λmax (A) = 274 (A = 0.874), 347 nm (A = 0.432); fluorescence (H₂O): λexc = 350.1 nm, λems = 488.8 nm; IR (KBr): ν = 3403.20 (NH), 3088 (C-H aromatic), 2965 (C-H aliphatic), 1611.40 (C=O), 1449.01 (C=O) 1170.20 cm⁻¹ (SO₂N); ¹H NMR (400 MHz, DMSO-d₆) δ = 11.87 (s, 1H, NH), 8.19 (d, J = 9.12 Hz, 1H, H-4), 7.92 (d, J = 3.52 Hz, 1H, H-1), 7.87 (m, J = 6.60 Hz and J = 4.20 Hz, 2H, H-2’ and H-6’), 7.82 (d, J = 9.08 Hz, 1H, H-5), 7.72 (d, J = 3.48 Hz, 1H, H-2), 7.62 (m, 1H, 4’-H), 7.60 (m, 2H, H-3’ and H-5’), 6.45 (d, J = 1.80 Hz, 1H, H-8), 3.50 ppm (s, 3H, SO₂CH₃); ¹³C NMR (101 MHz, DMSO-d₆) δ = 41.90 (SO₂CH₃), 109.08 (C-8), 109.78 (C-1), 116.19 (C-5), 117.30 (C-9a), 117.85 (C-4), 125.95 (C-9b), 127.56 (C-2), 127.80 (C-2’ and C-6’), 129.36 (C-3’ and C-5’), 129.56 (C-1’), 130.70 (C-4’), 133.82 (C-3a), 137.70 (C-5a), 143.87 (C-7), 177.48 ppm (C-9); HRMS (ESI-MS, 140 eV): m/z[M + H⁺] calculated for C₁₈H₁₅N₂O₅⁺, 339.0803; found, 339.0798; RP-C₁₈ HPLC: tₚ = 11.7 min, 97.5%.

4.1.5.2. 3-(p-Toluenesulfonyl)-7-phenyl-6H-pyrrolo[3,2-f]quinolin-9-one (4c)

Compound 4c was prepared as described for compound 4b by reacting 2.336 g (5.07 mmol) of the appropriate phenylacrylate derivative 3c to yield 1.197 g of a slightly brown solid product. Yield: 56.9%; Rf = 0.64 (blue fluorescent spot, eluent CHCl₃/MeOH, 9:1); mp: 277.6 °C (decomposition); UV-Vis (H₂O/MeOH, 99:1): λmax (A) = 275 (A = 0.985), 347 nm (A = 0.569); fluorescence (H₂O): λexc = 350.1 nm, λems = 489.1 nm; IR (KBr): ν = 3410.50 (NH), 3080 (C-H aromatic), 2965 (C-H aliphatic), 1611.40 (C=O),
1460.01 (C===C) 1170.17 cm⁻¹ (SO₃N); ¹H NMR (400 MHz, DMSO-d₆) δ = 11.85 (s, 1H, NH), 8.29 (d, J = 9.18 Hz, 1H, H-4), 7.90 (m, 1H, H-2), 7.89 (m, AA′BB′, J = 8.54 Hz, 2H, H-2’ and H-6’), 7.88 (m, 1H, H-1), 7.83 (m, 2H, H-2″ and H-6″), 7.79 (d, J = 9.18 Hz, 1H, H-5), 7.58 (m, 3H, H-3″, H-5″ and H-4″), 7.38 (m, 2H, H-3″ and H-5″), 6.40 (s, 1H, H-8), 2.30 ppm (s, 3H, -CH₃); ¹³C NMR (75 MHz, DMSO-d₆) δ = 21.91 (-CH₃), 109.62 (C-8), 111.66 (C-1), 117.04 (C-5), 118.36 (C-9a), 118.51 (C-4), 127.54 (C-2′ and C-6′), 127.61 (C-9b), 128.29 (C-2″ and C-6″), 128.49 (C-2), 129.84 (C-3″ and C-5″), 130.59 (C-4″), 131.14 (C-3′ and C-5′), 131.20 (C-1″), 134.94 (C-4′), 135.08 (C-3a), 139.07 (C-5a), 146.49 (C-1′), 149.65 (C-7), 178.48 ppm (C-9); HRMS (ESI-MS, 140 eV): m/z [M + H⁺] calculated for C₁₂₈H₉₉N₂O₅S²⁺, 415.1226; found, 415.1299; RP-C₁₈ HPLC: tᵣ = 14.7 min, 99.7%.

4.1.5.3. 3-[(4-(Trifluoromethyl)benzene)sulfonyl]-7-phenyl-6H-pyrrolo[3,2-f]quinolin-9-one (4d)

Compound 4d was described as prepared from compound 4b by reacting 1.555 g (3.02 mmol) of the appropriate phenylacrylate derivative 3d to yield 0.414 g of a brownish solid product. Yield: 29.3%; Rf = 0.65 (blue fluorescent spot, eluent CHCl₃/MeOH, 9:1); mp: 341 °C (decomposition); UV-Vis (H₂O/MeOH, 99:1): λₘₐₓ (A) = 270 (A = 0.279), 345 nm (A = 0.169); fluorescence (H₂O): λₘₐₓ = 345.7 nm, λₑₜₙ = 491.1 nm; IR (KBr): ν = 3402.50 (NH), 3078 (C-H aromatic), 2969 (C-H aliphatic), 1608.30 (C=O), 1458.01 (C–C) 1169.37 (SO₃N), 1325.12 cm⁻¹ (C-F); ¹H NMR (DMSO-d₆): δ = 11.90 (s, 1H, NH), 8.32 (dd, J = 9.16 Hz and 0.68 Hz, 1H, H-4), 8.23 (m, J = 8.28 Hz, 2H, H-2′ and H-6′), 7.99 (m, 1H, H-2), 7.99 (m, 2H, H-3″ and H-5″), 7.93 (d, J = 3.32 Hz, 1H, H-1) 7.84 (m, 2H, H-2″ and H6″) 7.82 (d, J = 9.08 Hz, 1H, H-5), 7.59 (m, 1H, 4′), 7.59 (m, 2H, H-3′ and H-5′), 6.46 ppm (bs, 1H, H-8); ¹³C NMR (DMSO-d₆): δ = 109.30 (C-8), 112.06 (C-1), 117.23 (C-5), 117.62 (C-9a), 118.03 (C-4), 126.41 (C-9b), 126.62 (q, J = 247.23, CF₃), 127.29 (C-2), 127.65 (q, J = 32.40 Hz, C-3″ and C-5″), 127.94 (C-2″ and C-6″), 128.10 (C-2′ and C-6′), 129.48 (C-3′ and C-5′), 130.85 (C-4′), 131.24 (C-3a), 134.52 (q, J = 32.40, C-4″), 135.70 (C-1′), 139.53 (C-5a), 141.19 (C-1″), 155.26 (C-7), 180.72 ppm (C-9); HRMS (ESI-MS, 140 eV): m/z [M + H⁺] calculated for C₁₂₈H₁₈F₃N₂O₅S²⁺, 469.1027; found, 469.1040; RP-C₁₈ HPLC: tᵣ = 16.45 min, 98.5%.

4.1.5.4. 3-Benzoyl-1,2-dihydro-7-phenyl-6H-pyrrolo[3,2-f]quinolin-9-one (7a) and 1n-benzoyl-2,3-dihydro-6-phenyl-5H-pyrrolo[2,3-g]quinolin-8-one (8a)

Compounds 7a and 8a were prepared as described for compound 4b by reacting 1.100 g (2.66 mmol) of the appropriate phenylacrylate derivative 6a to yield 0.443 g of a raw powdery solid consisting of the two isomers 7a and 8a. The two desired compounds were purified by liquid column chromatography (eluent CHCl₃/MeOH, 95:5).

3-benzoyl-1,2-dihydro-7-phenyl-6H-pyrrolo[3,2-f]quinolin-9-one (7a). 0.128 g were obtained. Yield: 15.3%; Rf = 0.49 (blue fluorescent spot, eluent CHCl₃/MeOH, 9:1); mp: 322 °C (decomposition); UV–Vis (H₂O/MeOH, 99:1): λₘₐₓ (A) = 221 (0.717), 293 (0.670), 343 nm (0.320); λₑₜₙ (A) = 258 (0.226), 318 nm (0.172); fluorescence
1N-benzoyl-2,3-dihydro-6-phenyl-5H-pyrrolo[2,3-g]quinolin-8-one (8a). 0.141 g were obtained. Yield: 16.8%; Rf = 0.37 (greenish fluorescent spot, eluent CHCl₃/MeOH, 9:1); mp: 319 °C (decomposition); UV–Vis (H₂O/MeOH, 99:1): λₑₓc = 222 (0.514), 284 (0.498), 350 nm (0.320); λₑₘᵦ = 258 (0.226), 318 nm (0.172); fluorescence (H₂O), λₑₓc = 350 nm, λₑₘᵦ = 468 nm; IR (KBr): ν = 3380, 3190, 2960, 1610, 1480 cm⁻¹; ¹H NMR (400 MHz, DMSO-d₆) δ = 11.72 (s, 1H, NH), 7.84 (m, 2H, H-2" and H-6"), 7.72 (m, 7H, H-3",-4",-5", H-2' and H-6', H-4 and H-9), 7.55 (m, 3H, H-3’, -5’ and 4’), 6.35 (s, 1H, H-7), 4.01 (t, J = 8.72 Hz, 2H, H-2), 3.28 ppm (t, J = 8.71 Hz, 2H, H-1); ¹³C NMR (DMSO-d₆): δ = 26.28 (C-3), 49.24 (C-2), 116.16 (C-4), 108.24 (C-7), 118.13 (C-8a), 120.34 (C-9), 121.04 (C-3a), 126.92 (C-2’ and C-6’), 127.04 (C-2’’ and C-6’’), 127.97 (C-3’’ and C-5’’), 128.18 (C-3’ and C-5’), 129.05 (C-4’), 130.14 (C-4’’), 131.14 (C-1’), 131.16 (C-9a), 136.14 (C-1’’), 140.36 (C-4a), 153.16 (C-6), 166.18 (C···O), 177.90 ppm (C-8); HRMS (ESI-MS, 140 eV): m/z [M + H⁺] calculated for C₉₂H₁₀₅N₃O₇, 367.1447; found, 367.1439; RP-C₁₈ HPLC: tₘₐₓ = 12.02 min, 97.5%.

4.1.5.5. 3-Methanesulfonyl-1,2-dihydro-7-phenyl-6H-pyrrolo[3,2-f]quinolin-9-one (7b) and 1N-methanesulfonyl-2,3-dihydro-6-phenyl-5H-pyrrolo[2,3-g]quinolin-8-one (8b). Compounds 7b and 8b were prepared as described for compound 4b by reacting 0.488 g (1.27 mmol) of the appropriate phenylacrylate derivative 6b to yield 0.210 g of a raw powdery solid consisting of the two isomers 7b and 8b. The two desired compounds were purified by liquid column chromatography (elucent CHCl₃/MeOH, 95:5)
(C-4'), 131.00 (C-1'), 131.28 (C-9b), 134.88 (C-3a), 138.74 (C-5a), 150.45 (C-7), 178.97 ppm (C-9); HRMS (ESI-MS, 140 eV): m/z [M + H+] calculated for C₁₈H₁₇N₂O₂S²⁺, 341.1211; found, 341.1250; RP-C18 HPLC: tᵣ = 10.99 min, 96.8%.

4.1.5.5.2. 1N-methanesulfonyl-2,3-dihydro-6-phenyl-5H-pyrrolo[2,3-g]quinolin-8-one (8b)

0.068 g of a yellowish solid were obtained. Yield: 15.7; Rᶠ = 0.10 (blue fluorescent spot, eluent CHCl₃/MeOH, 95:5); mp: 317.7 °C (decomposition); UV–Vis (H₂O/MeOH, 99:1): λmax (A) = 210 (1.137), 275 (0.875), 352 nm (0.447); λmin (A) = 190 (0.104), 247 (0.650), 314 nm (0.346); fluorescence (H₂O): λexc = 350 nm, λem = 475 nm; IR (ATR ZnSe): ν = 3250, 2986, 1615, 1476, 1055 cm⁻¹; ¹H NMR (400 MHz, DMSO-d₆) δ = 11.70 (s, 1H, NH), 7.91 (s, 1H, H-9), 7.81 (m, 2H, H-2' and H-6'), 7.64 (s, 1H, H-4), 7.59 (m, 3H, H-3', H-5' and H-4'), 6.30 (bs, 1H, H-6), 4.01 (t, J = 8.26 Hz, 2H, H-2), 3.28 (t, J = 8.16 Hz, 2H, H-1), 3.03 ppm (s, 3H, SO₂CH₃); ¹³C NMR (101 MHz, DMSO-d₆) δ = 27.91 (C-1), 34.40 (SO₂CH₃), 50.49 (C-2), 107.24 (C-7), 107.35 (C-9), 115.80 (C-4), 125.03 (C-9a), 127.80 (C-2' and C-6'), 129.48 (C-3' and C-5'), 130.84 (C-4'), 134.60 (C-1'), 137.83 (C-8a), 138.51 (C-3a), 139.01 (C-4a), 149.48 (C-7), 176.66 ppm (C-8); HRMS (ESI-MS, 140 eV): m/z [M + H⁺] calculated for C₁₈H₁₇N₂O₂S²⁺, 341.1211; found, 341.1226; RP-C18 HPLC: tᵣ = 10.52 min, 98.5%.

4.1.5.6. 3-[(4-(Trifluoromethyl)benzene)sulfonyl]-1,2-dihydro-7-phenyl-6H-pyrrolo[3,2-f]quinolin-9-one (7d) and 1N-[(4-(trifluoromethyl)benzene)sulfonyl]-2,3-dihydro-6-phenyl-5H-pyrrolo[2,3-g]quinolin-8-one (8d)

Compounds 7d and 8d were prepared as described for compound 4b by reacting 1.100 g (2.12 mmol) of the appropriate phenylacrylate derivative 6d to yield 0.920 g of a raw sticky viscous tar. The tar was triturated with Et₂O and purified by liquid column chromatography (eluent CHCl₃/MeOH, 9:1) yielding 0.587 g of a powdery white solid consisting of an irreproducible mixture of the two isomers 7d and 8d. Yield: 58.6%; Rᶠ = 0.61 (eluent CHCl₃/MeOH, 9:1); mp: 301 °C (decomposition); UV–Vis (H₂O/MeOH, 99:1): 275 nm (A = 0.337), 351 nm (A = 0.258); IR (KBr): ν = 3432.80 (NH), 3078 (C-H aromatic), 2980 (C-H aliphatic), 1600 (C=C), 1498 (C=C) 1171.51 (SO₂N), 1326.36 cm⁻¹ (C-F); ¹H NMR (400 MHz, DMSO-d₆): δ = 11.69 (s, 1H, NH), 11.67 (s, 1H, NH), 8.21 (m, J = 8.94 Hz, 2H), 8.14 (s, 1H), 8.02 (d, J = 9.17 Hz, 1H), 7.98 (m, 4H), 7.92 (d, J = 9.46 Hz, 1H), 7.80 (m, J = 8.22 Hz, 2H), 7.72 (d, J = 3.03 Hz, 1H), 7.69 (m, J = 8.23 Hz, 2H), 7.66 (d, J = 8.66 Hz, 2H), 7.61 (m, 2H), 7.41 (t, J = 2.05 Hz, 1H), 7.36 (s, 1H), 6.58 (s, 1H), 6.32 (s, 1H), 4.07 (t, J = 7.81 Hz, 4H), 3.54 (t, J = 7.83 Hz, 2H), 3.15 ppm (t, J = 8.23 Hz, 2H); ¹³C NMR (101 MHz, DMSO-d₆): δ = 27.33, 29.03, 50.21, 50.87, 104.52, 105.27, 106.95, 112.24, 113.50, 118.56, 119.39, 119.65, 121.42, 121.55, 124.83 (q, J = 4.01 Hz), 126.29, 126.67 (q, J = 249.26 Hz), 126.88 (q, J = 247.28 Hz), 127.38, 127.59, 127.92, 128.09, 128.48, 128.99, 129.06, 129.37, 129.99, 130.43, 130.74, 131.60, 133.72 (q, J = 33.01 Hz), 134.02 (q, J = 32.72 Hz), 136.93, 137.25, 139.86, 150.67, 152.13, 179.12, 180.02 ppm; HRMS (ESI-MS, 140 eV): m/z [M + H⁺] calculated for C₁₈H₁₇F₃N₂O₂S²⁺, 471.1234; found, 471.1221; RP-C18 HPLC:
4.1.5.7. 3-Benzoyl-7-phenyl-6H-pyrrolo[3,2-f]quinolin-9-one (4a)

Into a two-necked 50 mL round-bottomed flask, 0.041 g (1.7 mmol, 3 eq.) of NaH, 60% dispersion in mineral oil, was placed and washed with toluene (3 × 10 mL). With stirring, a solution of 7-phenyl-3H,6H-pyrrolo[3,2-f]quinolin-9-one (9, prepared as previously reported [8]), 0.150 g (0.57 mmol, 1 eq.) in 7 mL of anhydrous DMF, was dropped into the flask. After 30 min at room temperature, a solution of benzoyl chloride, 0.2 mL (1.7 mmol, d = 1.21 g/mL, 3 eq.) in 2 mL dry DMF, was added, and the reaction mixture was stirred for 2 h. The reaction was monitored by TLC analysis (eluent CHCl₃/MeOH, 9:1). At the end of the reaction, 25 mL of water was added, and the solvent was evaporated under reduced pressure, leaving a residue, which was extracted with EtOAc (3 × 50 mL). The organic phase, washed with water, a 10% Na₂CO₃ solution, and brine was dried over anhydrous Na₂SO₄ and concentrated under vacuum to yield a crude yellow solid (0.171 g). This crude product was purified with a silica gel chromatographic column (3 × 35 cm, 230–400 mesh, CHCl₃/MeOH, 9:1), yielding 0.057 g of a pure yellowish solid. Yield: 27.2%; Rf = 0.48 (blue fluorescent spot, eluent CHCl₃/MeOH, 9:1); mp: 316.8 °C (decomposition); UV–Vis (H₂O/MeOH, 99:1): λmax (A) = 204 (0.992), 279 (1.557), 352 nm (0.281); λmin (A) = 195 (0.322), 237 (0.443), 336 nm (0.225); fluorescence (H₂O), λexc = 277 nm, λem = 460 nm; IR (KBr): ν = 3327 (NH), 1790 (C=O amide), 1678 (C=O), 1660 cm⁻¹ (C=C); ¹H NMR (400 MHz, DMSO-d₆): δ = 11.87 (s, 1H, NH), 8.60 (d, J = 9.16 Hz, 1H, H-4), 7.88 (m, 2H, H-2′ and H-6′), 7.87 (d, J = 2.96 Hz, 1H, H-1), 7.82 (m, 2H, H-2″ and H-6″), 7.80 (d, J = 8.42 Hz, 1H, H-5), 7.73 (m, J = 7.51 Hz, J = 2.12 Hz and J = 1.22 Hz, 1H, H-4″), 7.65 (m, 2H, H-3″ and H-5″), 7.61 (m, 3H, H-3′, H-5′ and H-4′), 7.51 (d, J = 3.56 Hz, 1H, H-2), 6.44 ppm (d, J = 1.16 Hz, 1H, H-8); ¹³C NMR (101 MHz, DMSO-d₆): δ = 109.13 (C-8), 110.23 (C-1), 116.18 (C-5), 117.21 (C-9a), 120.64 (C-4), 127.33 (C-9b), 127.93 (C-2′ and C-6′), 129.23 (C-3′ and C-5′), 129.37 (C-2), 129.46 (C-3′ and C-5′), 129.75 (C-2″ and C-6″), 130.79 (C-4′), 131.51 (C-1′), 132.79 (C-4′′), 134.20 (C-1′′), 137.00 (C-3a), 138.96 (C-5a), 149.20 (C-7), 168.99 (NC=O), 178.45 ppm (C-9); HRMS (ESI-MS, 140 eV): m/z [M + H⁺] calculated for C₂₄H₁₇N₂O₄⁺, 365.1385; found, 365.1382; RP-C₁₈ HPLC: tᵣ = 13.95 min, 96.5%.

4.1.5.8. N-cyclopropyl-7-phenyl-6H-pyrrolo[3,2-f]quinolin-9-one-3-carboxamide (4e)

To a stirred slurry of NaH (0.041 g of a 60% mineral oil dispersion, 1.73 mmol, 3 eq.) in THF (2 mL) at 0 °C, under N₂, was cautiously added 7-phenyl-3H,6H-pyrrolo[3,2-f]quinolin-9-one (9, 0.15 g, 0.57 mmol, 1 eq., prepared as reported [8]) previously dissolved in THF (7 mL). The reaction mixture was stirred at 0 °C for 60 min, then transferred, via cannula, to a solution of 4-nitrophenvyl chloroformate (0.140 g, 0.69 mmol, 1.2 eq.) in THF (2 mL). The resultant reaction mixture was stirred at ambient temperature for 15 h (TLC, eluent n-Hex/EtOAc, 2:1), prior to removal of the solvent by concentration in vacuo. The residue obtained was
suspended in EtOAc (100 mL), then filtered and washed with EtOAc and Et₂O to yield 0.255 g (99.0%) of a pale yellow solid. The resulting activated N-(4-nitro)phenyl-7-phenyl-6H-pyrrolo[3,2-f]quinolin-9-1one-3-carboxamide was immediately used as follows: a 2.0 M solution of cyclopropylamine (0.382 mL, 5.48 mmol, d = 0.824 g/mL, 8 eq.) in THF (2.75 mL) was added to a solution of the activated carbamate (0.255 g, 0.60 mmol, 1eq.) in THF (5 mL). The resultant reaction mixture was stirred at ambient temperature for 4 h and monitored by TLC (eluent CHCl₃/MeOH, 85:15) prior to removal of the solvent by concentration to dryness, in vacuo. The residue obtained was partitioned between EtOAc (100 mL) and H₂O (100 mL). The layers were separated, and the aqueous phase was extracted with EtOAc (2 × 30 mL). The combined organic extract was washed with saturated NaHCO₃ (100 mL) and brine, dried over NaSO₄ and concentrated in vacuo to give a yellow solid that was suspended in Et₂O (35 mL), filtered and washed with Et₂O (2 × 20 mL) to give 0.062 g of a yellowish solid. Yield: 30.1%. Rf = 0.29 (eluent CHCl₃/MeOH, 85:15); mp: 314 °C (decomposition); UV–Vis (H₂O/MeOH, 99:1): λ_max (A) = 338 (0.081), 270 (0.162), 204 nm (0.169); λ_min (A) = 312 nm (0.046), 246 nm (0.079); fluorescence (H₂O): λ_exc = 215.00 nm, λ_em = 430.00 nm; IR (KBr): ν = 3311.56 (NH), 1618 (C=O) cm⁻¹; ¹H NMR (400 MHz, DMSO-d₆): δ = 11.80 (s, 1H, NH), 8.56 (d, J = 9.12 Hz, 1H, H-4), 8.40 (d, J = 2.91 Hz, 1H, C=O), 7.91 (d, J = 3.54 Hz, 1H, H-2), 7.85 (m, 2H, H-2’ and H-6’), 7.75 (d, J = 2.67 Hz, 1H, H-1), 7.68 (d, J = 9.15 Hz, 1H, H-5), 7.58 (m, 3H, H-3’, H-5’ and H-4’), 6.41 (bs, 1H, H-8), 2.81 (m, J = 3.24 Hz, 1H, NH-CH), 0.80–0.60 ppm (m, 4H, -CH₂CH₂-); ¹³C NMR (101 MHz, DMSO-d₆): δ = 6.31 (CH₂CH₂-), 23.82 (NH-CH), 108.05 (C-1), 108.93 (C-8), 115.15 (C-5), 117.86 (C-9a), 120.14 (C-4), 124.15 (C-9b), 125.81 (C-2), 127.86 (C-2’ and C-6’), 129.43 (C-3’ and C-5’), 130.59 (C-4’), 130.85 (C-1’), 131.39 (C-3a), 134.85 (C-5a), 153.12 (C-7), 165.21 (NC=ONH), 178.43 ppm (C-9); HRMS (ESI-MS, 140 eV): m/z [M + H⁺] calculated for C₁₂H₁₆N₂O₂⁺, 344.1999; found, 344.1994; RP-C₁₈ HPLC: t_R = 10.93 min, 98.5%.

4.2. Biological assays

4.2.1. Cell growth conditions and antiproliferative assay

Human T-leukemia (CCRF-CEM and Jurkat), human B-leukemia (RS4; 11, SEM) cells and human myeloid leukemia (HL-60, THP-1, MV4; 11) cells, were grown in RPMI-1640 medium (Gibco, Milano, Italy). Breast adenocarcinoma (MCF-7), human cervix carcinoma (HeLa), non small cell lung adenocarcinoma (A549) and human colon adenocarcinoma (HT-29) cells were grown in DMEM medium (Gibco, Milano, Italy), all supplemented with 115 units/mL penicillin G (Gibco, Milano, Italy), 115 µg/mL streptomycin (Invitrogen, Milano, Italy), and 10% fetal bovine serum (Invitrogen, Milano, Italy). Stock solutions (10 mM) of the different compounds were obtained by dissolving them in DMSO. Individual wells of a 96-well tissue culture microtiter plate were inoculated with 100 µL of complete medium containing 8 × 10⁵ cells. The plates were incubated at 37 °C in a humidified 5% CO₂ incubator for 18 h prior to the experiments. After medium removal, 100 µL of fresh medium containing the test compound at different concentrations was added to each well in
triplicate and incubated at 37 °C for 72 h. Cell viability was assayed by MTT test as previously described [26]. The GI<sub>50</sub> was defined as the compound concentration required to inhibit cell proliferation by 50%. Peripheral blood lymphocytes (PBL) from healthy donors were obtained by separation on Lymphoprep (Fresenius KABI Norge AS) gradient. After extensive washing, cells were resuspended (1.0 × 10<sup>6</sup> cells/mL) in RPMI-1640 with 10% fetal bovine serum and incubated overnight. For cytotoxicity evaluations in proliferating PBL cultures, non-adherent cells were resuspended at 5 × 10<sup>5</sup> cells/mL in growth medium, containing 2.5 μg/mL PHA (Irvine Scientific). Different concentrations of the test compounds were added, and viability was determined 72 h later by the MTT test. For cytotoxicity evaluations in resting PBL cultures, non-adherent cells were resuspended (5 × 10<sup>5</sup> cells/mL) and treated for 72 h with the test compounds, as described above.

4.2.2. Effects on tubulin polymerization and on colchicine binding to tubulin

To evaluate the effect of the compounds on tubulin assembly in vitro [15], varying concentrations of compounds were preincubated with 10 μM bovine brain tubulin in glutamate buffer at 30 °C and then cooled to 0 °C. After addition of 0.4 mM GTP (final concentration), the mixtures were transferred to 0 °C cuvettes in a recording spectrophotometer and warmed to 30 °C. Tubulin assembly was followed turbidimetrically at 350 nm. The IC<sub>50</sub> was defined as the compound concentration that inhibited the extent of assembly by 50% after a 20 min incubation. The ability of the test compounds to inhibit colchicine binding to tubulin was measured as described [16], except that the reaction mixtures contained 1 μM tubulin, 5 μM [³H]colchicine and 5 μM test compound.

4.2.3. Molecular modeling

Compounds in Table 1 were built and their partial charges calculated after semi-empirical (PM6) energy minimization using the MOE2015 [27] program. Fourteen crystallographic structures were selected to perform docking studies (see SI_Table 1). Only the ligands occupying the colchicine binding site and the protein chain in the proximity of 4.5 Å were considered and subjected to the structure preparation tool of MOE 2015. Finally, Protonate 3D tool was used to assign the ionic state of each complex [28]. To identify the more appropriate docking protocol for the eleven complexes, we performed a self-docking benchmark using DockBench 1.01, a tool that compared the performance of 14 different posing/scoring protocols [29]. The active site was defined using a radius of 12 Å from the center of mass of the co-crystallized ligand. Each ligand was docked 20 times. All synthesized compounds were docked using GOLD using PLP [30], using the virtual screening tool of DockBench adopting the parameters already used in the benchmark study. Finally, the obtained conformations were rescored with the dock_pKi MOE function. The similarity studies were carried out with vROCS considering the Tversky coefficient [31].
To facilitate the visualization and analysis of data obtained from the docking simulations, we implemented an in-house tool, named MMsDocking video maker, for the automated production of a video that shows the most relevant docking data, such as docking poses, per residue IEhyd and IEele data, experimental binding data and scoring values. Videos were mounted using MEncoder [32], starting from images obtained with the following procedure: the heat maps in the background were drawn with GNUPLOT 4.6 [33] starting from per residue IEhyd and IEele data computed with MOE. Two dimensional depictions of compounds were generated using the open-source cheminformatics toolkit RDKit [34]. Representations of docking poses within the binding site were constructed using CHIMERA [35].

4.2.4. Flow cytometric analysis of cell cycle distribution

5 × 10⁵ HeLa or Jurkat cells were treated with different concentrations of the test compounds for 24 h. After the incubation period, the cells were collected, centrifuged, and fixed with ice-cold ethanol (70%). The cells were then treated with lysis buffer containing RNase A and 0.1% Triton X-100 and stained with PI. Samples were analyzed on a Cytomic FC500 flow cytometer (Beckman Coulter). DNA histograms were analyzed using MultiCycle for Windows (Phoenix Flow Systems).

4.2.5. Apoptosis assay

Cell death was determined by flow cytometry of cells double stained with annexin V/FITC and PI. The Coulter Cytomics FC500 (Beckman Coulter) was used to measure the surface exposure of PS on apoptotic cells according to the manufacturer’s instructions (Annexin-V Fluos, Roche Diagnostics).

4.2.6. Analysis of mitochondrial potential and reactive oxygen species (ROS)

The mitochondrial membrane potential was measured with the lipophilic cation JC-1 (Molecular Probes, Eugene, OR, USA), while the production of ROS was followed by flow cytometry using the fluorescent dye H₂DCFDA (Molecular Probes), as previously described [11].

4.2.7. Evaluation of the metabolic stability of compound 4a in human liver microsomes

4.2.7.1. Incubation procedure

Compound 4a (final concentration, 10 μM) was incubated in a medium (final volume, 0.2 mL) containing 0.1 M KH₂PO₄ (pH 7.4) and 1.0 mg/mL of pooled mixed-gender human liver microsomes (Xenotech LLC, Lenexa, USA; HLMs), in the absence or presence of 1 mM NADPH (Sigma-Aldrich). Control incubations were performed in the absence of both HLMs and NADPH (buffer only-incubations). The reactions were started by adding the microsomes following a 3-min thermal equilibration at 37 °C, conducted at 37 °C for different time periods (i.e. 0, 15, 30 and 60 min), and terminated by adding 0.1 mL of ice-cold acetonitrile. Samples were
then centrifuged (4 °C) at 20,000g for 10 min, and aliquots of the supernatants were analyzed by HPLC with fluorescence detection, as described below.

4.2.7.2. HPLC analysis

The chromatographic system consisted of a Hewlett-Packard 1100 HPLC system (Agilent Technologies Inc., formerly Hewlett-Packard, Palo Alto, USA) equipped with a degasser, a quaternary pump, an autosampler, a column oven, and a fluorescence detector; chromatographic data were collected and integrated using the Agilent ChemStation software. Chromatographic conditions were as follows: column, Agilent Zorbax SB C18 (4.6 × 75 mm, 3.5 μm); mobile phase, 0.1% HCOOH in H2O (solvent A) and 0.1% HCOOH in acetonitrile (solvent B); elution program, isocratic elution with 95% solvent A for 2 min, linear gradient from 5 to 40% solvent B in 8 min, followed by a further linear gradient from 40 to 60% solvent B in 2 min, and an isocratic elution with 60% solvent B for 7 min; post-run time, 5 min; flow rate, 1.0 mL/min; injection volume, 50 μL; column temperature, 30 °C; detection, fluorescence (excitation wavelength, 344 nm; emission wavelength, 493 nm). Under the above conditions, the retention time of 4a was 13.4 min. Metabolic stability of 4a, expressed as percent of compound remaining, was calculated by comparing the corresponding chromatographic peak area at each time point relative to that at time 0 min.
References


30. GOLD suite, version 5.2. Cambridge Crystallographic Data Centre: 12 Union Road, Cambridge CB2 1EZ, UK. http://www.ccdc.cam.ac.uk

The role of 5-arylalkylamino- and 5-piperazino-moieties on the 7-aminopyrazolo[4,3-d]pyrimidine core in affecting adenosine A₁ and A₂A receptor affinity and selectivity profiles

Lucia Squarcialupi, Marco Betti, Daniela Catarzi, Flavia Varano, Matteo Falsini, Annalisa Ravani, Silvia Pasquini, Fabrizio Vincenzi, Veronica Salmaso, Mattia Sturlese, Katia Varani, Stefano Moro & Vittoria Colotta*


Abstract

New 7-amino-2-phenylpyrazolo[4,3-d]pyrimidine derivatives, substituted at the 5-position with aryl(alkyl)amino- and 4-substituted-piperazin-1-yl- moieties, were synthesized with the aim of targeting human (h) adenosine A₁ and/or A₂A receptor subtypes. On the whole, the novel derivatives 1–24 shared scarce or no affinities for the off-target hA₂B and hA₃ ARs. The 5-(4-hydroxyphenethylamino)- derivative 12 showed both good affinity (Ki = 150 nM) and the best selectivity for the hA₂A AR while the 5-benzylamino-substituted 5 displayed the best combined hA₂A (Ki = 123 nM) and A₁ AR affinity (Ki = 25 nM). The 5-phenethylamino moiety (compound 6) achieved nanomolar affinity (Ki = 11 nM) and good selectivity for the hA₁ AR. The 5-(N₄-substituted-piperazin-1-yl) derivatives 15–24 bind the hA₁ AR subtype with affinities falling in the high nanomolar range. A structure-based molecular modeling study was conducted to rationalize the experimental binding data from a molecular point of view using both molecular docking studies and Interaction Energy Fingerprints (IEFs) analysis.

1. Introduction

Adenosine receptors (ARs) are classified as A₁, A₂A, A₂B and A₃ subtypes [1, 2] and typically inhibit (A₁ and A₃) or activate (A₂A and A₂B) adenylyl cyclase. A₁ receptor is highly expressed in brain areas, such as the hippocampus and prefrontal cortex [3, 4], implicated in the control of emotions and cognition functions. Therefore, A₁ AR antagonists are investigated as therapeutic agents for mental dysfunctions, such as dementia and anxiety [3–5]. The A₂A AR subtype is present in the brain with the highest concentration in the striatum, nucleus accumbens, hippocampus and cortex, and its blockade has proven to be effective in neurodegenerative pathologies such as Parkinson’s disease (PD) [6–8]. The A₂A AR antagonist istradefylline has been recently approved for marketing in Japan for the treatment of PD patients [9]. In preclinical studies, dual A₁/A₂A antagonists have also turned out to be useful for PD therapy because they reduce both motor (A₂A) and cognitive (A₁) impairment associated with this pathology [5, 10–12].
Recent studies have highlighted new therapeutic applications of A2A AR antagonists [12]. If topically administered, they diminish scar size and promote restoration of skin integrity [13]. A2A AR antagonists have also demonstrated efficacy in enhancing immunologic response, especially by markedly improving anti-tumor immunity in mouse models, thus promoting tumor regression. A2A AR antagonists have been shown to improve the effect of tumor vaccines during T-cell activation, and may work in concert with other immune checkpoint inhibitors in cancer immunotherapy [12, 14].

In our laboratory, much research has been addressed to the study of AR antagonists belonging to different classes [15–26], including the 2-arylpyrazolo[4,3-d]pyrimidine derivatives [20,22,24,26] which display a broad range of affinity for the various AR subtypes, depending on the nature of the substituents at the 5- and 7-positions of the bicyclic scaffold. One recent study aimed at targeting the A1 and A2A ARs highlighted that the presence of a free 7-amino group, combined with a benzyl or, even better, a 3-phenylpropyl chain at the 5-position (Figure 1, compounds A and C) shifted affinity toward these two AR subtypes [24].

In our laboratory, much research has been addressed to the study of AR antagonists belonging to different classes [15–26], including the 2-arylpyrazolo[4,3-d]pyrimidine derivatives [20,22,24,26] which display a broad range of affinity for the various AR subtypes, depending on the nature of the substituents at the 5- and 7-positions of the bicyclic scaffold. One recent study aimed at targeting the A1 and A2A ARs highlighted that the presence of a free 7-amino group, combined with a benzyl or, even better, a 3-phenylpropyl chain at the 5-position (Figure 1, compounds A and C) shifted affinity toward these two AR subtypes [24].

Hence, to further explore the structural requirements for addressing affinity toward the A1 and/or A2A ARs, various aryl(alkyl)amino- and 4-substituted-piperazin-1-yl- moieties were appended at the 5-position of the scaffold (compounds 1–24, Figure 2). These substituents were selected since they are a common feature of potent A1 and/or A2A AR antagonists structurally correlated to our pyrazolopyrimidine derivatives [12,27,28] (such as the triazolotriazines ZM-241385 and D, Figure 1). The pyrazolopyrimidines 1–24 were tested in
binding assays to evaluate their affinity at cloned hA1, hA2A and hA3 ARs, stably expressed in CHO cells. Compounds were also tested at the hA2B receptor by measuring their inhibitory effects on NECA-stimulated cAMP levels in CHO cells.

A structure-based molecular modeling study was performed on the new derivatives to rationalize the experimental binding data from a molecular point of view, using molecular docking studies in tandem with Interaction Energy Fingerprints (IEFs) analysis.

2. Chemistry

The 7-amino-pyrazolo[4,3-d]pyrimidine derivatives 1–14, bearing an arylalkylamino moiety at the 5-position, were obtained as displayed in Scheme 1.

Both the 1-phenyl (26) and 1-alkyl substituted pyrazoles (27, 28) were synthesized from a common starting compound: the readily available 4-nitro-1H-pyrazole-3-carbonitrile 25 [26] which was a good substrate for both regioselective N-alkylation and Narylation. The latter was achieved by a cross-coupling reaction with phenylboronic acid in the presence of cupric acetate and activated molecular sieves. The 1-phenyl-pyrazole derivative 26 was thus prepared with higher yield than those previously obtained in our laboratory through another synthetic pathway [22]. The 1-methyl- and 1-benzyl-pyrazoles 27 and 28 were prepared from compound 25 as already described [26]. The 4-nitropyrazolo-3-carbonitriles 26–28 were transformed into
the corresponding 4-amino derivatives 29–31 [26] by reduction with cyclohexene and Pd/C, under microwave-assisted conditions. Reaction of compounds 29–31 with isothiocyanates in anhydrous DMF yielded the corresponding N-(1-substituted-3-cyano-pyrazol-4-yl)thiourea derivatives 32–42. Phenyl-, 4-methoxyphenyl-, 2,4-dichlorophenyl and benzyl-isothiocyanates were commercially available, the others were synthesized as previously reported, i.e. allowing the corresponding arylalkylamines to react with CS2, in 30% hydrogen peroxide aqueous solution (phenylethyl-, phenylpropyl- and 3,4-dimethoxyphenyl-isothiocyanates) [29,30] or with thiophosgene and potassium carbonate, in CH2Cl2 under nitrogen atmosphere (4-methoxyphenylisothiocyanate) [31].

Compounds 32–42 were reacted with iodomethane in anhydrous DMF to give the corresponding S-methylisothiourea derivatives 43–53 which were cyclized to the desired 7-amino-5-arylalkylamino-pyrazolo[4,3-d]pyrimidines 1–3, 5–12 by reaction with ammonium chloride in formamide, under microwave irradiation. The methoxy-substituted derivatives 2, 11 and 12 were transformed into the corresponding hydroxy derivatives 4, 13 and 14 by treatment with BBr3 in anhydrous CH2Cl2.

The 7-amino-pyrazolo[4,3-d]pyrimidine derivatives 15–22, bearing N-substituted-piperazine moieties at the 5-position, were obtained utilizing the synthetic route as described in Scheme 2.

Scheme 2 Reagents and conditions: (a) N,N-dimethylaniline, POCl3, 150 °C, mw; (b) 33% aqueous NH3, 100 °C, mw; (c) benzylationmine, ethyldiisopropylamine, tert-butanol, 200 °C, mw; (d) ethyldiisopropylamine, N-methylpyrrolidone, 130–150 °C, mw; (e) compound 20, LiAlH4, anhydrous THF, room temperature.

Allowing the 1-phenylpyrazolo[4,3-d]pyrimidine-5,7-dione 54 [20] to react with phosphorus oxychloride and N,N-dimethylaniline under microwave irradiation, the 5,7-dichloro-derivative 55 was prepared, which was reacted with 33% aqueous ammonia solution under microwave irradiation at 100°C to give the 7-amino-5-
chloro-pyrazolopyrimidine 56 as the only regioisomer. The 7-amino structure of 56 was expected on the basis of the well-known different mobility of the two chlorine atoms in the pyrimidine ring, also condensed with diverse heterocyclic systems [32–34]. To confirm the structure, derivative 56 was treated with benzylamine in tert-butanol, in the presence of diisopropylethylamine, and the 7-amino-5-benzylaminopyrazole derivative 5, already synthesized through the unambiguous synthesis as depicted in Scheme 1, was obtained. This reaction was carried out under prolonged microwave irradiation (about 1 h at 200°C) but conversion of derivative 56 into 5 occurred with unsatisfactory yields. The 1H NMR spectrum of the crude reaction (data not shown) displayed the presence of both the 5-benzylamino derivative 5 and the starting material 56 (ratio about 3.5:1), besides degradation compounds, thus indicating the poor reactivity of the C5 atom toward the primary benzyl ammine group. Instead, microwave-assisted reaction of the 5-chloro derivative 56 with the N-substituted piperazines 57–63, in N-methylpyrrolidone and in the presence of diisopropylethylamine, proceeded to completion, thus giving the desired pyrazolopyrimidine derivatives 15–20 with good yields (48–85%). The piperazine derivatives 57, 58, 62 and 63 were commercially available, while derivatives 59 and 61 were prepared as previously described [35, 36]. The piperazine derivative 60 was synthesized starting from the reductive alkylation of N-Boc-piperazine 63 with 2,4,6-trifluorobenzaldehyde and triacetoxy sodium borohydride. The obtained tert-butyl 4-(2,4,6-trifluorobenzyl)piperazine-1-carboxylate was hydrolyzed with trifluoroacetic acid to give the 1-(2,4,6-trifluorobenzyl)piperazine 60, isolated as trifluoroacetate salt.

Reduction of the 2-furoyl carbonyl group of compound 20 with LiAlH4 in anhydrous THF provided derivative 22. Finally, the pyrazolopyrimidines 23–24, bearing an acyl moiety on the piperazine nitrogen, were synthesized as depicted in Scheme 3. Treatment of the N-Boc derivative 21 with trifluoroacetic acid furnished compound 64 which was reacted with suitable acyl chlorides, in the presence of triethylamine in anhydrous tetrahydrofuran, to provide the desired 23–24.
3. Results and discussion

3.1 Structure–affinity relationship studies

The results of binding experiments and cAMP assays carried out on the new 5-substituted-pyrazolopyrimidines 1–14 and 15–24 are displayed, respectively, in Tables 1 and 2. Table 1 also includes the affinity data of the pyrazolopyrimidines A–C and of ZM-241385 reported as references.

Table 1 Binding affinity at hA1, hA2A and hA3 Rs and potencies at hA2B Rs.

<table>
<thead>
<tr>
<th></th>
<th>R</th>
<th>R2</th>
<th>Binding experimentsa</th>
<th>cAMP assays</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Ki (nM) or I%</td>
<td>IC50 (nM) or I%</td>
</tr>
<tr>
<td>1</td>
<td>Ph</td>
<td>Ph</td>
<td>67±5</td>
<td>412±37</td>
</tr>
<tr>
<td>2</td>
<td>4-OMe-C6H4</td>
<td>Ph</td>
<td>33%</td>
<td>8%</td>
</tr>
<tr>
<td>3</td>
<td>2,4-diIC6H3</td>
<td>Ph</td>
<td>1%</td>
<td>1%</td>
</tr>
<tr>
<td>4</td>
<td>4-OH-C6H4</td>
<td>Ph</td>
<td>1%</td>
<td>1%</td>
</tr>
<tr>
<td>5</td>
<td>Ph-CH2</td>
<td>Ph</td>
<td>25±3</td>
<td>123±11</td>
</tr>
<tr>
<td>6</td>
<td>Ph-CH2CH2</td>
<td>Ph</td>
<td>11.5±1.2</td>
<td>40%</td>
</tr>
<tr>
<td>7</td>
<td>Ph-CH2CH2</td>
<td>Ph</td>
<td>785±72</td>
<td>29%</td>
</tr>
<tr>
<td>8</td>
<td>Ph</td>
<td>Me</td>
<td>1%</td>
<td>4%</td>
</tr>
<tr>
<td>9</td>
<td>Ph-CH2</td>
<td>Me</td>
<td>9%</td>
<td>1%</td>
</tr>
<tr>
<td>10</td>
<td>Ph-CH2</td>
<td>CH2Ph</td>
<td>19%</td>
<td>1%</td>
</tr>
<tr>
<td>11</td>
<td>4-OMe-C6H4</td>
<td>CH2CH2</td>
<td>19%</td>
<td>1%</td>
</tr>
<tr>
<td>12</td>
<td>4-OH-C6H4</td>
<td>CH2CH2</td>
<td>27%</td>
<td>1%</td>
</tr>
<tr>
<td>13</td>
<td>3,4-diOMe-C6H3-(CH2)2</td>
<td>Ph</td>
<td>153±11</td>
<td>26%</td>
</tr>
<tr>
<td>14</td>
<td>3,4-diOH-C6H3-(CH2)2</td>
<td>Ph</td>
<td>415±39</td>
<td>189±18</td>
</tr>
<tr>
<td>A</td>
<td>Ph</td>
<td>-</td>
<td>150±12</td>
<td>110±10</td>
</tr>
<tr>
<td>B</td>
<td>Ph-CH2</td>
<td>-</td>
<td>15%</td>
<td>35%</td>
</tr>
<tr>
<td>C</td>
<td>Ph-CH2CH2</td>
<td>-</td>
<td>5.31±0.42</td>
<td>55±5</td>
</tr>
</tbody>
</table>

A Ki values are means ± SEM of four separate assays each performed in duplicate. Percentage of inhibition (%) are determined at 1 μM concentration of the tested compounds.

b Displacement of specific [3H]DPCPX competition binding assays to hA1 CHO cells.

c Displacement of specific [3H]ZM241385 competition binding to hA2A CHO cells.

d Displacement of specific [125I]AB-MECA competition binding to hA3 CHO cells.

e cAMP experiments in hA2B CHO cells, stimulated by 200 nM NECA. Percentage of inhibition (%) are determined at 1 μM concentration of the tested compounds.

f Ref. 24.

g Ref. 5.

h Ki value obtained from binding experiments at recombinant hA2B.
Table 2 Binding affinity at hA₁, hA₂A and hA₃ ARs and potencies at hA₂B ARs.

<table>
<thead>
<tr>
<th>R</th>
<th>Ki (nM) hA₁</th>
<th>Ki (nM) hA₂A</th>
<th>Ki (nM) hA₃</th>
<th>Ki (nM) hA₂B</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>647±53</td>
<td>20%</td>
<td>20%</td>
<td>1%</td>
</tr>
<tr>
<td>16</td>
<td>162±14</td>
<td>1%</td>
<td>1%</td>
<td>1%</td>
</tr>
<tr>
<td>17</td>
<td>518±42</td>
<td>40%</td>
<td>40%</td>
<td>1%</td>
</tr>
<tr>
<td>18</td>
<td>204±18</td>
<td>34%</td>
<td>39%</td>
<td>1%</td>
</tr>
<tr>
<td>19</td>
<td>193±17</td>
<td>29%</td>
<td>1%</td>
<td>2%</td>
</tr>
<tr>
<td>20</td>
<td>580±47</td>
<td>16%</td>
<td>13%</td>
<td>2%</td>
</tr>
<tr>
<td>21</td>
<td>615±49</td>
<td>33%</td>
<td>21%</td>
<td>1%</td>
</tr>
<tr>
<td>22</td>
<td>92±8</td>
<td>38%</td>
<td>5%</td>
<td>16%</td>
</tr>
<tr>
<td>23</td>
<td>29%</td>
<td>2%</td>
<td>1%</td>
<td>2%</td>
</tr>
<tr>
<td>24</td>
<td>429±36</td>
<td>3%</td>
<td>1%</td>
<td>3%</td>
</tr>
</tbody>
</table>

a Ki values are means ± SEM of four separate assays each performed in duplicate. Percentage of inhibition (I%) are determined at 1 μM concentration of the tested compounds.

b Displacement of specific [³H]DPCPX competition binding assays to hA₁CHO cells.

c Displacement of specific [³H]ZM241385 competition binding to hA₂A CHO cells.

d Displacement of specific [³H]AB-MECA competition binding to hA₃CHO cells.

e cAMP experiments in hA₂B CHO cells, stimulated by 200 nM NECA. Percentage of inhibition (%) are determined at 1 μM concentration of the tested compounds.

As expected, the new derivatives 1–24 shared scarce or no affinities for the off-target hA₂B and hA₃ ARs, except the 5-anilino and 5-benzylamino derivatives 1–3 and 5, respectively which displayed nanomolar affinity for the hA₃ subtype (Ki = 13–61 nM). In particular, compounds 2 and 3 are worth noting, being also highly hA₃ selective.

Since the purpose of the work was to target hA₁ and hA₂A ARs, SAR discussion was focused on hA₁ and hA₂A binding data. In this respect, results of some interest have been obtained from the 5-arylalkylamino-pyrazolopyrimidines 1–14. In fact, compound 12 showed both good affinity and the best selectivity for the hA₂A AR, while compounds 1, 5, 13 and 14 were able to bind both the hA₁ and hA₂A ARs. Moreover, a derivative having nanomolar affinity and high selectivity for the hA₁ AR subtype was identified (compound 6).

The new 5-phenyl(alkyl)amino derivatives 1, 5 and 6 were designed as analogs of our previously reported antagonists 5-phenyl(alkyl) derivatives A, B and C [24] whose methylene linker at the 5-position of the bicyclic core was replaced with an NH. This modification, suggested by the structure of potent A₂A antagonists bearing arylalkylamino moieties as key substituents [12], was thought to change the flexibility of the 5-lateral chain
and, hopefully, to increase the affinity for the targeted ARs. Actually, the NH linker enhanced the hA1 AR affinity (compare 1 and 5 to A and B, respectively) or maintained it in the nanomolar range (compare 6 to C). Instead, the hA2A AR binding was ameliorated in one case, i.e. the 5-benzylamino derivative 5 which was more active than the corresponding phenylalkyl-derivative B.

Analyzing the hA1 and hA2A AR binding data of 1–6 in detail, it can be observed that 5-phenylamino derivative 1 binds to the hA2A and hA1 AR subtypes with scarce (Ki = 412 nM) and good affinity (Ki = 67 nM), respectively. Introduction of either a 4-methoxy group or 2,4-dichloro substituents on the 5-aniline moiety of 1 (compounds 2 and 3) dropped affinity for hA2 and hA2A ARs. Instead, the presence of a 4-hydroxy residue (compound 4) reduced the hA2A affinity while conserving some ability to bind the hA1 receptor (Ki = 481 nM). Homologation of the 5-phenylamino moiety (derivative 1) to the 5-benzylamino group (derivative 5) produced some improvement in the binding activity at both hA1 (Ki = 25 nM) and hA2A ARs (Ki = 123 nM). Quite unexpectedly, homologation of the alkyl chain of compound 5, to obtain the 5-phenethylamino and the 5-phenylpropylamino derivatives 6 and 7, caused a drastic reduction of the hA2A AR affinity and, in the former, it increased the hA1 one, thus affording a selective hA1 receptor ligand (Ki = 11.5 nM).

Replacement of the 2-phenyl group of derivatives 1 and 5 with a methyl residue, to give compounds 8 and 9, was performed to verify whether a reduction in the volume of the molecule might permit a better accommodation inside the recognition site of the targeted hARs. This modification, instead, annulled the capability to bind the target hARs. The same detrimental effect was obtained when the 2-phenyl ring of 5 was replaced with the more flexible benzyl moiety (derivative 10).

Insertion of the para hydroxy substituent on the 5-phenethylamino moiety of derivative 6, to give compound 12, was based on the structure of the well-known potent and selective hA2A AR antagonist ZM-241385 [5,12] (Figure 1). Accordingly, we also thought it would be interesting to evaluate the 3,4-dihydroxy substitution (compound 14), as well as the 4-methoxy- and the 3,4-dimethoxysubstituents (derivatives 11 and 13). As expected, the presence of the 4-hydroxy group was able to shift the affinity toward the hA2A AR. In fact, the 4-hydroxy-substituted derivative 12 showed good hA2A affinity (Ki = 150 nM) and the best selectivity among all the ligands reported here. In contrast, reversed selectivity was demonstrated by the 4-methoxy derivative 11, which displayed good affinity for the hA1 AR but not for the hA2A subtype. Instead, the 3,4-dimethoxy substituted derivative 13 bound both hA1 and hA2A receptors and also the 3,4-dihydroxy derivative 14 showed quite good affinity for both the receptors, but especially for the hA1 one.

Finally, to further explore the SARs in this class of AR ligands, various N-substituted piperazine moieties were appended at the 5-position (derivatives 15–24, Table 2), in accordance with the structure of known potent and selective hA2A AR antagonists [27, 28]. In contrast to our expectations, none of the 5-(N4-R-piperazin-1-yl) derivatives 15–24 were able to bind effectively the A2A AR while they possessed affinity for the hA1 AR
subtype, falling in the high nanomolar range. The most active compounds proved to be 22 (Ki = 92 nM) and 16 (Ki = 162 nM) which bear, respectively, the (2-furyl)-methyl and 2-benzyl pendant on the N4-piperazine moiety. Introduction of halogen atoms on the benzyl moiety of 16 left almost unchanged the hA1 AR affinity (compounds 18 and 19) while elongation of the benzyl chain decreased it (compound 17). Also the other substituents evaluated on the piperazine ring, i.e. acyl moieties (derivatives 20, 23, 24) and the tert-butoxycarbonyl group (derivative 21) did not ameliorate the hA1 AR affinities.

3.2 Molecular modeling studies

A structure-based molecular modeling study was conducted to rationalize the experimental binding data from a molecular point of view. Minor attention was devoted to the hA2B AR subtype, since no significant binding affinity has been estimated for any of the compounds under investigation. Docking was performed on hA1, hA2A and hA3 AR subtypes, and the resulting poses were evaluated according to the van der Waals and electrostatic interactions, as previously reported [37,38] and described in detail in the “Experimental” section. Positive electrostatic and van der Waals values were used as filters to reject unfavorable docking poses. One pose for each ligand was selected on the basis of the Interaction Energy Fingerprints (IEFs) and by visual inspection.

An overview of the most favorable poses of all compounds on hA1, hA2A and hA3 ARs is reported in video SM1-SM2-SM3, included in Supplementary Material. The heat map depicted in the background reports the electrostatic and hydrophobic contributions of the residues mainly involved in binding (“ele” and “hyd” labels identify the major contribution type of the residue) by a colorimetric scale going from blue to green for negative to positive values. These crucial residues are mainly positioned on the superior half of TM6 and TM7 and EL2, and the overall binding modes of the compounds under examination are very consistent among them. Here, we describe in detail the poses of compound 1 as an example, because of its high binding affinity for all three AR subtypes taken into consideration (Ki = 67 nM for hA1, Ki = 412 nM for hA2A and Ki = 13 nM for hA3).

With regard to the hA1 AR, Glu172 (EL2) and Asn254 (6.55), represented by blue bars on electrostatic IEFs (Figure 3, panel A on the left), emerge as important residues for electrostatic contribution, together with a slight contribution of Trp247 (6.48) and His251 (6.52). Asn254 (6.55) and Glu172 (EL2) are engaged in a three hydrogen bond pattern with N1 of pyrazole and with the exocyclic amine group at position 7 of compound 1, as shown in Figure 4, panel A. The aromatic pyrazolopyrimidine scaffold is involved in a π–π stacking interaction with Phe171 (EL2), which is one of the residues appearing to have the strongest hydrophobic interaction on the hydrophobic IEFs (green bars in Figure 3, panel A on the right). Val87 (3.32), Leu88 (3.33),
Trp247 (6.48), Leu250 (6.51), Tyr271 (7.36) and Ile274 (7.39) are also involved in significant hydrophobic contacts, with Val87 (3.32), Leu88 (3.33), Trp247 (6.48) defining the bottom of the binding pocket.

The residues involved in binding at hA2A AR are positioned equivalently to those just described for the hA1 subtype. Glu169 (EL2) and Asn253 (6.55) are involved in hydrogen bonds and Phe168 (EL2) makes a π–π stacking interaction, as can be seen in Figure 4, panel B. Trp246 (6.48) and His250 (6.52), together with Glu169 and Asn253, give stabilizing electrostatic contributions to the binding of 1, while Leu85 (3.33), Leu167 (EL2), Phe168 (EL2), Trp246 (6.48), Leu249 (6.51), Tyr271 (7.36) are interested by hydrophobic contacts (Figure 3, panel B).

Fig. 3 Interaction Energy Fingerprints (IEFs) comparison between compound 1 and compound ZM-241385 used as reference. Panels A, B and C report the comparison analysis for hA1, hA2A and hA3 receptor subtypes, respectively. On the left side is shown the electrostatic contribution comparison, while on the right the hydrophobic one. In each subsection, the IEFs of compound 1 are shown above the IEFs of the reference ZM-241385.
The binding of compound 1 to the hA3 subtype mainly engages Trp243 (6.48), Ser247 (6.52) and Asn250 (6.55) for electrostatic interactions, and Leu91 (3.33), Phe168 (EL2), Val169 (EL2), Trp243 (6.48), Leu246 (6.51), Leu264 (7.35), Tyr265 (7.36), Ile268 (7.39) for hydrophobic interactions, as can be seen in Figure 3, panel C. In this case only Asn250 can be involved in the hydrogen bond network (Figure 4, panel C), since in the A3 AR the position equivalent to Glu172 of the hA1 and Glu169 of the A2A AR is occupied by Val169, which cannot establish a hydrogen bond with the amino group at position 5 of compound 1.

Most of the poses resemble the conformation that ZM-241385 assumes in the binding site of the hA2A AR crystal structure and of hA1 and hA3 AR models. The benzene ring at position 2 occupies the position of the furan ring of ZM-241385, the 7-amino-pyrazolopyrimidine scaffold is well superimposed on the reference 7-amino-triazolotriazine and the arylalkylamino group at position 5 points in the same direction as the para-hydroxyphenyl-ethylamino fragment. The similarity of the binding modes confirms the expectation provided by the IEFs comparison between 1 and ZM-241385 (Figure 3, panels A, B and C). To quantitatively compare the calculated IEFs profiles, two novel analyses have been proposed called RMSD and RMSD\textsubscript{trend} analysis (see the Experimental Section for more details). In the case of derivative 1 both RMSD and RMSD\textsubscript{trend} between electrostatic and hydrophobic IEFs on each hAR subtype are quite low. However, the electrostatic RMSD (2.18 kcal/mol) and the electrostatic RMSD\textsubscript{trend} (0.47 kcal/mol) for the hA1 subtype are higher than the values observed for hA2A and hA3 ARs. This does not seem to fit with the low Ki (67 nM) for the hA1 receptor;
however, it appears that the major unfavorable contribution is provided by Glu170, which may probably be corrected by a slight rotation of the phenyl group of the compound.

Subsequently, we compared the binding behavior of compound 1 to that of its analog derivative A (Ki = 150 nM for hA1, Ki = 110 nM for hA2A and I% = 39 at 1 mM for hA3), having a methylene instead of the NH linker at the 5-position. The IEFs comparison did not allow a complete rationalization of the different selectivity profiles of compounds 1 and A (Figure SM1). Electrostatic RMSD and RMSD\(_{\text{trend}}\) values on the hA3 receptor (1.10 and 0.18 kcal/mol, respectively) are higher than those of compound 1 (0.60 and 0.06 kcal/mol, respectively), in accordance with the lower potency of derivative A (I% = 39% at 1 mM) compared with 1 (Ki = 13 nM). On the other hand, we have to honestly observe that also compound A presents higher RMSD and RMSD\(_{\text{trend}}\) values (1.49 and 0.18 kcal/mol, respectively) on the hA2A receptor as compared with compound 1 (0.96 and 0.08 kcal/mol, respectively), but in this case the affinity of the former (110 nM) is higher than that of the latter (412 nM). The result of the IEFs comparison is confirmed by the similarity of the binding modes of derivatives 1 and A at all receptor binding sites, as reported in Figure 4 (panels A, B and C). In this case docking is not sufficient to rationalize the difference in binding affinities. In fact, the mere examination of the final state of the binding process may not be sufficient to explain differences in the activity or selectivity profiles. The presence of water molecules and the entropic effect are only two among the pool of binding contributions that we are not taking into consideration during our docking simulations.

![Fig. 5 Results of the IEFs comparison between all compounds and reference compound ZM-241385. RMSDs and RMSD\(_{\text{trend}}\) between electrostatic (panels A and B, respectively) and hydrophobic (panels C and D, respectively) Energy Fingerprints of each compound (y-axis) and reference ZM-241385 are reported for hA1, hA2A and hA3 receptors (x-axis). A colorimetric scale going from blue to red represents favorable to unfavorable values. An exclamation point identifies those poses that have a positive van der Waals and/or electrostatic potential (and for which was not possible to select an alternative pose with negative values).](image-url)

Similar considerations can be made observing the results of IEFs comparison for all the dataset compounds on the different AR subtypes (Figure 5). We would have expected to find blue and red rectangles associated with good and bad binders, respectively, but this prevision was not satisfied: a major similarity of the IEFs between the target and the reference compounds are not always related to good binding affinity of the ligand. However, an interesting example is provided by compounds 8, 9 and 10, which have no affinity for any of the receptors. Red rectangles cross horizontally almost the whole hydrophobic RMSD and RMSD\textsubscript{trend} table, meaning that there is a considerable loss in the binding hydrophobic contribution in comparison with the reference. As a control experiment, ZM-241385 has been docked into the three AR subtypes, the IEFs have been computed for the selected poses and compared with that of the reference pose of ZM241385: as expected, the electrostatic and hydrophobic RMSD and RMSD\textsubscript{trend} values are close to zero (Figure 5).

The 5-(N4-arylalkylamino-1-yl) compounds 15–24 are hA\textsubscript{1} AR selective. These derivatives find a steric hindrance in the hA\textsubscript{3} binding site and the van der Waals values of the selected poses are positive (as indicated by exclamation points in Figure 5). However, from the IEFs comparison analysis (Figure 5), we would have predicted a hA\textsubscript{2A} versus hA\textsubscript{1} selectivity (blue versus red rectangles). In fact, while at the hA\textsubscript{2A} binding site the predicted poses of these compounds behave like ZM-241385, at the hA\textsubscript{1} binding site they deviate a little from the reference position, losing some of the canonical interactions (Video SM1-SM2). Interestingly, this diversion results in a gain for compounds 16, 17, 18, 19 and 22: the protonated amine at position 4 of the piperazine moiety is involved in an ionic interaction with Glu170 (EL2), which is confirmed by a highly negative electrostatic contribution reported on the heat map in the background of Video SM1. The absence of a negatively charged residue at a position equivalent to Glu170 on hA\textsubscript{2A} (Leu167) and hA\textsubscript{3} (Gln167) receptors may be associated with the hA\textsubscript{1} selectivity of these compounds.

4. Conclusion

The herein reported structural investigation was carried out to identify new antagonists targeting the hA\textsubscript{2A} AR or both the hA\textsubscript{1}/hA\textsubscript{2A} ARs. Hence, various arylalkylamino- and 4-substituted-piperazin-1-yl- moieties were appended at the 5-position of the pyrazolo[4,3-d]pyrimidine scaffold. The 4-hydroxyphenylethylamino group was the most profitable, since the ZM-241385-based compound 12 showed both good hA\textsubscript{2A} affinity (Ki = 150 nM) and the highest selectivity among all the ligands reported here. The 5-benzylamino moiety (compound 5) achieved the best combined hA\textsubscript{2A} (Ki = 123 nM) and hA\textsubscript{1} affinity (Ki = 25 nM) while the 5-phenethylamino pendant (compound 6) afforded nanomolar affinity (Ki = 11 nM) and good selectivity for the hA\textsubscript{1} AR. The 5-(N4 -substituted-piperazin-1-yl) derivatives 15–24 were inactive at the hA\textsubscript{2A} AR while the hA\textsubscript{1} affinities spanned the high nanomolar range. These outcomes provide new insights about the structural requirements of our pyrazolopyrimidine series for hA\textsubscript{2A}- and hA\textsubscript{1}-receptor ligand interaction. Nevertheless, the obtained
results do not prompt us to synthesize further derivatives of this series featured by 5-arylalkyamino- and 5-piperazino- moieties.

A structure-based molecular modeling study was conducted to rationalize the experimental binding data from a molecular point of view using molecular docking studies in tandem with Interaction Energy Fingerprints (IEFs) analysis. Moreover, to quantitatively compare IEFs profiles and, consequently, to address the similarity of the binding modes of different compounds in different receptor subtypes, two novel analyses have been proposed, called RMSD and RMSD\textsubscript{trend} analyses. Even if, we are conscious that the simple inspection of the final state of the binding process may not be sufficient to explain differences in the activity or selectivity profiles, these novel tools can facilitate the mode of representation and interpretation of the docking data obtained by analyzing simultaneously several compounds against different receptor subtypes.

5. Experimental section

5.1 Chemistry

The microwave-assisted syntheses were performed using an Initiator EXP Microwave Biotage instrument (frequency of irradiation: 2.45 GHz). Analytical silica gel plates (Merck F254), preparative silica gel plates (Merck F254, 2 mm) and silica gel 60 (Merck, 70–230 mesh) were used for analytical and preparative TLC, and for column chromatography, respectively. All melting points were determined on a Gallenkamp melting point apparatus and are uncorrected. Elemental analyses were performed with a Flash E1112 Thermofinnigan elemental analyzer for C, H, N and the results were within ±0.4% of the theoretical values. All final compounds revealed a purity not less than 95%. The IR spectra were recorded with a Perkin-Elmer Spectrum RX I spectrometer in Nujol mulls and are expressed in cm\(^{-1}\). The \(^1\)H NMR spectra were obtained with a Bruker Avance 400 MHz instrument. The chemical shifts are reported in \(\delta\) (ppm) and are relative to the central peak of the solvent which was CDCl\(_3\) or DMSO-d\(_6\). The assignment of exchangeable protons (OH, and NH) was confirmed by addition of D\(_2\)O. The following abbreviations are used: s = singlet, d = doublet, t = triplet, m = multiplet, br = broad and ar = aromatic protons.

4-Nitro-1-phenyl-1H-pyrazole-3-carbonitrile 26 [26]

The title compound was prepared with a different procedure from that previously described by us [26]. Briefly, phenylboronic acid (2.4 mmol), cupric acetate (1.8 mmol) and activated 4 Å molecular sieves (750 mg) were added to a solution of 4-nitro- 1H-pyrazole-3-carbonitrile [26] (1.2 mmol) in anhydrous dichloromethane (8 mL) and pyridine (2.4 mmol). The mixture was stirred at room temperature, under air, in a loosely capped flask for two days, then it was diluted with chloroform (20–30 mL) and filtered through celite. The solution was extracted with 0.1 M HCl (15 ml for three times), the organic phase was anhydrided (Na\(_2\)SO\(_4\)) and evaporated at reduced pressure to give a solid which was collected by suction, washed with
water and then cyclohexane and recrystallized. Yield 75%; m.p. 143–145°C (cyclohexane/EtOH); 1H NMR (DMSO-d$_6$) 7.54–7.63 (m, 3H, ar), 7.74–7.76 (m, 2H, ar), 8.71 (s, 1H, H-5).

**General procedure for the synthesis of 3-substituted-1-[3-cyano-1-R2-1H-pyrazol-4-yl]thioureas 32–42**

The commercially available phenyl-, 4-methoxyphenyl-, 2,4-dichlorophenyl- and benzyl-isothiocyanates or the suitably synthesized phenethyl- [29], 4-methoxyphenethyl- [31], 3,4-dimethoxyphenethyl- [30], phenylpropyl-isothiocyanates [29] (1.97 mmol) were added to a solution of the 1-substituted-4-amino-pyrazole-3-carbonitrides 29–31 [26] (1.64 mmol) in anhydrous DMF (1.5 mL). The mixture was stirred at room temperature for 3–4 h (compounds 32–34, 38, 39), for 16 h (compounds 35, 40, 42) and for 24 h (compounds 36, 37, 41).

The obtained dark slurry was treated with water (20 mL) and, in the case of compounds 33–35 and 39, a solid precipitated which was collected by filtration. For derivatives 32, 36–38, 40–42, the aqueous mixture was extracted with EtOAc (30 mL X3). The combined organic extracts were anhydridified (Na$_2$SO$_4$) and the solvent evaporated at reduced pressure. The obtained solid was treated with Et$_2$O (5–10 mL) and isolated by filtration. Crude compound 42 was purified by column chromatography (eluent:cyclohexane/ EtOAc/MeOH 6:4:1). Derivatives 32, 38–40, as well as 42, were unstable upon recrystallization, hence they were used as such for the next step.

**1-(3-Cyano-1-phenyl-1H-pyrazol-4-yl)-3-phenylthiourea 32**

Yield 89%; 1H NMR (DMSO-d$_6$) 7.19 (t, 1H, ar, J = 7.4 Hz), 7.36 (t, 2H, ar, J = 7.6 Hz), 7.46 (t, 1H, ar, J = 7.4 Hz), 7.51 (d, 2H, ar, J = 7.6 Hz), 7.58 (t, 2H, ar, J = 7.5 Hz), 7.88 (d, 2H, ar, J = 7.7 Hz), 9.59 (s, 1H, pyrazole proton), 9.68 (br s, 1H, NH), 10.04 (br, s, 1H, NH).

**1-(3-Cyano-1-phenyl-1H-pyrazol-4-yl)-3-(4-methoxyphenyl)thiourea 33**

Yield 95%; m.p. 165–167°C (cyclohexane/EtOAc); 1H NMR (DMSO-d$_6$) 3.75 (s, 3H, OCH$_3$), 6.95 (d, 2H, ar, J = 8.9 Hz), 7.33 (d, 2H, ar, J = 8.9 Hz), 7.46 (t, 1H, ar, J = 7.6 Hz), 7.58 (t, 2H, ar, J = 7.3 Hz), 7.89 (d, 2H, ar, J = 7.6 Hz), 8.95 (s, 1H, pyrazole proton), 9.56 (br s, 1H, NH), 9.89 (br s, 1H, NH). Anal. Calc. for C$_{18}$H$_{15}$N$_5$OS.

**1-(3-Cyano-1-phenyl-1H-pyrazol-4-yl)-3-(2,4-dichlorophenyl)thiourea 34**

Yield 98%; m.p. 166–169°C (cyclohexane/EtOAc); 1H NMR (DMSO-d$_6$) 7.44–7.60 (m, 5H, ar), 7.80 (s, 1H, ar), 7.96 (d, 2H, ar, J = 7.9 Hz), 9.02 (s, 1H, pyrazole proton), 9.82 (br s, 1H, NH), 9.98 (br s, 1H, NH). Anal. Calc. for C$_{17}$H$_{15}$Cl$_2$N$_5$S.
1-Benzyl-3-(3-cyano-1-phenyl-1H-pyrazol-4-yl)thiourea 35
Yield 74%; m.p. 180–183°C (EtOH). \(^1\)H NMR (DMSO-d6) 4.75 (d, 2H, CH2, J = 4.6 Hz), 7.26–7.34 (m, 5H, ar), 7.45 (t, 1H, ar, J = 7.3 Hz), 7.57 (t, 2H, ar, J = 7.4 Hz), 7.87 (d, 2H, ar, J = 8.1 Hz), 8.49 (br s, 1H, NH), 8.99 (s, 1H, H-5), 9.55 (br s, 1H, NH). Anal. Calc. for C\(_{18}\)H\(_{15}\)N\(_5\)S.

1-Phenylethyl-3-(3-cyano-1-phenyl-1H-pyrazol-4-yl)thiourea 36
Yield 55%; m.p. 161–164°C (cyclohexane/EtOAc). \(^1\)H NMR (DMSO-d6) 2.88 (t, 2H, CH2, J = 7.5 Hz), 3.69–3.70 (m, 2H, CH2), 7.23–7.34 (m, 5H, ar), 7.46 (t, 1H, ar, J = 7.0 Hz), 7.58 (t, 2H, ar, J = 7.7 Hz), 7.89 (d, 2H, ar, J = 8.1 Hz), 8.09 (br s, 1H, NH), 8.91 (s, 1H, pyrazole proton), 9.50 (s, 1H, NH). Anal. Calc. for C\(_{19}\)H\(_{17}\)N\(_5\)S.

1-(3-Cyano-1-phenyl-1H-pyrazol-4-yl)-3-phenylpropythiourea 37
Yield 57%; m.p. 133–136°C (cyclohexane/EtOAc). \(^1\)H NMR (DMSOD\(_6\)) 1.82–1.90 (m, 2H, CH2), 2.63 (t, 2H, CH2, J = 7.3 Hz), 3.49–3.50 (m, 2H, CH2), 7.17–7.31 (m, 5H, ar), 7.45 (t, 1H, ar, J = 7.4 Hz), 7.57 (t, 2H, ar, J = 7.7 Hz), 7.88 (d, 2H, ar, J = 7.9 Hz), 8.09 (br s, 1H, NH), 8.95 (s, 1H, pyrazole proton), 9.42 (s, 1H, NH). Anal. Calc. for C\(_{20}\)H\(_{19}\)N\(_5\)S.

1-(3-Cyano-1-methyl-1H-pyrazol-4-yl)-3-phenylthiourea 38
Yield 45%; \(^1\)H NMR (DMSO-d6) 3.92 (s, 3H, CH\(_3\)), 7.17 (t, 1H, ar, J = 7.3 Hz), 7.36 (t, 2H, ar, J = 7.8 Hz), 7.48 (d, 2H, ar, J = 7.7 Hz), 7.95 (s, 1H, pyrazole proton), 8.39 (br s, 1H, NH), 12.11 (br s, 1H, NH).

1-Benzyl-3-(3-cyano-1-phenyl-1H-pyrazol-4-yl)thiourea 39
Yield 56%; \(^1\)H NMR (DMSO-d6) 3.95 (s, 3H, CH3), 5.84 (br s, 2H, CH2), 7.19–7.39 (m, 5H, ar), 7.70 (s, 1H, pyrazole proton), 8.39 (br s, 1H, NH), 9.47 (br s, 1H, NH).

1-Benzyl-3-(1-benzyl-3-cyano-1H-pyrazol-4-yl)thiourea 40
Yield 50%; \(^1\)H NMR (DMSO-d6) 4.71 (d, 2H, CH2, J = 4.5 Hz), 5.40 (s, 2H, CH2), 7.26–7.40 (m, 10H, ar), 8.44 (br s, 1H, NH); 8.51 (s, 1H, pyrazole proton), 9.47 (s, 1H, NH).

1-(3-Cyano-1-phenyl-1H-pyrazol-4-yl)-3-(4-methoxyphenylethyl)thiourea 41
Yield 62%; m.p. 260–262°C (cyclohexane/EtOAc). \(^1\)H NMR (DMSO-d6) 4.71 (d, 2H, CH2, J = 7.2 Hz), 3.65–3.67 (m, 2H, CH2), 3.72 (s, 3H, CH3), 6.88 (d, 2H, ar, J = 8.9 Hz), 7.17 (d, 2H, ar, J = 8.9 Hz), 7.46 (t, 1H, ar, J = 7.0 Hz), 7.58 (t, 2H, ar, J = 7.7 Hz), 7.87 (d, 2H, ar, J = 7.7 Hz), 8.05 (br s, 1H, NH), 8.91 (s, 1H, pyrazole proton), 9.48 (s, 1H, NH). Anal. Calc. for C\(_{20}\)H\(_{19}\)N\(_5\)S.
1-(3-Cyano-1-phenyl-1H-pyrazol-4-yl)-3-[2-(3,4-dimethoxyphenyl)ethyl]thiourea 42

Yield 55%; $^1$H NMR (DMSO-d$_6$) 2.81 (t, 2H, CH$_2$, J = 7.2 Hz), 3.69–3.72 (m, 5H, OCH$_3$ + CH$_2$), 3.75 (s, 3H, OCH$_3$), 6.76 (d, 1H, ar, J = 8.1 Hz), 6.84 (s, 1H, ar), 6.89 (d, 1H, ar, J = 8.2 Hz), 7.46 (t, 1H, ar, J = 7.2 Hz), 7.58 (t, 2H, ar, J = 7.1 Hz), 7.86 (d, 2H, ar, J = 8.0 Hz), 8.04 (br s, 1H, NH), 8.91 (s, 1H, pyrazole proton), 9.48 (br s, 1H, NH).

Anal. Calc. for C$_{21}$H$_{21}$N$_5$O$_2$S.

General procedure for the synthesis of S-methylisothiourea derivatives 43–53

A mixture of the suitable thiourea derivatives 32–42 (0.92 mmol) and iodomethane (3.69 mmol) in 0.1 N NaOH solution (11.8 mL) was stirred at room temperature until the disappearance of the starting material (12–24 h). Then, glacial acetic acid was added until pH 6. The solid which precipitated was collected by filtration and dried, except compounds 46 and 51 which were isolated from the reaction mixture by extraction with EtOAc (30 mL X3). Evaporation of the anhydrified (Na$_2$SO$_4$) organic phase gave a solid which was collected by filtration. These S-methylisothiourea derivatives were unstable upon recrystallization, thus they were used for the next step without further purification. It was observed that derivatives 43–45 and 51 exist in two tautomeric forms in DMSO solution. In fact, in their $^1$H NMR spectra there are two signals assignable to the SCH$_3$ and to the pyrazole proton. Compounds 44 and 51 also display, two signals assignable to the OCH$_3$ and SCH$_3$ substituents, respectively (see below for details).

$N$-(3-Cyano-1-phenylpyrazolo-4-yl)-$N'$-phenyl-S-methylisothiourea 43

Yield 86%; $^1$H NMR (DMSO-d$_6$) mixture of two tautomers (ratio about 1:2.7) 2.35 (s, SCH$_3$), 2.38 (s, SCH$_3$), 7.30–7.31 (m, ar), 7.42–7.60 (m, ar +2 NH), 7.98 (d, ar, J = 8.0 Hz), 8.89 (s, pyrazole proton), 8.93 (s, pyrazole proton).

$N$-(3-Cyano-1-phenylpyrazolo-4-yl)-$N'$-4-methoxyphenyl-S-methylisothiourea 44

Yield 98%; $^1$H NMR (DMSO-d$_6$) mixture of two tautomers (ratio about 1:3.2) 2.41 (s, SCH$_3$), 2.43 (s, SCH$_3$), 3.86 (s, OCH$_3$), 3.88 (s, OCH$_3$), 7.41–7.72 (m, ar), 7.96–8.01 (m, ar), 8.93 (s, pyrazole proton), 9.01 (br s, pyrazole proton + NH).

$N'$-2,4-Dichlorophenyl-$N$-(3-cyano-1-phenylpyrazolo-4-yl)-S-methylisothiourea 45

Yield 95%; $^1$H NMR (DMSO-d$_6$) mixture of two tautomers (ratio about 1:3.6) 2.36 (s, SCH$_3$), 2.42 (s, SCH$_3$), 7.09–7.70 (m, ar), 8.00–8.16 (m, ar + NH), 9.00 (s, pyrazole proton), 9.18 (s, pyrazole proton).
$N'$-Benzyl-$N$-(3-cyano-1-phenylpyrazolo-4-yl)-S-methylisothiourea 46

Yield 98%; $^1$H NMR (DMSO-$d_6$) 2.48 (s, 3H, SCH$_3$), 5.47 (br s, 2H, CH$_2$) 7.24–7.34 (m, 5H, ar), 7.44 (t, 1H, ar, J = 7.3 Hz), 7.57 (t, 2H, ar, J = 7.6 Hz), 7.97 (d, 2H, ar, J = 8.0 Hz), 8.13 (br s, 1H, NH), 8.80 (s, 1H, pyrazole proton).

$N$-(3-Cyano-1-phenylpyrazolo-4-yl)-$N'$-phenylethyl-S-methylisothiourea 47

Yield 86%; $^1$H NMR (DMSO-$d_6$) 2.56 (s, 3H, CH$_3$), 3.04 (t, 2H, CH$_2$, J = 7.1 Hz), 4.33 (t, 2H, CH$_2$, J = 7.1 Hz), 7.25–7.41 (m, 5H, ar), 7.43 (t, 1H, ar, J = 9.2 Hz), 7.57 (t, 2H, ar, J = 9.2 Hz), 7.98 (d, 2H, ar, J = 9.2 Hz), 8.14 (s, 1H, pyrazole proton).

$N$-(3-Cyano-1-phenylpyrazolo-4-yl)-$N'$-phenylpropyl-S-methylisothiourea 48

Yield 86%; $^1$H NMR (DMSO-$d_6$) 2.03–2.05 (m, 2H, CH$_2$), 2.51 (s, 3H, CH$_3$), 2.71 (t, 2H, CH$_2$, J = 7.4 Hz), 4.16 (t, 2H, CH$_2$, J = 7.4 Hz), 7.21 (t, 1H, ar, J = 9.0 Hz), 7.27–7.32 (m, 4H, ar), 7.42 (t, 1H, ar, J = 9.0 Hz), 7.56 (t, 2H, ar, J = 9.0 Hz), 7.96 (d, 2H, ar, J = 9.0 Hz), 8.03 (s, 1H, NH), 8.83 (s, 1H, pyrazole proton).

$N$-(3-Cyano-1-methylpyrazolo-4-yl)-$N'$-phenyl-S-methylisothiourea 49

Yield 87%; $^1$H NMR (DMSO-$d_6$) 2.23 (s, 3H, SCH$_3$), 3.39 (s, 3H, CH$_3$), 7.33–7.35 (m, 2H, ar), 7.52–7.60 (m, 3H, ar), 8.07 (s, 1H, pyrazole proton).

$N'$-Benzyl-$N$-(3-cyano-1-methylpyrazolo-4-yl)-S-methylisothiourea 50

Yield 84%; $^1$H NMR (DMSO-$d_6$) 2.43 (s, 3H, SCH$_3$), 3.99 (s, 3H, CH$_3$), 5.42 (br s, 2H, CH$_2$), 7.20–7.30 (m, 5H, ar), 7.75 (s, 1H, NH), 8.05 (s, 1H, pyrazole proton).

$N'$-Benzyl-$N$-(3-cyano-1-benzylpyrazolo-4-yl)-S-methylisothiourea 51

Yield 81%; $^1$H NMR (DMSO-$d_6$) mixture of two tautomers (ratio about 1:6) 2.40 (s, SCH$_3$) 2.43 (s, SCH$_3$), 5.41 (br s, CH$_2$), 5.48 (s, CH$_2$), 7.14–7.39 (m, ar), 7.81 (br s, NH), 8.25 (s, pyrazole proton), 8.17 (s, pyrazole proton).

$N'$-4-Methoxyphenylethyl-$N$-(3-cyano-1-phenylpyrazolo-4-yl)-Smethylisothiourea 52

Yield 86%; $^1$H NMR (DMSO-$d_6$) 2.56 (s, 3H, SCH$_3$), 2.97 (t, 2H, CH$_2$, J = 7.3 Hz), 3.75 (s, 3H, OCH$_3$), 4.29 (t, 2H, CH$_2$, J = 7.3 Hz), 6.91 (d, 2H, ar, J = 9.5 Hz), 7.24 (d, 2H, ar, J = 9.5 Hz), 7.43 (t, 1H, ar, J = 9.5 Hz), 7.57 (t, 2H, ar, J = 9.3 Hz), 7.98 (d, 2H, ar, J = 9.3 Hz), 8.11 (s, 1H, NH), 8.85 (s, 1H, pyrazole proton).

$N'$-3,4-Dimethoxyphenylethyl-$N$-(3-cyano-1-phenylpyrazolo-4-yl)-Smethylisothiourea 53

Yield 73%; $^1$H NMR (DMSO-$d_6$) 2.58 (s, 3H, SCH$_3$), 2.98 (t, 2H, CH$_2$, J = 8.5 Hz), 3.73 (s, 3H, OCH$_3$), 3.76 (s, 3H, OCH$_3$), 4.34 (t, 2H, CH$_2$, J = 7.2 Hz), 6.83–6.93 (m, 4H, 3 ar + NH), 7.45 (t, 1H, ar, J = 7.3 Hz), 7.59 (t, 2H, ar, J = 7.8 Hz), 7.99 (d, 2H, ar, J = 8.3 Hz), 8.95 (s, 1H, pyrazole proton).
General procedure for the synthesis of 5-aryl(alkyl)amino-7-amino-2H-pyrazolo[4,3-d]pyrimidine derivatives 1–3, 5–12

A mixture of the suitable S-methylisothioureas 43–53 (1 mmol) and NH₄Cl (20 mmol) in formamide (2 mL) was microwave irradiated at 110°C for 20 min (compounds 9, 10), at 130°C for 40 min (compound 12) and for 2 h (compounds 2, 3), at 150°C for 15 min (compounds 1, 5, 8) and for 20 min (compounds 6, 7, 11). The suspension was then treated with NaHCO₃ saturated solution until pH 7 and the obtained solid was collected by filtration to give compounds 1–3. To isolate derivatives 5–12, the mixture was extracted with CHCl₃ (15 mL × 3), the organic phase was washed with water (15 mL × 2) and anhydrified (Na₂SO₄). Evaporation of the solvent at reduced pressure afforded a residue which was taken up with diethyl ether (2–3 mL) and collected by filtration. The crude derivatives were purified by recrystallization, except compounds 1, 6, 7, 10, 11 which were first purified by column chromatography or preparative TLC (see below for details).

7-Amino-5-phenylamino-2-phenyl-2H-pyrazolo[4,3-d]pyrimidine 1

Purified by column chromatography (Et₂O/cyclohexane/EtOAc 3:1:1). Yield 66%; m.p. 252–254°C (EtOH). ¹H NMR (DMSO-d₆) 6.86 (t, 1H, ar, J = 7.3 Hz), 7.23 (t, 2H, ar, J = 7.6 Hz), 7.41–7.45 (m, 3H, 1 ar + NH₂), 7.58 (t, 2H, ar, J = 7.6 Hz), 7.90 (d, 2H, ar, J = 8.5 Hz), 8.03 (d, 2H, ar, J = 8.5 Hz), 8.74 (s, 1H, H–3), 8.76 (br s, 1H, NH). Anal. Calc. for C₁₇H₁₄N₆.

7-Amino-5-(4-methoxyphenyl)amino-2-phenyl-2H-pyrazolo[4,3-d]pyrimidine 2

Yield 62%; m.p. 253–255°C (cyclohexane/EtOAc); ¹H NMR (DMSO-d₆) 3.82 (s, 3H, OCH₃), 6.84 (d, 2H, ar, J = 8.9 Hz), 7.45–7.40 (m, 3H, 1 ar + NH₂), 7.58 (t, 2H, ar, J = 7.6 Hz), 7.90 (d, 2H, ar, J = 8.5 Hz), 8.03 (d, 2H, ar, J = 8.5 Hz), 8.74 (s, 1H, H–3), 8.76 (br s, 1H, NH). Anal. Calc. for C₁₈H₁₆N₆O.

7-Amino-5–(2,4-dichlorophenyl)amino-2-phenyl-2H-pyrazolo[4,3-d]pyrimidine 3

Yield 58%; m.p. 251–252°C (cyclohexane/EtOAc); ¹H NMR (DMSO-d₆) 7.40–7.48 (m, 3H, 2 ar + NH), 7.54 (s, 1H, ar), 7.53–7.61 (m, 2H, ar), 7.61–7.79 (br s, 2H, NH₂), 8.02 (d, 2H, J = 7.9 Hz), 8.67 (d, 1H, J = 8.9 Hz), 8.82 (s, 1H, H–3). Anal. Calc. for C₁₇H₁₂Cl₂N₆.

7-Amino-5-benzylamino-2-phenyl-2H-pyrazolo[4,3-d]pyrimidine 5

Purified by preparative TLC (Et₂O/cyclohexane/EtOAc 3:1:1). Yield 60%; m.p. 143–145°C (cyclohexane/EtOAc). ¹H NMR (DMSO-d₆) 4.49 (d, 2H, CH₂, J = 6.3 Hz), 6.71 (t, 1H, NH, J = 6.3 Hz), 7.18 (t, 1H, ar, J = 7.1 Hz), 7.26–7.34 (m, 6H, 4 ar + NH₂), 7.39 (t, 1H, ar, J = 7.4 Hz), 7.55 (t, 2H, ar, J = 7.5 Hz), 7.95 (d, 2H, ar, J = 7.6 Hz), 8.48 (s, 1H, H–3). Anal. Calc. for C₁₈H₁₆N₆.
7-Amino-5-(2-phenylethyl)amino-2-phenyl-2H-pyrazolo[4,3-d]pyrimidine 6

Purified by column chromatography (cyclohexane/ EtOAc/MeOH 6:4:1). Yield 65%; m.p. 168–171°C (cyclohexane/ EtOAc). 1H NMR (DMSO-d6) 2.86 (t, 2H, CH2, J = 7.1 Hz), 3.45–3.50 (m, 2H, CH2), 6.12 (br s, 1H, NH), 7.18–7.32 (m, 7H, 5 ar + NH2), 7.39 (t, 1H, ar, J = 7.5 Hz), 7.56 (t, 2H, ar, J = 7.5 Hz), 7.98 (d, 2H, ar, J = 7.7 Hz), 8.52 (s, 1H, H-3). Anal. Calc. for C19H18N6.

7-Amino-5-(3-phenylpropyl)-2-phenyl-2H-pyrazolo[4,3-d]pyrimidines 7

Purified by column chromatography (eluent cyclohexane/ EtOAc/MeOH 6:4:1). Yield 58%; m.p. 159–162°C (EtOAc). 1H NMR (DMSO-d6) 1.83–1.86 (m, 2H, CH2), 2.64 (t, 2H, CH2, J = 7.4 Hz), 3.27 (m, 2H, CH2), 6.20 (br s, 1H, NH), 7.19–7.30 (m, 7H, 5 ar + NH2), 7.40 (t, 1H, ar, J = 9.0 Hz), 7.56 (t, 2H, ar, J = 7.6 Hz), 7.97 (d, 2H, ar, J = 7.9 Hz), 8.50 (s, 1H, H-3). Anal. Calc. for C20H20N6.

7-Amino-2-methyl-5-phenylamino-2H-pyrazolo[4,3-d]pyrimidine 8

Yield 42%; m.p. 252–254°C (EtOH). 1H NMR (DMSO-d6) 4.05 (s, 3H, Me), 6.84 (t, 1H, ar, J = 7.2 Hz), 7.19–7.21 (m, 4H, 2 ar + NH2), 7.75 (d, 2H, ar, J = 7.2 Hz), 7.87 (s, 1H, H-3), 8.62 (br s, 1H, NH). Anal. Calc. for C12H12N6.

7-Amino-2-methyl-5-benzylamino-2H-pyrazolo[4,3-d]pyrimidine 9

Yield 80%; m.p. 213–214°C (EtOH); 1H NMR (DMSO-d6) 3.98 (s, 3H, CH3), 4.45 (d, 2H, CH2, J = 6.4 Hz), 6.47 (br s, 1H, NH), 7.05 (br s, 2H, NH2), 7.16–7.32 (m, 5H, ar), 7.70 (s, 1H, H-3). IR: 3326, 3179, 1658. Anal. Calc. for C13H14N6.

7-Amino-2-benzyl-5-benzylamino-2H-pyrazolo[4,3-d]pyrimidine 10

Yield 40%; m.p. 174–175°C (cyclohexane/EtOAc). 1H NMR (DMSO-d6) 4.45 (d, 2H, CH2, J = 6.3 Hz), 5.45 (s, 2H, CH2), 6.58 (br s, 2H, NH2), 7.11 (br s, 2H, NH2), 7.15–7.36 (m, 10H, ar), 7.87 (s, 1H, H-3). Anal. Calc. for C19H18N6 O.

7-Amino-2-(4-methoxyphenyl)ethylamino-2-phenyl-2H-pyrazolo[4,3-d]pyrimidine 11

Purified by column chromatography (cyclohexane/EtOAc/MeOH 6:4:1), Yield 58%; m.p. 142–145°C (cyclohexane/EtOAc). 1H NMR (DMSO-d6) 2.73 (t, 2H, CH2, J = 7.3 Hz), 3.38–3.41 (m, 2H, CH2), 3.73 (s, 3H, OCH3), 6.03 (br s, 1H, NH), 6.69 (d, 2H, ar, J = 8.9 Hz), 7.04 (d, 2H, ar, J = 8.9 Hz), 7.22 (br s, 2H, NH2), 7.41 (t, 1H, ar, J = 7.3 Hz), 7.57 (t, 2H, ar, J = 7.6 Hz), 7.98 (d, 2H, ar, J = 7.9 Hz), 8.51 (s, 1H, H-3). Anal. Calc. for C20H20N6O.

7-Amino-2-(3,4-dimethoxyphenyl)ethylamino-2-phenyl-2H-pyrazolo[4,3-d]pyrimidine 12

Purified by preparative TLC (cyclohexane/EtOAc/MeOH 6:4:1). Yield 45%; m.p. 100–102°C (H2O/MeOH). 1H NMR (DMSO-d6) 2.79 (t, 2H, CH2, J = 7.2 Hz), 3.43–3.48 (m, 2H, CH2), 3.72 (s, 3H, OCH3), 3.74 (s, 3H, OCH3), 3.74 (s, 3H, OCH3),
6.10 (br s, 1H, NH), 6.75 (d, 1H, ar, J = 6.5 Hz), 6.84–6.87 (m, 2H, ar), 7.27 (br s, 2H, NH2), 7.39 (t, 1H, ar, J = 7.2 Hz), 7.56 (t, 2H, ar, J = 7.7 Hz), 7.97 (d, 2H, ar, J = 7.8 Hz), 8.51 (s, 1H, H-3). Anal. Calc. for C_{21}H_{22}N_{6}O_{2}.

**General procedure for the synthesis of the pyrazolo[4,3-d]pyrimidine-7-amine derivatives 4, 13 and 14**

To a suspension of the methoxy-substituted pyrazolopyrimidine derivatives 2, 11 and 12 (1.02 mmol) in anhydrous CH₂Cl₂ (20 mL), a 1 M BBr₃ solution (2.60 mL for 2, 11 and 5.2 mL for 12) in CH₂Cl₂ was added at 0°C, under nitrogen atmosphere. The mixture was stirred at room temperature for 20–24 h (compounds 4, 13) or 16 h (compound 14), then was diluted with water (10 mL) and neutralized with NaHCO₃ saturated solution. The organic solvent was removed under reduced pressure and the obtained precipitate was collected by filtration and recrystallized. The crude derivative 4 was first purified by column chromatography (eluent CHCl₃/MeOH 9:1) and then recrystallized.

**7-Amino-5-(4-hydroxyphenyl)amino-2-phenyl-2H-pyrazolo[4,3-d]pyrimidine 4**

Yield 67%; m.p. 223–225°C (EtOAc/cyclohexane); ¹H NMR (DMSO-d₆) 6.65 (d, 2H, ar, J = 8.8 Hz), 7.30–7.45 (m, 3H, 2 ar + NH₂), 7.57 (t, 2H, ar, J = 7.6 Hz), 7.63 (d, 2H, ar, J = 8.8 Hz), 8.01 (d, 2H, ar, J = 8.3 Hz), 8.41 (s, 1H, NH), 8.65 (s, 1H, H-3), 8.85 (s, 1H, OH). Anal. Calc. for C_{17}H_{14}N_{6}O.

**7-Amino-5-(4-hydroxyphenethyl)amino-2-phenyl-2H-pyrazolo[4,3-d]pyrimidine 13**

Yield 89%; m.p. 241–244°C (EtOAc/EtOH). ¹H NMR (DMSO-d₆) 2.73 (t, 2H, CH₂, J = 7.3 Hz), 3.38–3.40 (m, 2H, CH₂), 6.03 (br s, 1H, NH), 6.68 (d, 2H, ar, J = 8.3 Hz), 7.04 (d, 2H, ar, J = 8.3 Hz), 7.22 (br s, 2H, NH₂), 7.41 (t, 1H, ar, J = 7.4 Hz), 7.56 (t, 2H, ar, J = 7.7 Hz), 7.97 (d, 2H, ar, J = 7.8 Hz), 8.51 (s, 1H, H-3), 9.14 (s, 1H, OH). Anal. Calc. for C_{19}H_{18}N_{6}O.

**7-Amino-5-(3,4-dihydroxyphenethyl)amino-2-phenyl-2H-pyrazolo[4,3-d]pyrimidine 14**

Yield 50%; m.p. 242–243°C (EtOH). ¹H NMR (DMSO-d₆) 2.65 (t, 2H, CH₂, J = 7.1 Hz), 3.37–3.41 (m, 2H, CH₂), 6.03 (t, 1H, NH, J = 5.8 Hz), 6.59–6.47 (d, 1H, ar, J = 8.0 Hz), 6.63–6.66 (m, 2H, ar), 7.23 (br s, 2H, NH₂), 7.40 (t, 1H, ar, J = 7.4 Hz), 7.56 (t, 2H, ar, J = 7.6 Hz), 7.97 (d, 2H, ar, J = 7.7 Hz), 8.51 (s, 1H, H-3), 8.62 (br s, 1H, OH), 8.74 (br s, 1H, OH). Anal. Calc. for C_{19}H_{18}N_{6}O₂.

**Synthesis of 5,7-dichloro-2-phenyl-2H-pyrazolo[4,3-d]pyrimidine 55**

A suspension of the pyrazolopyrimidine-5,7-dione derivative 54 [20] (2 mmol) and N,N-dimethylaniline (3.95 mmol) in phosphorus oxychloride (5 mL) was microwave irradiated at 150°C for 20 min. The excess of phosphorus oxychloride was distilled off under reduced pressure and the residue was treated with water (about 5-10 mL). The crude product was collected by filtration and recrystallized. Yield 96%; m.p. 252–254°C (EtOH). ¹H NMR (DMSO-d₆) 7.62 (t, 1H, ar, J = 9.1 Hz), 7.69 (t, 2H, ar, J = 9.3 Hz), 8.17 (d, 2H, ar, J = 9.1 Hz), 9.69 (s, 1H, H-3). Anal. Calc. for C_{11}H_{10}N_{6}Cl₂.
A suspension of the suitable 5,7-dichloropyrazolopyrimidine derivative 55 (1.72 mmol) in aqueous 33% ammonia solution (10 mL) was microwave irradiated at 100°C for 30 min. The suspension was cooled at room temperature and the solid was collected by filtration and recrystallized. Yield 90%; m.p. 260–261°C (2-ethoxyethanol). $^1$H NMR (DMSO-d$_6$) 7.50 (t, 1H, ar, J = 8.0 Hz), 7.62 (t, 2H, ar, J = 8.0 Hz), 8.05 (d, 2H, ar, J = 8.0 Hz), 8.35 (br s, 1H, NH$_2$), 8.38 (br s, 1H, NH$_2$), 9.05 (s, 1H, H-3). Anal. Calc. for C$_{11}$H$_8$ClN$_5$.

**General procedure for the synthesis of 5-(4-R-piperazin-1-yl)-substituted pyrazolo[4,3-d]pyrimidines 15–21**

A mixture of the 5-chloro-pyrazolopyrimidine derivative 56 (0.41 mmol), the suitable N-substituted piperazine 57–63 (0.82 mmol) and ethyldiisopropylamine (0.49 mmol) in N-methylpyrrolidone (2 mL) was heated by microwave irradiation in the conditions described below for each compound. The obtained slurry was poured dropwise into water (50 mL) under vigorous stirring. The solid which precipitated was collected by filtration, purified by chromatography (column or preparative TLC, as reported below for each derivative) and then recrystallized, except derivative 16 which was directly recrystallized. The not commercially available 1-substituted piperazines were prepared as reported below (60) or as previously described (59, 61) [35,36].

**7-Amino-5-(4- phenylpiperazin-1-yl)-2H-pyrazolo[4,3-d]pyrimidine 15**

The reaction mixture was microwave irradiated at 150°C for 15 min. Column chromatography, eluent: acetonitrile. Yield 48%; m.p. 183–185°C (cyclohexane/EtOAc). $^1$H NMR (DMSO-d$_6$) 3.18–3.20 (m, 4H, piperazine protons), 3.80–3.85 (m, 4H, piperazine protons), 6.80 (t, 1H, ar, J = 7.1 Hz), 7.00 (d, 2H, ar, J = 7.3 Hz), 7.24 (t, 2H, ar, J = 7.3 Hz), 7.40–7.46 (m, 3H, 1 ar + NH$_2$), 7.57 (t, 2H, ar, J = 7.5 Hz), 8.00 (d, 2H, ar, J = 8.4 Hz), 8.59 (s, 1H, H-3). Anal. Calc. for C$_{21}$H$_{21}$N$_7$.

**7-Amino-5-(4- benzylpiperazin-1-yl)-2H-pyrazolo[4,3-d]pyrimidine 16**

The reaction mixture was microwave irradiated at 130°C for 25 min. Yield 74%; m.p. 201–202°C (diisopropyl ether/MeOH). $^1$H NMR (DMSO-d$_6$) 2.39–2.42 (m, 4H, piperazine protons), 6.80 (t, 1H, ar, J = 7.1 Hz), 7.00 (d, 2H, ar, J = 7.3 Hz), 7.24–7.29 (m, 1H, ar), 7.33–7.35 (m, 4H, ar), 7.39–7.42 (m, 3H, 1 ar + NH$_2$), 7.56 (t, 2H, ar, J = 7.6 Hz), 7.97 (d, 2H, ar, J = 7.7 Hz), 8.55 (s, 1H, H-3). Anal. Calc. for C$_{22}$H$_{23}$N$_7$.

**7-Amino-5-(4-phenylethylpiperazin-1-yl)-2H-pyrazolo[4,3-d]pyrimidine 17**

The reaction mixture was microwave irradiated at 150°C for 1 h. Column chromatography: eluent EtOAc/CH$_2$Cl$_2$/MeOH, 8:3:1. Yield 65%; m.p. 198–200°C (cyclohexane/EtOAc). $^1$H NMR (DMSO-d$_6$) 2.63–2.70 (m, 6H, 4 piperazine protons + CH$_2$), 2.87–2.91 (m, 2H, CH$_2$), 3.88–3.90 (m, 4H, piperazine protons), 5.53 (br s, 2H, NH$_2$), 7.23–7.34 (m, 5H, ar), 7.41 (t, 1H, ar, J = 7.4 Hz), 7.53 (t, 2H, ar, J = 8.2 Hz), 7.81 (d, 2H, ar, J = 7.6 Hz), 8.08 (s, 1H, H-3). Anal. Calc. for C$_{23}$H$_{26}$N$_7$. 

---

The reaction mixture was microwave irradiated at 150°C for 1 h and 45 min. Column chromatography: eluent cyclohexane/EtOAc 7:3. Yield 85%; m.p. 233–235°C (cyclohexane/EtOAc). 1H NMR (DMSO-d6) 2.43 (br s, 4H, piperazine protons), 3.57 (s, 2H, CH2), 3.67 (br s, 4H, piperazine protons), 7.20 (t, 1H, ar, J = 8.3 Hz), 7.41 (t, 2H, ar, J = 7.2 Hz), 7.50 (br s, 2H, NH2), 7.56 (t, 2H, ar, J = 7.7 Hz), 7.97 (d, 2H, ar, J = 8.0 Hz), 8.55 (s, 1H, H-3). Anal. Calc. for C22H20N7F3

7-Amino-5-(4-(2,4,6-trifluoro)benzyl)piperazin-1-yl)-2-phenyl-2H-pyrazolo[4,3-d]pyrimidine 18

The reaction mixture was microwave irradiated at 150°C for 1 h and 45 min. Column chromatography: eluent cyclohexane/EtOAc 7:3. Yield 85%; m.p. 233–235°C (cyclohexane/EtOAc). 1H NMR (DMSO-d6) 2.43 (br s, 4H, piperazine protons), 3.57 (s, 2H, CH2), 3.67 (br s, 4H, piperazine protons), 7.20 (t, 1H, ar, J = 8.3 Hz), 7.41 (t, 2H, ar, J = 7.2 Hz), 7.50 (br s, 2H, NH2), 7.56 (t, 2H, ar, J = 7.7 Hz), 7.97 (d, 2H, ar, J = 8.0 Hz), 8.55 (s, 1H, H-3). Anal. Calc. for C22H20N7F3

7-Amino-5-(4-(2-chloro-4-fluoro)benzyl)piperazin-1-yl)-2-phenyl-2H-pyrazolo[4,3-d]pyrimidine 19

The reaction mixture was microwave irradiated at 150°C for 1 h and 15 min. Column chromatography: eluent cyclohexane/EtOAc 7:3. Yield 53%; m.p. 203–205°C. 1H NMR (DMSO-d6) 2.47 (br s, 4H, piperazine protons), 3.57 (s, 2H, CH2), 3.70 (br s, 4H, piperazine protons), 7.24 (t, 1H, ar, J = 6.2 Hz), 7.41–7.60 (m, 7H, ar + NH2), 7.98 (d, 2H, ar, J = 7.8 Hz), 8.56 (s, 1H, H-3). Anal. Calc. for C22H21N7ClF

7-Amino-2-phenyl-5-(4-(2-furyl)piperazin-1-yl)-2H-pyrazolo[4,3-d]pyrimidine 20

The reaction mixture was microwave irradiated at 150°C for 1 h. Preparative TLC: eluent cyclohexane/EtOAc/MeOH 3:6:1. Yield 84%; m.p. 263–264°C (EtOH). 1H NMR (DMSO-d6) 3.74–3.79 (m, 8H, CH2), 6.65–6.66 (m, 1H, furan proton), 7.03–7.04 (m, 1H, furan proton), 7.42 (t, 1H, ar, J = 7.4 Hz), 7.51 (br s, 2H, NH2), 7.57 (t, 2H, ar, J = 7.8 Hz), 7.87 (m, 1H, furan proton), 7.99 (d, 2H, ar, J = 7.6 Hz), 8.60 (s, 1H, H-3). Anal. Calc. for C20H19N7O2

Tert-butyl 4-(7-amino-2-phenyl-2H-pyrazolo[4,3-d]pyrimidin-5-yl)piperazine 1-carboxylate 21

The reaction mixture was microwave irradiated at 150°C for 1 h and 30 min. Column chromatography: eluent cyclohexane/EtOAc/MeOH 3:6:1. Yield 76%; m.p. 263–264°C (EtOH). 1H NMR (DMSO-d6) 3.74–3.79 (m, 8H, CH2), 6.65–6.66 (m, 1H, furan proton), 7.03–7.04 (m, 1H, furan proton), 7.42 (t, 1H, ar, J = 7.4 Hz), 7.51 (br s, 2H, NH2), 7.57 (t, 2H, ar, J = 7.8 Hz), 7.87 (m, 1H, furan proton), 7.99 (d, 2H, ar, J = 7.6 Hz), 8.60 (s, 1H, H-3). Anal. Calc. for C20H25N7O2

Synthesis of 2-phenyl-5-(4-(methyl-2-furyl)piperazin-1-yl)-2H-pyrazolo[4,3-d]pyrimidin-7-amine 22

A solution of the 5-(4-(2-furyl)piperazin-1-yl) derivative 20 (1 mmol) in anhydrous THF (5 mL) was added to a suspension of LiAlH4 (3 mmol) in anhydrous THF (20 mL) at 0°C. The suspension was stirred for 16 h at room temperature, then treated with water (10 mL) and the solid which precipitated was filtered off. The clear solution was diluted with water (about 30 mL) and extracted with EtOAc (3 X 20 mL). The organic phase was anhydried and the solvent removed at reduced pressure to give a solid which was purified by preparative TLC (eluent: cyclohexane/EtOAc/MeOH 5:5:0.4). Yield 65%; m.p. 199–200°C. 1H NMR (DMSO-d6) 2.40–2.43 (m, 4H, piperazine protons), 3.53 (s, 2H, CH2), 3.67–3.69 (m, 4H, piperazine protons), 6.30–6.32 (m, 1H, furan
proton), 6.41–6.43 (m, 1H, furan proton), 7.38–7.43 (m, 3H, 1 ar + NH2), 7.56–7.70 (m, 3H, 2 ar +1 furan proton), 7.98 (d, 2H, ar, J = 7.4 Hz), 8.54 (s, 1H, H-3). Anal. Calc. for C20H21N7O.

Synthesis of 1-(2,4,6-trifluorobenzyl)piperazinium trifluoroacetate 60

A solution of N-(Boc)piperazine 63 (2.06 mmol) and 2,4,6-trifluorobenzaldehyde (1.87 mmol) in anhydrous CH2Cl2 (20 mL) was stirred at room temperature for 1.5 h, then triacetoxy sodium borohydride (7.47 mmol) was added portion wise. The mixture was refluxed for 48 h, then treated with iced water (10 mL) and diluted with CH2Cl2 (15 mL). The aqueous phase was extracted with CH2Cl2 (10 mL X 3) and the organic phases were collected and anhydrified (Na2SO4). Evaporation of the solvent at reduced pressure gave the crude tert-butyl-4-(2,4,6-trifluorobenzyl)piperazine-1-carboxylate which was purified by preparative TLC (eluent: CH2Cl2/acetonitrile/cyclohexane, 9:1:1) and obtained as a yellow oil. Yield 91%; 1H NMR (CDCl3) 1.46 (s, 9H, t-But), 2.41–2.45 (m, 4H, piperazine protons), 3.42–3.46 (m, 4H, piperazine protons), 3.67 (s, 2H, CH2), 6.69 (t, ar, J = 7.8 Hz). This derivative was then transformed into the title compound as follows. A solution of concentrated trifluoroacetic acid (2.5 mL) in anhydrous CH2Cl2 (2.5 mL) was added dropwise to a solution of tert-butyl-4-(2,4,6-trifluorobenzyl)piperazine-1-carboxylate (1.04 mmol) in anhydrous CH2Cl2 (20 mL). The solution was stirred at room temperature for 3 h, then the solvent and the excess of the acid were removed at reduced pressure. The residue was treated with Et2O (5 mL) to give a solid which was collected by filtration and dried. The crude compound was used for the next step without further purification. Yield 67%; 1H NMR (CDCl3) 2.29–3.31 (m, 4H, piperazine protons), 3.49–3.52 (m, 4H, piperazine protons), 4.12 (s, 2H, CH2), 6.73 (t, 2H, ar, J = 7.8 Hz).

Synthesis of 4-(7-amino-2-phenyl-2H-pyrazolo[4,3-d]pyrimidin-5-yl)piperazin-1-ium trifluoroacetate 64

The title compound was obtained by treatment of compound 21 (1.04 mmol) with trifluoroacetic acid, in the conditions described previously to prepare compound 60 from the corresponding N-Boc-derivative. The crude compound was used directly for the next step without purification. Yield 68%; 1H NMR (DMSO-d6) 3.22–3.29 (m, 4H, piperazine protons), 3.98–4.02 (m, 4H, piperazine protons), 7.51 (t, 1H, ar, J = 7.3 Hz), 7.63 (t, 2H, ar, J = 7.8 Hz), 8.03 (d, 2H, ar, J = 8.2 Hz), 8.69 (s, 1H, H-3), 9.25 (br s, 2H, NH2 ).

General procedure for the synthesis of 5-(4-acylpiperazin-1-yl)substituted-2H-pyrazolo[4,3-d]pyrimidin-7amines 23 and 24

A mixture of derivative 64 (0.98 mmol) and triethylamine (1.96 mmol) in anhydrous THF (20 ml) was stirred at room temperature for 1 h. Then, 3,3-dimethylbutiryl chloride (1.17 mmol) or phenylacetyl chloride (1.17 mmol) was added and the solution was stirred at room temperature for 5 h or 3 h, respectively. The mixture was diluted with water (15 ml) and extracted with EtOAc (20 X 3 ml). The organic phase was anhydrifed (Na2SO4) and the solvent evaporated at reduced pressure to give a solid which was taken up with cyclohexane.
and EtOAc, collected by filtration and purified by column chromatography (elucent CHCl₃/MeOH 10:0.5 for compound 23, MeOH for derivative 24).

7-Amino-2-phenyl-5-(4-(3,3-dimethylbutiryl)piperazin-1-yl)-2H-pyrazolo[4,3-d]pyrimidine 23

Yield 32%; m.p. 178–180°C. ¹H NMR (DMSO-d₆) 1.02 (s, 9H, t-But), 2.29 (s, 2H, CH₂), 3.58–3.60 (m, 4H, piperazine protons), 3.69–3.72 (m, 4H, piperazine protons), 7.42–7.46 (m, 3H, 1 ar þ NH₂) 7.59 (t, 2H, ar, J = 7.4 Hz), 8.00 (d, 2H, ar, J = 7.9 Hz), 8.60 (s, 1H, H-3). Anal. Calc. for C₂₁H₂₇N₇O.

7-Amino-2-phenyl-5-(4-phenylacetyl)piperazin-1-yl)-2H-pyrazolo[4,3-d]pyrimidine 24

Yield 69%; m.p. 207–209°C (CH₃NO₂). ¹H NMR (DMSO-d₆) 3.52–3.57 (m, 4H, piperazine protons), 3.60–3.66 (m, 4H, piperazine protons), 3.77 (s, 2H, CH₂), 7.21–7.34 (m, 5H, ar), 7.42–7.50 (m, 3H, 1 ar + NH₂), 7.57 (t, 2H, ar, J = 7.9 Hz), 7.98 (d, 2H, ar, J = 8.2 Hz), 8.59 (s, 1H, H-3). Anal. Calc. for C₂₃H₂₃N₇O.

5.2 Molecular modeling studies

5.2.1 Software overview

MOE suite (Molecular Operating Environment, version 2015.1001) [39] was used to perform most general molecular modeling operations.

Docking simulations were performed using the GOLD (Genetic Optimization for Ligand Docking, version 5.2) suite [40]. Quantum mechanical calculation of PM3 charges was carried out with the software MOPAC [41] as implemented in the MOE suite.

Analyses of docking poses in terms of energy calculation and visual inspection were executed taking advantage of the MOE suite.

Molecular modeling studies have been performed on a 8 CPU (Intel® Xeon® CPU E5-1620 3.70 GHz) linux workstation.

5.2.2 Three-dimensional structures of adenosine receptors

Among all the available crystallographic structures of hA₂A AR cocrystallized with a ligand in the orthosteric binding site, we opted for a complex with the antagonist ZM-241385 because of the structural similarity of its ([1,2,4]triazolo[1,5,a][1,3,5]triazin-5,7-yl)diamine scaffold with the (pyrazolo[4,3-d]pyrimidin-5,7-yl)diamine scaffold of the compounds under investigation. The crystallographic structure identified with 4EIY PDB code [42] was selected among all the structures co-crystallized with ZM-241385, because of its highest resolution (1.80 Å).

Since to date there are no crystallographic structures available for hA₃ and hA₁ ARs, we retrieved from the Adenosiland web-platform [43,44] previously developed by our research group, their homology models.
constructed using 4EIY structure as template. Those models were constructed in the presence of ZM-241385 as environment for induced fit, so the resulting structures consist in complexes between each AR subtype and the antagonist ZM-241385.

The residues are identified according to the generic Ballesteros Weinstein numbering system [45].

5.2.3 Molecular docking

Three-dimensional structures of ligands were built taking advantage of the MOE-Builder tool and ionization states were predicted using the MOE-Protonate-3D tool [46]. Ligand structures were subjected to MMFF94x energy minimization until the root mean square (rms) gradient fell below 0.05 kcal mol\(^{-1}\) Å\(^{-1}\). GOLD docking tool [40] was selected as conformational search program and GoldScore as scoring function, thanks to a docking benchmark study previously carried out in our laboratory [38,47]. For each compound, 10 docking runs were performed on each receptor subtype, searching in a sphere of 20 Å radius centered on the coordinates of the center of mass of ZM-241385 in complex with the receptor. Along with the compounds under investigation, docking simulations were conducted also for ZM-241385 as a reference example.

After computing atomic partial charges both of ligand poses, using PM3/ESP method, and receptors, using Amber10EHT force field, electrostatic and van der Waals contributions to the binding energy were calculated with MOE.

5.2.4 Interaction energy fingerprints (IEFs)

Individual electrostatic and hydrophobic interactions, hereinafter identified as IEele and IEhyd, respectively, were computed between ligand poses and each protein residue involved in binding [37,38]. Both these contributions were computed using MOE and, in particular, IEele were calculated as non-bonded electrostatic interactions energy term of the force field, so they are expressed in kcal/mol. Instead, IEhyd were computed as contact hydrophobic surfaces and are associated to an adimensional score (the higher the better). The data obtained by this analysis were reported in a graphic, called Interaction Energy Fingerprints (IEFs), representing residues (x-axis) in the form of equally high rectangles rendered according to a colorimetric scale. As regards IEele, colors from blue to red represent energy values ranging from negative to positive values; for IEhyd, colors from white to dark green depict scores going from 0 to positive values. More precisely, we retrieved the coordinates of the center of mass of ZM-241385 in the structure of each AR subtype complex. Only residues within 10 Å from this point were retained as belonging to the binding site, and plotted in the IEFs.
5.2.5 Interaction Energy Fingerprints comparison (IEFs comparison)

A new method has been introduced to evaluate docking results, which rests on the observation that ligands able to bind the same site of a protein often share a similar interaction pattern, too. The new method consists in the comparison of the IEFs of the pose of a candidate ligand (hereinafter called “docked”) with the IEFs of a ligand whose bound conformation is considered known (hereinafter called “reference”).

A quantitative estimation of the similarity of IEFs is computed as root mean square deviation (RMSD) between per residue interaction energies of the docked and the reference poses, both for electrostatic and hydrophobic interactions. This would inform about the average divergence of the docked from the reference: in particular a high RMSD value corresponds to large differences.

So far, there is no information about the direction of the divergence thus, along with RMSD, another analysis, named RMSDtrend, has been proposed. This consists of the sum of differences between per residue interaction energies of the docked and the reference, weighted by the number of residues of the binding site. A more favorable interaction energy profile would correspond to a negative RMSD\text{trend} in the case of electrostatic interactions, while to a positive one in the case of hydrophobic interactions.

In summary, low RMSD values, along with negative electrostatic RMSD\text{trend} and high hydrophobic RMSD\text{trend} could be interpreted as an indication of a higher “stability” of the docked pose respect the reference in the orthosteric binding state.

Moreover, this approach could be expanded to compare the behavior of the same ligand on different receptor subtypes, in order to have a preliminary “selectivity” profile based on the stabilities of the docked poses in their corresponding orthosteric binding states. In that case, RMSD and RMSD\text{trend} are computed for a docked compound against a reference on each receptor subtype. The reference compound should be a known good binder for each subtype and, at best, the crystallographic structure of the complex should be known.

In our case, ZM-241385 was chosen as reference compound, since it is a ligand for all ARs, having a Ki of 774 nM for hA\text{1} AR, of 1.6 nM for hA\text{2A} AR and of 743 nM for hA\text{3} AR. As regards the hA\text{2A} receptor, 4EIY crystallographic complex could be employed, while, for hA\text{1} and hA\text{3} ARs, we used the homology models, that are receptor-ZM-241385 complexes, since they were constructed considering ZM-241385 as environment for induced fit.

An additional graph was added, which allows to compare electrostatic and hydrophobic IEFs RMSDs and RMSD\text{trend} for different ligands on the different AR subtypes. RMSD and RMSD\text{trend} for the ligands (y-axis) on the various receptors (x-axis) were reported on a heat map, where they are represented by a colorimetric scale going from red to blue from unfavorable to favorable values. Finally, if a ligand presents blue rectangles...
on all receptors, it is expected to be “non-selective”, otherwise red and blue rectangles should describe lower and higher stability values, respectively, among the different receptor subtypes.

5.2.6 MMsDocking video maker

To facilitate the visualization and analysis of data obtained from the docking simulations, we have implemented a in-house tool, named MMsDocking video maker, for the automated production of a video that shows the most relevant docking data, such as docking poses, per residue IEhyd and IEele data, experimental binding data and scoring values. Videos were mounted using MEncoder [48] starting from images obtained with the following procedure: the heat maps in the background were drawn with GNUPlot 4.6 [49] starting from per residue IEhyd and IEele data computed with MOE. 2d depictions of compounds were generated using the open-source cheminformatics toolkit RDKit [50]. Representations of docking poses within the binding site were constructed using CHIMERA [51].

5.3 Pharmacological assays

5.3.1 Human cloned A1, A2A and A3 AR binding assay

All synthesized compounds were tested to evaluate their affinity at human A1, A2A and A3 ARs. Displacement experiments of [3H]DPCPX (1 nM) to hA1 CHO membranes (50 mg of protein/assay) and at least 6–8 different concentrations of antagonists for 120 min at 25°C in 50 mM Tris-HCl buffer pH 7.4 were performed [52]. Non-specific binding was determined in the presence 1 µM of DPCPX (≤10% of the total binding). Binding of [3H]ZM-241385 (1 nM) to hA2A CHO membranes (50 µg of protein/assay) was performed by using 50 mM Tris-HCl buffer, 10 mM MgCl2 pH 7.4 and at least 6–8 different concentrations of antagonists studied for an incubation time of 60 min at 4°C [53]. Non-specific binding was determined in the presence of 1 µM ZM-241385 and was about 20% of total binding. Competition binding experiments to hA3 CHO membranes (50 µg of protein/assay) were performed incubating 0.5 nM [125I]AB-MECA, 50 mM Tris-HCl buffer, 10 mM MgCl2, 1 mM EDTA, pH 7.4 and at least 6–8 different concentrations of examined ligands for 60 min at 37°C [54]. Non-specific binding was defined as binding in the presence of 1 µM AB-MECA and was about 20% of total binding. Bound and free radioactivity were separated by filtering the assay mixture through Whatman GF/B glass fiber filters by using a Brandel cell harvester. The filter bound radioactivity was counted by Scintillation Counter Packard Tri Carb 2810 TR with an efficiency of 58%.

5.3.2 Measurement of cyclic AMP levels in CHO cells transfected with hA2B AR

CHO cells transfected with hA2B AR subtypes were washed with phosphate-buffered saline, diluted trypsin and centrifuged for 10 min at 200 g. The cells (1 X 106 cells/assay) were suspended in 0.5 ml of incubation mixture (mM): NaCl 15, KCl 0.27, NaH2PO4 0.037, MgSO4 0.1, CaCl2 0.1, Hepes 0.01, MgCl2 1, glucose 0.5, pH 7.4 at 37°C, 2 IU/ml adenosine deaminase and 4–(3-butoxy-4-methoxybenzyl)-2-imidazolidinone (Ro 20–
1724) as phosphodiesterase inhibitor and preincubated for 10 min in a shaking bath at 37°C. The potency of antagonists to the A2B AR was determined by the inhibition of NECA (200 nM)-induced cyclic AMP production [55]. The reaction was terminated by the addition of cold 6% trichloroacetic acid (TCA). The TCA suspension was centrifuged at 2000 g for 10 min at 4°C and the supernatant was extracted four times with water saturated diethyl ether. The final aqueous solution was tested for cyclic AMP levels by a competition protein binding assay. Samples of cyclic AMP standard (0–10 pmoles) were added to each test tube containing [3H] cyclic AMP and incubation buffer (trizma base 0.1 M, aminophylline 8.0 mM, 2-mercaptoethanol 6.0 mM, pH 7.4). The binding protein prepared from beef adrenals was added to the samples previously incubated at 4°C for 150 min, and, after the addition of charcoal, was centrifuged at 2000 g for 10 min. The clear supernatant was counted in a Scintillation Counter Packard Tri Carb 2810 TR with an efficiency of 58%.

5.3.3 Data analysis

The protein concentration was determined according to a Bio-Rad method [56] with bovine albumin as a standard reference. Inhibitory binding constant (Ki) values were calculated from those of IC50 according to Cheng & Prusoff equation \( Ki = IC_{50}/(1 + [C^*]/K_{D^*}) \), where \([C^*]\) is the concentration of the radioligand and \(K_{D^*}\) its dissociation constant [57]. A weighted non-linear least-squares curve fitting program LIGAND [58] was used for computer analysis of inhibition experiments. IC50 values obtained in cyclic AMP assay were calculated by non-linear regression analysis using the equation for a sigmoid concentration–response curve (Graph-PAD Prism, San Diego, CA).
References


40. GOLD suite, version 5.2. Cambridge Crystallographic Data Centre: 12 Union Road, Cambridge CB2 1EZ, UK. http://www.ccdc.cam.ac.uk


57. Cheng Y, Prusoff WH (1973) Relationship between the inhibition constant (K1) and the concentration of inhibitor which causes 50 per cent inhibition (I50) of an enzymatic reaction. Biochem Pharmacol 22:3099–3108

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 when aiming to interpret the mechanism of known active substances as well as to develop novel active candidates. Unfortunately, simulating the binding process is still a challenging task because it requires classical MD experiments in a long microsecond time scale that are 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), and successfully applied it to 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 present an extension of the SuMD application domain including different types of proteins in comparison with GPCRs. In particular, we have deeply analyzed the ligand–protein recognition pathways of six different case studies that we grouped into two different classes: globular and membrane proteins. Moreover, we introduce the SuMD-Analyzer tool that we have specifically implemented to help the user in the analysis of the SuMD trajectories. Finally, we emphasize the limit of the SuMD applicability domain as well as its strengths in analyzing the complexity of ligand–protein recognition pathways.

1. Introduction

The essential features of ligand–protein interaction are very often summarized under the expression “molecular recognition” incorporating both thermodynamic aspects (quantified by $K_d$, the equilibrium dissociation constant) and kinetic aspects (reflected by the rate constants $k_{on}$ and $k_{off}$) of ligand binding. 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 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 physicochemical bases governing the optimization of thermodynamic aspects of ligand binding are relatively well acknowledged, but unluckily, they remain still poorly understood for binding kinetics. In fact,
the $K_d$ 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 a transition state is 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 the target 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 time scale, 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 reproducing a ligand–protein binding or unbinding event on an accessible time scale 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 us to explore the ligand–protein approaching path in a 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 exploits a tabu-like algorithm to monitor the distance between the center of masses of the ligand atoms and the protein binding site in standard short 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**

In the present work, we present an extension of the SuMD application domain to types of proteins beyond GPCRs. In particular, we deeply analyzed the ligand–protein recognition pathways of six different case studies that we grouped into two different classes: globular and membrane proteins (Table 1). Moreover, we introduce the SuMD-Analyzer tool that we have specifically implemented to help, also nonexpert users, in the analysis of the SuMD trajectories.
**Table 1** Structural Summary of Selected Ligand–Protein Complexes.

<table>
<thead>
<tr>
<th>PDB</th>
<th>Protein</th>
<th>Ligand</th>
<th>Resolution [Å]</th>
<th>Affinity</th>
<th>Ligand MW</th>
<th>ref</th>
</tr>
</thead>
<tbody>
<tr>
<td>2ZJW</td>
<td>CK2</td>
<td>Ellagic Acid</td>
<td>2.40</td>
<td>Ki = 0.04 µM</td>
<td>302.197</td>
<td>41</td>
</tr>
<tr>
<td>13GS</td>
<td>GSTP1-1</td>
<td>SASP</td>
<td>1.90</td>
<td>Ki = 24 µM</td>
<td>398.39</td>
<td>44</td>
</tr>
<tr>
<td>4K7I</td>
<td>PRDX5</td>
<td>Benzen-1,2-diol</td>
<td>2.25</td>
<td>KD = 1500 µM</td>
<td>110.11</td>
<td>45</td>
</tr>
<tr>
<td>2VDB</td>
<td>HSA (S)</td>
<td>(S)-naproxen</td>
<td>2.52</td>
<td>Ka = 1.2 1.8 µM⁻¹</td>
<td>230.25</td>
<td>49,59</td>
</tr>
</tbody>
</table>

**transmembrane systems**

<table>
<thead>
<tr>
<th>PDB</th>
<th>Protein</th>
<th>Ligand</th>
<th>Resolution [Å]</th>
<th>Affinity</th>
<th>Ligand MW</th>
<th>ref</th>
</tr>
</thead>
<tbody>
<tr>
<td>3GWW</td>
<td>LeuT</td>
<td>(S)-fluoxetine</td>
<td>2.46</td>
<td>IC₅₀ = 355 mM</td>
<td>345.79</td>
<td>51</td>
</tr>
<tr>
<td>2YDV</td>
<td>hA₂AR</td>
<td>NECA</td>
<td>2.60</td>
<td>Ki = 13.8 nM</td>
<td>308.29</td>
<td>55</td>
</tr>
</tbody>
</table>

*a* In the affinity column, different constants were reported according the method by which they were inferred; $K_i$, $K_d$, and $K_a$ correspond, respectively, to the inhibition, dissociation, and association constants. For the complex 3GWW, only the half maximal inhibitory concentration (IC₅₀) was experimentally available.

2. Materials and methods

2.1 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 parametrization, and (iii) solvated system setup and equilibration. Two different protocols based on AMBER12 [13]/general Amber force field (GAFF) [14] and the CHARMM27 [15]/CHARMM general force field (CGenFF) force fields combinations were adopted for globular and transmembrane systems, respectively [16, 17].

2.2 Systems Preparation

Protein–ligand complexes were retrieved from the RCSB PDB database [18]. Protein 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 CHARMM27 [15] force field topology. Missing loops were modeled by the default homology modeling 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 atoms.
2.3 Ligand Parametrization

2.3.1 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] (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.

2.3.2 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].

2.4 Solvated System Setup and Equilibration

2.4.1 Globular Systems

Protein–ligand complexes were assembled with the tleap tool using AMBER14SB [29] as the 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 the water model [30]. To neutralize the total charge, Na+/Cl− counterions were added to a final salt concentration of 0.150 M. The systems were energy minimized by 2000 steps with the conjugate-gradient method and then 50,000 steps of NVE (100 ps) followed by 1 ns of NPT simulation, both using a 2 fs as the time step and applying a 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 nonbonded cutoff distance of 9 Å with a switching distance of 7.5 Å was used.

2.4.2 Transmembrane Systems

Transmembrane proteins were embedded in a 1-palmitoyl-2-oleoyl-snglycerol-3-phosphocholine (POPC) lipid bilayer according to the suggested orientation reported in the Orientations of Proteins in Membranes (OPM) database [35]. Initial POPC atoms were placed through the VMD [28] membrane builder plugin, and lipids within 0.6 Å from amino acid atoms were removed. The systems were solvated with TIP3P [30] water using
the program Solvate 1.0 [36] and neutralized by Na\textsuperscript{+}/Cl\textsuperscript{−} counterions to a final concentration of 0.154 M. The systems were then equilibrated through a two-step procedure: In the first stage, after 2000 cycles of a 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\textsuperscript{−1} Å\textsuperscript{−2}. The temperature was maintained at 310 K using a Langevin thermostat with a low damping constant of 1 ps\textsuperscript{−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 time step 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\textsuperscript{−1} Å\textsuperscript{−2} for the next 10 ns of MD simulation.

2.5 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.

2.5.1 SuMD Input Files

SuMD requires a configuration file (selection.dat, Figure S1A) 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 three-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: 31) 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, and (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 (Section 2.5.2) of nonsupervised 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.

2.5.2 SuMD Main Code

The workflow of the SuMD main code is reported in Figure 2. 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 \(d_{\text{cm}}(L,R)\) that will be monitored. Then, a series of 600 ps classical MD simulations are
performed. This is a crucial parameter that we have empirically set up to guarantee to a ligand to significantly translate its center of mass during this short time of conventional MD simulation. After each simulation, five dcm\_(L-R) distance points are collected at regular intervals of 75 ps. 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.

2.5.3 SuMD Log File

At each SuMD simulation step, a log file (Figure S1B) is updated collecting information about (i) the step number, (ii) the dcm\_(L-R) distance, (iii) the slope value (m), and (iv) the electrostatic and van der Waals potential energy contributions of the ligand–receptor interaction energy (IE). A counter keeps track of how
many times each SuMD step has been attempted. Furthermore, three counters corresponding to the \(d_{cm(L-R)}\) distance ranges 0–2, 2–5, and 5–9 Å are reported. These distances monitor how many times the binding site is approached, i.e., how often the \(d_{cm(L-R)}\) 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.

2.5.4 SuMD Termination Criteria

A SuMD simulation is stopped when one of the following four counters reaches the default maximum value:

(i) Counter 1 is incremented when the step is not productive, otherwise it is restored to zero (default maximum value: 17 steps correspond to 10.2 consecutive ns).

(ii) Counter 2 monitors the \(d_{cm(L-R)}\), and its value is incremented when the distance is between 9 and 5 Å (default maximum value: 19 corresponding to 11.4 nonconsecutive ns).

(iii) Counter 3 monitors the \(d_{cm(L-R)}\), and its value is incremented when the distance is between 5 and 2 Å (default maximum value: 19 corresponding to 11.4 nonconsecutive ns).

(iv) Counter 4 monitors the \(d_{cm(L-R)}\), and its value is incremented when the distance is lower than 2 Å (default maximum value: 19 corresponding to 11.4 nonconsecutive ns).

2.6 SuMD-Analyzer Tool

The SuMD-Analyzer is a plugin written in python, tcl, and bash to analyze the SuMD trajectories (Figure 3). 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.

![Fig. 3 Overview of the analyses provided by SuMD-Analyzer. The analysis are colored based on the class they belong to. Green: ligand position analysis. Blue: interaction energy analysis. Magenta: per residue interaction. Violet: per replicas analysis.](image-url)
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 five through the VMD animate module.

2.6.1 Ligand Position

Two analyses follow the coordinates explored by the ligand during the SuMD trajectory (Figure 3, 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 3A). The Pollicino analysis is a simplified representation of the trajectory, highlighting only the most relevant phases of the recognition pathway explored by the ligand. The analysis collects the last frame of each SuMD simulation step (one point each 600 ps) and clusters the coordinates of the ligand mass center according to the corresponding dcm(L-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. According to this procedure, each sphere can collect states arisen from different moments of the trajectory. Arrows indicate the chronological order onto which the regions where the sphere reside are approached by the ligand mass center (Figure 3B).

2.6.2 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 3, 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 dcm(L-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 3C). 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 3D).

2.6.3 Per Residue Interactions

A further set of analyses was developed to highlight the most important residues involved in the ligand recognition pathway (Figure 3, magenta boxes): (i) the “protein–ligand contacts count” and (ii) the “ligand–protein recognition map”. In the first graph (Figure 3E), 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 3F), the residues 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 than 4 Å from ligand atoms, while green dots correspond to a contact event. 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 UCSF Chimera [38] (Figure 3G,H). 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 3G), the ribbons are colored according the so-derived B-factor values. A similar representation, “chimera_time” (Figure 3H), 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.

2.6.4 Replicas Analysis

The replica analysis (Figure 3, violet box) tool integrated in the SuMD-Analyzer was developed taking advantage of VMD and Chimera as graphical visualization tools. The tool collects the raw data from all the simulated replicas and merges the data for the above-described plots and graphical representations.

3. Results and Discussion

3.1 Case Studies Selection

As already anticipated, in this work, the 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-pyridylamino)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 A2A receptor (hA2A 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 4 and briefly described in the following.
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 processes 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 the “hinge region”. The N-terminal lobe presents five β-strands, and the α-helix C is involved in the substrate recognition, whereas the C-terminal lobe is composed of α-helices. All PKs present a glycine-rich loop (Ploop), an activation loop, and a catalytic loop [40]. The X-ray complex highlights that the inhibitor binds to Lys49, Ser51, and His160 as shown in Figure 4A [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 toward numerous anticancer drugs [43]. SASP, which is able to inhibit GSTs without acting as a cosubstrate for the conjugation reaction with GSH, has been cocrystallized with GSTP1-1 and represents a starting point for structure-based design of new anticancer drugs [44]. The X-ray complex (Figure 4B) highlights that the inhibitor binds to a hydrophobic pocket formed by Phe8, Val10, Val35, Ile10,4 and Tyr108 side-chains. The 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 carboxylate group of the ligand 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 whose role relies on the hydrogen peroxide and alkyl hydroperoxides reduction. PRDX5 plays a remarkable role in postischemic 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 ($K_d = 1.5 \pm 0.5$ mM). 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 4C), 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 side chain of Arg127 is oriented toward the hydroxyl moiety and contributes to the binding with an additional hydrogen bond. Similarly, the thiol group of Cys47 is faced to the catechol. Pro40, Leu116, and Phe120 establish hydrophobic interactions with the aromatic ring.

Human serum albumin (HSA) is a deeply investigated protein for its ability to 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 the 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 ligands 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 4D). The carboxylic group is partially exposed to the solvent but is surrounded by several charged residues forming the entrance of the pocket: Arg145, Lys 190, and in particular, Arg186.

The 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 the recent past, the crystallographic structure of the LeuT from *Aquifex aeolicus* (a NSS family member) has been disclosed with the aim of better understanding the basis of selective serotonin reuptake inhibitors (SSRIs) activity toward serotonin transporters [51]. The LeuT-(S)-fluoxetine X-ray complex (Figure 4E) highlights hydrophobic contacts between the inhibitor and Leu29, Arg30, Tyr108, and Phe253 side chains. The (S)-fluoxetine secondary amino group points toward the extracellular space and engages 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 GPCRs superfamily. The known four subtypes, termed adenosine $A_1$, $A_2A$, $A_2B$, and $A_3$ receptors, are widely distributed in the human body, involved in several physio-pathological processes, and represent potential targets for the treatment of several...
In diseases [52]. In the past decade, X-ray structures of the hA2A 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 4F) 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.

A summary of the SuMD simulation performed on the selected case studies is reported in Table 2. For each replica, the productive simulation time is reported, and for the simulation in which the ligand reached the binding site (dcm(L-R) < 5 Å), a brief statistical analysis about the comparison with the final state and the experimentally solved structure is reported.

<table>
<thead>
<tr>
<th>Table 2 SuMD Results Summary</th>
<th>System</th>
<th>Replica</th>
<th>Time [ns]</th>
<th>RMSDmin [Å]</th>
<th>RMSDmax [Å]</th>
<th>RMSDave [Å]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ellagic Acid</td>
<td>1</td>
<td>25.21</td>
<td>4.85</td>
<td>6.1</td>
<td>5.39</td>
</tr>
<tr>
<td></td>
<td>Ellagic Acid</td>
<td>2</td>
<td>19.81</td>
<td>4.62</td>
<td>6.55</td>
<td>5.25</td>
</tr>
<tr>
<td></td>
<td>Ellagic Acid</td>
<td>3</td>
<td>24.01</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SASP – GSTP1-1</td>
<td>1</td>
<td>21.01</td>
<td>10.08</td>
<td>13.87</td>
<td>11.76</td>
</tr>
<tr>
<td></td>
<td>SASP – GSTP1-1</td>
<td>2</td>
<td>27.01</td>
<td>2.04</td>
<td>7.98</td>
<td>5.74</td>
</tr>
<tr>
<td></td>
<td>SASP – GSTP1-1</td>
<td>3</td>
<td>19.21</td>
<td>2.07</td>
<td>7.55</td>
<td>5.3</td>
</tr>
<tr>
<td></td>
<td>Benzen-1,2-diol</td>
<td>1</td>
<td>17.41</td>
<td>1.25</td>
<td>5.88</td>
<td>3.04</td>
</tr>
<tr>
<td></td>
<td>Benzen-1,2-diol</td>
<td>2</td>
<td>31.21</td>
<td>1.12</td>
<td>7.35</td>
<td>2.99</td>
</tr>
<tr>
<td></td>
<td>Benzen-1,2-diol</td>
<td>3</td>
<td>18.01</td>
<td>0.69</td>
<td>5.95</td>
<td>2.98</td>
</tr>
<tr>
<td></td>
<td>(S)-naproxen</td>
<td>1</td>
<td>13.21</td>
<td>4.77</td>
<td>11.99</td>
<td>7.54</td>
</tr>
<tr>
<td></td>
<td>(S)-naproxen</td>
<td>2</td>
<td>26.41</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(S)-naproxen</td>
<td>3</td>
<td>27.01</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(S)-fluoxetine</td>
<td>1</td>
<td>19.21</td>
<td>7.65</td>
<td>8.46</td>
<td>8.13</td>
</tr>
<tr>
<td></td>
<td>(S)-fluoxetine</td>
<td>2</td>
<td>18.61</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(S)-fluoxetine</td>
<td>3</td>
<td>12.01</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>NECA – hA2A AR</td>
<td>1</td>
<td>27.61</td>
<td>4.41</td>
<td>7.59</td>
<td>5.98</td>
</tr>
<tr>
<td></td>
<td>NECA – hA2A AR</td>
<td>2</td>
<td>22.21</td>
<td>4.35</td>
<td>7.67</td>
<td>5.59</td>
</tr>
<tr>
<td></td>
<td>NECA – hA2A AR</td>
<td>3</td>
<td>27.01</td>
<td>4.44</td>
<td>6.44</td>
<td>5.34</td>
</tr>
</tbody>
</table>

*For each system, the result of three replicas are reported within their simulation time. The RMSD minimum, maximum, and average are computed considering only the frames that present a dcm(L-R) lower than 5 Å.
3.2 Globular Systems

3.2.1 Ellagic Acid–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 5A 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 by Lys49 that directs the ligand to the P-loop of the kinase. As shown by the Pollicino analysis (Figure 5B), the ellagic acid approaches the region of the P-loop and mostly interacts with the Arg47, Lys49, Glu53, and the Lys71 (Figure 5A). These residues describe an interaction site at 10.5 Å where the ligand resides for about 6 ns. In fact, the ligand RMSD plot (Figure 5C) 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 5E) highlights that the above-mentioned residues are those establishing the greatest number of contacts, whereas the corresponding 3D models helps in identifying their location (Figure 5F) and the chronological order at which they have been approached by the ligand (Figure 6A). 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 a new interaction site described by Asp120, Arg47, and Met163. The permanency in this site is about of 2 ns with an IE of –51 kcal/mol (Figure 5C,D). Consistently, the RMSD plot presents another plateau in the time range of 8–10 ns (Figure 5C) that corresponds to the swarm of dots in the IE Landscape at dcm\(_{(L,R)}\) = 11 Å (Figure 5D). 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). As shown in the cumulative ligand–protein IE (Figure 6B) and its corresponding decomposition into electrostatic and van de Waal contributions (Figure 6C 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 contributions (Figure 6C and D, respectively), the stabilization can be ascribed by the establishment of an electrostatic interaction with Asp175. As result of the new interaction, the ligand moves into the orthosteric site (Figure 6E) and interacts with Lys159, Val66, Val116, Val53, His115, and Lys68 by maintaining the same position until 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 and –70 kcal/mol.
The simulation was replicated three times, and the replicas analysis results are reported in Figure 7. In particular, the RMSD plot indicates that one replica does not reach the orthosteric site (Figure 7A, green line), whereas the others reach the same final RMSD value. The same conclusion arises from the investigation of the Pollicino analysis where the ligand pathway of the two replicas converge in the proximity of the protein (Figure 7B, 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 7C, the ligand reaches a different interaction site with an IE of ~60 kcal/mol, a value close to the IE of the replicas that converges into in the orthosteric site (Figure 7C, 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 7D, green line), indicating a very strong interaction.
Fig. 6 Ellagic acid–CK2 recognition pathway: (A) chimera time, B) cumulative IE, (C) cumulative IE electrostatic contribution, (D) cumulative IE van der Waals contribution, and (E) superimposition between SuMD endpoint conformation (gold) and the X-ray binding mode (green). The residues interacting with the ligand are labeled in black, except the ones detected only in the X-ray complex that are labeled in green.

Fig. 7 Ellagic acid–CK2 recognition pathway: (A) per replica ligand RMSD, (B) per replica Pollicino analysis, (C) per replica IE landscape, and (D) per replica cumulative IE.
In order to compare the role of the supervision in reducing the computation time, we performed 1 μs of classical MD simulation using the same starting geometry of the SuMD simulation (Video S7) in which the ligand was placed at a distance of 50 Å from the binding site. As expected, during the classical simulation, the ligand did not approach the protein in agreement with the results previously obtained also for other systems [7].

3.2.2 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 8A,B). 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 8C).

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 8A).

The energy stabilization of the complex increases, and after 8 ns of simulation, SASP proceeds toward a farther conformation, able to gain a more favorable 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 8A,B) is retained until the end of SuMD simulation, with the exception of conformational changes occurring to the pyridylsulfamoyl moiety, able to fit in the hydrophobic pocket delimited by Phe8, Val35, and Trp38. During the SASP–GSTP1-1 recognition event, GSH remains in the catalytic G site of the enzyme, not interacting with the inhibitor.

Figure 8D highlights all the residues involved in the interaction with SASP during the SuMD simulations; the selective contacts toward 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 a geometrical reference, the ligand RMSD analysis (Figure 8E) reaches a minimum after 15 ns of simulation (Figure 8F) before stabilizing around a value of about 5 Å. Figure S2 reports other ligand–protein IE analyses. The replicas analysis (Figure S3) depicts a recognition event with no metastable binding sites, 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.
3.2.3 Benzen-1,2-diol–PRDX5 Recognition Pathway

The simulations were repeated on both the monomeric and dimeric forms yielding similar results. However, here we 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 PRDX5 binding site (dcm_{L-R} = 78 Å). As reported in Figure 9A (point b), B, and C, 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 9A and B, respectively), this site engages the ligand in 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 through a series of molecular interactions, including residues (chronologically sorted) Glu91, Glu16, Glu18, and Phe79 belonging to the first monomer unit (Figure S4). Finally, the fragment accesses the binding site where it 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 (Figure S4). During the fluctuation, the catechol contacts most of the residue forming the site, in particular (sorted by number of molecular contacts during the trajectory) Thr147, Thr44, Arg127, Phe120, Leu116, Gly46, and Cys47 (Figure 9C,D). The main conformation observed corresponds to the crystallographic one, as reported in Figure 9E and F where the RMSD reaches a minimum value 0.69 Å at 17.3 ns.

The simulation was repeated in triplicate 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 of 1.12 and 1.24 Å for replica 2 and 3, respectively.

**Fig. 9** Benzen-1,2-diol–PRDX5 recognition pathway: (A) landscape, (B) Pollicino analysis, (C) ligand–protein recognition map, (D) chimera, (E) ligand–RMSD, and (F) superimposition between SuMD endpoint conformation (gold) and the X-ray binding mode (green). The residues interacting with the ligand are reported.
3.2.4 (S)-Naproxen–HAS Recognition Pathway

The SuMD simulation was performed maintaining decanoic ligand in the IB site according to the crystallographic geometries. (S)-Naproxen was separated from the HSA-decanoid acid complex by placing it 32 Å far from the IB site (point a, Figure 10A,B). In the first SuMD step, the ligand fluctuates until 50 Å from the IB site. As reported in Figure 10C, after a couple of nanoseconds, the ligand approaches the first protein site by engaging Lys510 and Thr564 (Video S4). Shortly after, the ligand establishes a network of interaction for 1 ns (from 2.3 to 3.2 ns) in a site located at around dcm\(_{L-R}\) = 20 Å (point b, Figure 10A), defined by residues Val116, Pro118, Val122, Thr133, and Phe134. Then, the molecule approaches a second site, 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). 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, and Tyr161) as shown by Figure 10B and E. Within the primary site, the ligand is able to place the methyl ether group in the proximity of Phe157 with an orientation very similarly to the one observed in the crystal structure. On the other hand, the naphthalene core and, in particular, the carboxylic group adopts a different position due to the presence of the extended loop. This different orientation abolishes the ionic interaction between the carboxyl group and the Arg117 observed in the crystallographic structure (Figure 10F). At the end of the simulation, the RMSD fluctuates around 5 Å, reaching the lowest value of 4.76 at 12.70 ns (Figure 10E,F).

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 that the loop might have a crucial role in the recognition process (Figure S7).
3.3 Transmembrane Systems

3.3.1 (S)-Fluoxetine–LeuT Recognition Pathway

The (S)-fluoxetine recognition pathway highlights, after 1 ns of SuMD simulation, a first electrostatic interaction between the Asp 158 side chain and the charged secondary amine group of the ligand (Video S5). The energetic stabilization characterizing this complex corresponds to the IE landscape minimum reported in Figure 11A (point a) and B. This preliminary complex is able to favor the ligand approach toward an inner pocket of LeuT, topologically defined by Tyr471 and the aliphatic chains of Lys474 and Glu478, 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 with the Glu402 side chain.
This scenario anticipates the ligand repositioning inside an inner hydrophobic site, where the ligand engages for almost 7 ns Tyr471, Trp406, Ile475, and Phe405 side chains in lipophilic interactions through its phenyl ring (point b, Figure 11A,B).

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 11A,B, and continuous lines corresponding to the last 4 ns of SuMD simulation, Figure 11C), both located at the protein extracellular gate and involved in an ionic lock that sterically obstructs the SSRIs binding site disclosed by the LeuT crystallographic structure. Figure 11D summarizes all the amino acids involved in the (S)-fluoxetine recognition event during the SuMD simulation.

The RMSD plot (Figure 11E) outlines the inhibitor difficulty in reproducing the experimental pose (Figure 11F). Investigation of the LeuT crystal structure without a cocrystallized inhibitor reveals an alternative conformation of the Arg30 side chain and the absence of the ionic lock (Figure S12) [57]; it is possible to speculate that the extracellular gate in LeuT, during the SuMD simulation time scale, 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 preparation of the system for SuMD simulations.

Replicas analysis (Figure S9) highlights two alternative recognition pathways through the extracellular vestibule that do not reproduce the binding mode observed in the crystallographic complex and are characterized by accentuated energy variations in proximity of the extracellular transporter gate.

3.3.2 NECA–hA2A AR Recognition Pathway

NECA establishes the first stabilizing contacts with the hA2A AR after about 4 ns of SuMD simulation (Video S6). During this initial scenario (point a, Figure 12A,B), the ligand approaches the protein topological structure defined by ECL2, the N-terminus, and the residues located at top of TM5 and TM6. More precisely, NECA engages the Phe257 (TM6) side chain in a π–π stacking interaction through its purine scaffold and locates the N-ethylcarboxamido moiety toward a pocket delimited by Trp143 (ECL2), Pro173 (ECL2), and Asn175 (TM5) side chains, as highlighted by the first stripes in Figure 12C and the yellow and violet ribbons in Figure 12D.

Fig. 12 NECA–hA2A AR recognition pathway: (A) landscape, (B) Pollicino analysis, (C) ligand–protein recognition map, (D) chimera, (E) ligand–RMSD, and (F) superimposition between SuMD endpoint conformation (gold) and X-ray binding mode (green). The residues interacting with the ligand are labeled in black, except the ones detected only in the X-ray complex that are labeled in green.
This complex anticipates a repositioning that allows the ligand to reach a meta-stable 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 a hydrogen bond interaction between its C2’ hydroxide group and Asn253 (TM6) (point b, Figure 12A,B).

During the time slot rising from 14 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 it engages the Phe168 (ECL2) side chain in a π–π stacking interaction and an Asn253 (TM6) side chain in hydrogen bond interactions through its exocyclic amine and the N7 position of the purine scaffold (point c, Figures 12A,B). This complex orientation (associated with the minimum RMSD value in Figure 12E, with respect to the NECA crystallographic conformation) is followed by an alternative stabilized conformation (point d, Figure 12A,B) which involves also hydrophobic interactions with Leu249 (TM6), Leu85 (TM3), and Val84 (TM3). Data from mutagenesis experiments confirm the involvement of some residues highlighted by the SuMD simulations. More precisely, there is strong evidence about the recruitment of Phe257, Asn253, and Phe168 side chains during agonists recognition [58].

During the remaining SuMD simulation time, the protein–ligand complex geometry remains 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 S10, other ligand–protein energy interaction analyses are reported.

At the minimum RMSD value, NECA pyrimidine scaffold coincides with the crystallographic orientation, while the ribose moiety is oriented in an alternative conformation (Figure 12F).

Replicas analysis (Figure S11) highlights also a different NECA recognition pathway, which involves residues located at the ECL2 and characterized by comparable energetic stabilizations.

4. Conclusions

In the present work, we have demonstrated the general applicability of SuMD simulations using different types of targets, including both globular and membrane proteins. Moreover, we have presented the SuMD-Analyzer tool that helps, also a nonexpert 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 the nanosecond simulation time scale. Furthermore, SuMD simulations enable the investigation of ligand–protein binding events independently from the starting position and 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). This 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 the expectations, not all SuMD trajectories converge to the structure of the complex obtained by X-ray crystallography. Indeed, there are several plausible reasons that may be argued to describe this particular unexpected behavior: (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 with respect to its apo-form. This could be interpreted as the sign of an important induce-fit process during the ligand recognition. (c) The boundary conditions that led to the formation of the crystallographically ligand–protein complex (solvent and cosolvent, 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 making any conjecture 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.

One of the key aspects is the notably reduction of the time needed to obtain a SuMD trajectory in comparison to classical MD; the computation time is in the range from a few hours to tens of hours for the presented case studies. Thanks to these performances, a second effort will be addressed to extend the number of replicas with the aim to investigate the convergence of the pathway in sampling a bigger number of states. 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 to expand our perspectives in several scientific areas from molecular pharmacology to drug discovery.
References


42. Wilce MCJ, Parker MW (1994) Structure and function of glutathione S-transferases. Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology 1205:1–18


SCIENTIFIC PUBLICATIONS


Exploring Protein-Peptide Recognition Pathways Using a Supervised Molecular Dynamics Approach

Veronica Salmaso, Mattia Sturlese, Alberto Cuzzolin, and Stefano Moro


Abstract

Peptides have gained increased interest as therapeutic agents during recent years. The high specificity and relatively low toxicity of peptide drugs derive from their extremely tight binding to their targets. Indeed, understanding the molecular mechanism of protein-peptide recognition has important implications in the fields of biology, medicine, and pharmaceutical sciences. Even if crystallography and nuclear magnetic resonance are offering valuable atomic insights into the assembling of the protein-peptide complexes, the mechanism of their recognition and binding events remains largely unclear. In this work we report, for the first time, the use of a supervised molecular dynamics approach to explore the possible protein-peptide binding pathways within a timescale reduced up to three orders of magnitude compared with classical molecular dynamics. The better and faster understating of the protein-peptide recognition pathways could be very beneficial in enlarging the applicability of peptide-based drug design approaches in several biotechnological and pharmaceutical fields.

1. Introduction

Protein-peptide recognition has a crucial role in various fundamental aspects of cellular homeostasis, such as signal transduction, protein-trafficking, and immune response. Moreover, protein-peptide recognition has an important impact on various biotechnological and pharmaceutical applications, such as peptide-based therapeutics, biosensors, biomarkers, and functional modulators of proteins. In particular, nowadays peptide-based drug discovery could be a serious option for addressing new therapeutic challenges [1-3]. In fact, novel chemical strategies for limiting metabolism and alternative routes of administration have emerged in recent years and resulted in an increasing number of peptide-based drugs that are now being marketed [1-3].

Understanding the molecular mechanism of protein-peptide recognition has, and surely will have even more in the future, important applications in the fields of biology, medicine, and pharmaceutical sciences. High-resolution structure determination methods, such as X-ray crystallography and nuclear magnetic resonance, are offering valuable atomic insights into the assembling of the protein-peptide complexes. However, the molecular mechanism of the recognition and binding events that occur between the bound and unbound
states remains largely unclear. Computational modeling offers the opportunity to directly inspect the binding event and understand the key features of protein-peptide recognition. Molecular docking and molecular dynamic (MD) simulations have already been proposed as suitable strategies to explore protein-peptide interactions [4-15]. Unfortunately, the whole recognition process from the unbound to the bound state 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 simulation in a long microsecond timescale. Recently, we have overcome this limiting factor by implementing an alternative MD approach, named supervised molecular dynamics (SuMD), that notably speeds up the complete simulation of a protein-ligand recognition process compared with classical MD [16-17]. As described previously, SuMD enables the investigation of ligand-receptor binding events independently from the starting position, from the chemical structure of the ligand, and also from its receptor binding affinity.

Starting from the original implementation of SuMD, for the first time, we have extended the applicability domain of SuMD toward the exploration of the protein-peptide recognition pathway (pepSuMD) within a timescale reduced up to three orders of magnitude compared with classical MD. In particular, to evaluate the performance and robustness of pepSuMD, three well-renowned complexes were selected from a subset of characterized protein-protein interaction targets [18], as pilot key studies: two of them containing natural peptides (Bcl-XL/BAD and MDM2/p53) [19, 20] and one containing a stapled peptidomimetic (MDM2/SAH-p53-8) [21], as summarized in Table 1. Both Bcl-XL/BAD and MDM2/p53 complexes play a key role in the regulation of the apoptotic pathway. In fact, one of the ways that cancer cells can evade physiological cellular regulation and chemotherapeutic-induced cell death is by overexpression of pro-survival proteins, such as Bcl-XL and MDM2, or by amplification of their genes. The BAD protein disrupts the heterodimer normally formed by Bcl-XL with specific pro-apoptotic proteins such as BAK and BAX [22]. Similarly, MDM2 exerts its oncogenic effects primarily by interacting with the p53 tumor suppressor protein [23].

The preliminary results collected in this pilot key study are very encouraging. In fact, pepSuMD methodology allows the simulation of the whole process of protein-peptide recognition (from the unbound to the bound state) in a nanosecond timescale, with an appreciable capability to reproduce the crystallographic structures of the native complexes. The better and faster understating of the protein-peptide recognition pathways

<table>
<thead>
<tr>
<th>Complex</th>
<th>Peptide length</th>
<th>PDB ID</th>
<th>Method</th>
<th>Affinity and Reference</th>
<th>RMSD(MIN (Å))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bcl-XL - BAD</td>
<td>25</td>
<td>1G5J</td>
<td>Solution NMR</td>
<td>(Kd=0.6 nM)[19]</td>
<td>4.72</td>
</tr>
<tr>
<td>MDM2 - p53</td>
<td>17</td>
<td>1YCQ</td>
<td>X-Ray diffraction</td>
<td>(Kd=1 µM) [20]</td>
<td>4.15</td>
</tr>
<tr>
<td>MDM2 - SAH-p53-8</td>
<td>10 (stapled)</td>
<td>3V3B</td>
<td>X-Ray diffraction</td>
<td>(Kd=55 nM) [21]</td>
<td>1.87</td>
</tr>
</tbody>
</table>

The preliminary results collected in this pilot key study are very encouraging. In fact, pepSuMD methodology allows the simulation of the whole process of protein-peptide recognition (from the unbound to the bound state) in a nanosecond timescale, with an appreciable capability to reproduce the crystallographic structures of the native complexes. The better and faster understating of the protein-peptide recognition pathways
could be very beneficial in enlarging the applicability of peptide-based drug design approaches in several biotechnological and pharmaceutical fields.

2. Results

From a computational point of view, pepSuMD is based on the same supervision approach used in the previously published SuMD methodology [16, 17]. In brief, a pepSuMD simulation is composed of a number of consecutive short unbiased MD simulations (600 ps) in which a supervision strategy, based on a tabu search-like strategy, is applied at the end of each simulation. The supervised variable is the distance between the center of mass of the peptide and the center of mass of its binding site on the protein. In a nutshell, if this distance is likely to be shortened during the simulation, the MD simulation is prolonged, otherwise, it is stopped, and the simulation is restarted from the previous set of coordinates. The supervision is maintained until the protein-peptide distance reaches a preset threshold value, then the simulation proceeds as a conventional unbiased MD simulation. This threshold is user-definable and in this pilot key study has been set at 10 Å. A completely automated set of tools has been developed to analyze a pepSuMD trajectory from a geometric and energetic point of view, including a self-production video recorder to reproduce the whole pepSuMD trajectory.

The most remarkable results for each of the three key studies will be briefly described below.

2.1 Test Case 1: Bcl-XL/BAD Complex

In this case, as well as for all other described herein, to avoid any bias in reproducing the protein-peptide bound state, the initial unbound state was prepared randomly distancing the peptide very far from its protein recognition site and, as well, very far from the long-range interaction cutoff (i.e., 9 Å). In the case of the Bcl-XL/BAD complex, the BH3 domain of BAD (Asn103-Lys127) was positioned 72 Å away from the binding cleft of Bcl-XL ($d_{cm(P-T)} = 72$ Å). As reported previously, the unbound state was hydrated, neutralized, and equilibrated before running the pepSuMD protocol [16, 17].

As shown in Movie S1, in the first part of the simulation (0–7 ns) the peptide tumbles in solution before approaching the cutoff threshold for the attractive interactions with Bcl-XL. The evolution of the distance between the centers of mass (peptide-protein) is reported in Figure S1.

The flexibility of the BH3 α helix of BAD is evident between 2.4 and 3.4 ns when it temporarily bends at about 90° with respect to the helix axis; at 2.82 ns the peptide highly deviates from the reference conformation, with a Cα-root-mean-square deviation (Cα-RMSD) value of 6.27 Å (Figure S5).

The first peptide-protein interaction occurs at around 7 ns and is mediated by the peptide N terminus and the loop connecting the α3 and α4 helices of Bcl-XL, involving residues Gln111- Gln121 (Figure 1D).
After 2 ns, BAD establishes a greater contact with the α2 and α3 helices of Bcl-Xₐ in particular through the interactions of Glu113 (BAD) with Arg102/103 (Bcl-XL) and of Asn103 (BAD) with Asp107 (Bcl-XL) (Figures 1E and 1F). A series of hydrophobic contacts are established between Tyr101 (α3 helix of Bcl-Xₐ) and Tyr110 and Leu114 (BAD) leading to the reorientation of the BAD peptide with the helix axis parallel to the Bcl-Xₐ cleft.

Up to 10 ns, the C-terminal helix of Bcl-Xₐ explores a wide pool of conformations (Figure S4C). The Cα-RMSD of Bcl-Xₐ reaches the maximum value of 6.24 Å at 5.42 ns (Figure S4A), with the greater contribution provided by the C-terminal portion, as indicated by per-residue RMSD analysis (Figure S4B). After 10 ns, the C-terminal portion of Bcl-Xₐ makes contacts with BAD, driving it toward the binding site, and this enhances the stabilization of the Bcl-Xₐ C terminus close to the bound reference to the end of the simulation (Figures S4D–S4F).

Fig. 1 Bcl-XL-BAD Recognition Pathway (A) Superimposition between the experimental nuclear magnetic resonance complex (PDB:1G5J) (orange-colored BAD peptide) and the pepSuMD conformation with lowest RMSD along the trajectory (blue-colored BAD peptide). The superposition was performed considering only the target protein residues. The peptide is shown using a ribbon style, while the protein is represented using both ribbon (gray) and surface (white, transparent rendering). The nomenclature of the most relevant Bcl-Xₐ helices is reported on the corresponding α helix segment. (B) RMSD of simulated BAD peptide Cα carbon atoms against PDB references. (C) Interaction energy landscape. (D) Peptide-target recognition map. (E and F) Dynamic electrostatic interaction energy, on the peptide and protein side, respectively.
Between 12 ns and 18 ns, the peptide does not undergo significant movement; the whole peptide is involved in contacts with α2, α3, and C-terminal helices of Bcl-Xₐ with a strong electrostatic interaction between Glu120 (BAD) and Arg100 (Bcl-Xₐ). The protein region involved in this prolonged interaction may be defined as a metastable binding site, as revealed by the stability of MMGBSA energy values (Movie S1, lower left). A metastable binding site is a sort of stopover with a sufficient residence time, that breaks the progressive and continual approach of the peptide.

The molecular mechanism leading BAD to reach the final binding site is the interaction between Arg115 and Glu129 of BAD and Asp133 at the C terminus of α4 helix of Bcl-XL (Figures 1E and 1F). A dynamic qualitative and quantitative analysis of the target residues mainly involved in BAD binding is reported in Movie S1, lower right, in which the cumulative electrostatic interactions highlight residues Asp107, Asp133, Glu129, Arg100, and Arg103 as important in the binding process.

Finally, at 25 ns the pattern of interactions found in the experimental structure is achieved: the hydrophobic residues Tyr110, Leu114, Phe121, and Phe125 of BAD are inserted into a series of hydrophobic pockets within the binding cleft.

In the final 20 ns the peptide fluctuates without changing its orientation, as can be observed by the RMSD profile (Figure 1B; Movie S1, upper right). The fluctuations are restricted to side chains, which lead to the optimization of intermolecular peptide-protein interactions, as shown in the MMGBSA energy profile (Movie S1, lower left). Moreover, the estimated MMGBSA energy values that can be associated to the bound state are in accordance with the extended surface involved in binding and are compatible with the sub-nanomolar value of the experimental Bcl-XL/BAD complex dissociation constant [19].

Summarizing, the recognition pathway of BAD with Bcl-XL depicted by the analysis of SuMD trajectories is compatible with a two-step mechanism of binding: a first intermediate binding state that anticipates the final bound state, as shown by the interaction energy landscape in which the profile clearly shows a nonmonotonic trend (Figure 1C).

2.2 Test Case 2: MDM2/p53 Complex

In this second key study, the p53 peptide was positioned 28 Å away from the MDM2 binding cleft away (d_{cm(P-T)} = 28 Å).

The whole recognition pathway can also be appreciated in this case by browsing Movie S2. The peptide-protein centers of mass distance decreases from the initial 28 Å to about 10 Å during the first 15 ns of SuMD simulation. After that, it is stabilized at about 8 Å, with a 7.7 Å value in the last frame (Figure S2). The first significant contacts between p53 and MDM2 occur after the first 1.5 ns, involving the loop connecting α1 and α2 (Gln40-Thr45) and the peptide C terminus (Movie S2, upper left, and Figure 2D). At around 3 ns of
simulation, the peptide gets closer to helix α2 due to the interactions of Ser20, Asp21, Trp23, and Lys 24 with Tyr51 and Gln55 of MDM2 (Figures 2E and 2F). These interactions lead to a peptide reorientation assuming a parallel position to helix α2 of MDM2.

At this point, p53 needs to roll over the surface of helix α2 to finally reach the binding cleft. This recognition mechanism is nicely described by the MMGBSA energy plot (Movie S2, lower left) in which the two minima at 4 and 8 ns are separated by an energetic barrier (5.5 ns).

Finally, at 13 ns the p53 peptide joins the orientation found in the crystallographic structure, with Phe19 and Trp23 inserted in the binding cleft of MDM2, and the RMSD drops at around 5 Å. In the next 10 ns, the
position of the peptide maintains the same orientation as highlighted by the RMSD profile (Figure 2B; Movie S2, upper right), while the side chains of both protein and peptide continue in an induced-fit recognition mechanism that can be evinced by the MMGBSA profile (Movie S2, lower left) and, in a phenomenological manner, by the molecular trajectory (Movie S2, upper left).

Although the orientation of the helix of p53 is nicely reproduced in the last part of the simulation, the helix axis is slightly shifted and the characteristic hydrogen bond between the indolic nitrogen of Trp23 and the backbone of Ile50 of MDM2 is not observed, due to a flipping by 180° of the indole ring (Figure 2A).

From these preliminary simulations, the p53-MDM2 recognition pathway extracted by the analysis of SuMD trajectories is again compatible with a two-step mechanism of binding, as supported by the non-monotonic trend of the energy interaction landscape (Figure 2C). In particular, the preliminary binding to the α2 helix could be considered as an intermediate binding site.

### 2.3 Test Case 3: MDM2/SAH-p53-8 Complex

The last key study has been chosen in order to evaluate the capability of the pepSuMD approach to appropriately deal with simulations in which a peptide is replaced by a peptidomimetic. Specifically, in this case we have explored the recognition of MDM2 by a stapled p53 peptide, named SAH-p53-8 [21]. This peculiar peptidomimetic has been designed to stabilize the helical conformation, crucial for the interaction with MDM2, through the introduction of an olefinic crosslinking moiety placed before Asn20 and after Leu26 in the native p53 sequence numbering.

In these simulations, the SAH-p53-8 p53 mimetic was positioned 27 Å away from the MDM2 binding cleft ($d_{cm(P-T)} = 27$ Å), and the possible recognition pathway is summarized in Movie S3. The evolution of the distance between the centers of mass (peptide–protein) supervised by pepSuMD is reported in Figure S3. In the first part of the trajectory, the peptide randomly tumbles in the water box and requires 2 ns of pepSuMD simulation to approach the protein. A preliminary stable recognition occurs at around 1.6 ns (Movie S3, upper left) and involves the two loops surrounding the binding site (Val93-Arg97, earlier, and Gln71-His73) with the C-terminus of the peptidomimetic (Figure 3D). Immediately thereafter, helix α2 of MDM2 (Ly51-Gln59) is also engaged in the interaction and the peptide assumes an orientation parallel to the binding cleft, compatible with the native interaction. At 3.7 ns, a well-known induced-fit mechanism occurs [24, 25]: the side chain of Tyr67 (MDM2) rotates enlarging the binding pocket to accommodate Phe19 (SAH-p53-8) and the consecutive settlement of the peptide. Here, SAH-p53-8 reaches a stable conformation almost identical to the crystal structure (Figure 3A): Phe19, Trp23, and Leu26 side chains are oriented in the hydrophobic cleft of MDM2 and the olefinic chain interacts with helix α2 of MDM2. The coordinates of the indolic nitrogen atom of Trp23 and the backbone carbonyl group of Leu54 are compatible with the hydrogen bond formation found in the
crystallographic structure. Gln72 is relevant in the complex stabilization for the formation of a hydrogen bond with the N-terminal portion of SAH-p53-8, which emerges especially between about 8 and 10 ns (Figure 3E; Movie S3, lower right).

Fig. 3 MDM2/SAH-p53-8 Recognition Pathway (A) Superimposition between the experimental crystallographic complex (PDB: 3V3B) (orange-colored peptidomimetic) and the pepSuMD conformation with lowest RMSD along the trajectory (blue-colored peptidomimetic). The superposition was performed considering only the target protein residues. The protein is represented using both ribbon (gray) and surface (white, transparent rendering). The peptidomimetic is represented using full heavy-atom style. The position of the most relevant residues for the interaction, Phe19 and Thr23, are indicated by black arrows. (B) RMSD of simulated SAH-p53-8 ligand heavy atoms against PDB references. (C) Interaction energy landscape. (D) Peptidomimetic-target recognition map. (E) Dynamic electrostatic interaction energy, on the protein side.

The RMSD reaches the minimum value of 1.9 Å, at 4.5 ns (Figure 3B; Movie S3, upper right) when computed on all the peptide heavy atoms in the peptide, while it drops to 0.7 Å considering only the side chains of Phe19 and Trp23. The strength of the interaction is evident in the MMGBSA profile (Movie S3, lower left), which reaches values close to 60 kcal/mol.
In this trajectory, the absence of stable metastable binding sites is evident from the interaction energy landscape (Figure 3C), in which the profile can be approximated by a monotonic function differently to the p53 peptide of the previous test case. In addition, the modest dispersion of the energy values emphasizes the stability of the complex.

The effect of the introduction of the olefinic bridge is clear when the energetic profiles are compared with the natural peptide p53: it results in a greater stability of the complex (less dispersed points) and a stronger interaction (more negative values) that perfectly fits with experimental evidence.

3. Discussion

The preliminary results obtained from pepSuMD simulations are promising. Although we are aware that, in this first pilot study, the number of protein-peptide complexes taken into account is limited, in all the key studies analyzed, the pepSuMD approach succeeded in reproducing the native binding mode, even when starting from a random and very distant position of the peptide from its binding site. Considering the preliminary reproducibility analysis of SuMD trajectories, the distance between the centers of mass of the simulated peptide in the final bound state and the center of mass of the experimental peptide fell below 5 Å in all the examples, as reported in Figures S1–S3. A summary of the preliminary statistical analysis is reported in Table S1 and Movies S4, S5, and S6. However, even if encouraging, a robust statistical analysis based on a large campaign of pepSuMD simulations is in progress to appropriately analyze the strengths and weakness of the approach and to explore the limits of the applicability domain of this new technique. Moreover, the final conformations of simulated peptides are comparable with those observed in the experimental bound state, as demonstrated by Cα-RMSD values lower than 6 Å in all cases (Figures 1B, 2B, and 3B).

In addition, pepSuMD trajectories were able to reveal the presence of multiple intermediate states (metabinding sites) that chronologically anticipate the native bound site, as observed, for example, for Bcl-XL/BAD and MDM2/p53 systems. Finally, the energetic profiles extracted from pepSuMD trajectories are very useful to analyze the localization, the nature and the intensity of the most crucial peptide-protein interactions (hotspots).

Another crucial aspect that must be carefully taken into account is related to those intrinsically disordered peptides that fold upon binding. In fact, in our pilot study, all peptides in their unbound states are pre-folded and they are recognized by pre-folded proteins through a conformational selection mechanism. In particular, the pepSuMD approach is particularly efficient in dealing with conformational constrained peptides and peptidomimetics. However, minor induced-fit phenomena can be observed during pepSuMD trajectories, as can be seen from the Cα-RMSD profiles in Figures S4 and S5. As already noted, for example, the C-terminal portion of Bcl-XL explores multiple conformations when unbound, while it is stabilized, close to the native
conformation in the bound state (Figures S4C–S4F). Indeed, considerable induced-fit and folding-upon-binding mechanisms have not been extensively explored and, in principle, it could be quite difficult to observe relevant folding phenomena within the short timescale of our supervised binding process. However, further implementations of the pepSuMD approach dealing specifically with the induced-fit phenomena are under development.

Understanding the intimate protein-peptide recognition process remains a charming challenge for structural biology and peptide-based drug discovery.

For the first time, we reported the application of a pepSuMD approach on three different peptide-protein key studies with the aim of verifying the effectiveness and robustness of this method in depicting their possible binding pathways leading to the final bound state as described by the corresponding experimental high-resolution structures. pepSuMD was able to reproduce experimental peptide-protein complexes and to reduce the timescale of a peptide-protein binding event by up to three orders of magnitude in comparison with classical MD.

Insights from pepSuMD simulations can be helpful to explain the mechanistic evidence of recognition processes and they could be very beneficial in enlarging the challenges of peptide-based drug discovery.

4. Method details

4.1 General

All simulations were carried out on a hybrid CPU/GPU cluster. MD simulations were performed with the ACEMD engine [26] on a GPU cluster equipped with 20 NVIDIA GTX graphics cards. Prior to run pepSuMD simulations, the following preliminary steps were accomplished: (i) protein-peptide system preparation; (ii) peptidomimetic parameterization, if necessary; (iii) solvated system setup and equilibration.

4.2 Protein-Peptide Systems Preparation

Protein-peptide complexes were retrieved from the RCSB PDB database [27] and processed with the protein preparation tool as implemented in MOE [28]: hydrogen atoms were added to X-ray derived complexes and appropriate ionization states were assigned by Protonate-3D tool [29]. Missing atoms in protein side chains were built according to AMBER12 [30] force field topology. Non-natural N-terminal and C-terminal were capped to mimic the previous residue. To avoid protein-ligand long range interactions in the starting geometry, the peptide was then moved far from the protein at a distance bigger than the electrostatic cut-off term used in the simulation (9 Å with Amber force field).
4.3 Peptidomimetic Parameterization

The non-natural portion of the peptidomimetic ligand was parametrized with GAFF [31] as implemented in ambertools2014 [32] by using antechamber and parmchk tools. RESP partial charges were calculated with Gaussian 09 [33] following the procedure suggested by Antechamber [34].

4.4 Solvated System Setup and Equilibration

Complexes were assembled with tleap tool using AMBER14SB [32] as force field for the protein. 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. To neutralize the total charge, Na+/Cl- counterions were added to a final salt concentration of 0.150 M. The systems were energy minimized by 2000 steps with conjugate-gradient method, then 50000 steps of NVE (100 ps) followed by 1 ns of NPT simulations were carried out, both using 2 fs as time step and applying harmonic positional constraints on protein and peptide/peptidomimetic heavy atoms by a force constant of 1 kcal mol⁻¹ Å⁻², gradually reduced with a scaling factor of 0.1. During this step, the temperature was maintained at 310 K by a Langevin thermostat and the pressure at 1 atm by a Berendsen barostat.

4.5 Peptidic Supervised Molecular Dynamics (pepSuMD)

pepSuMD ensues from SuMD code recently developed [17], in order to be applicable to peptides and peptidomimetics. The entire protocol is written in Python and bash and operates the supervision of MD trajectories according to the algorithm that has been previously described. Similarly to the original implementation, also in this protocol the supervision algorithm monitors the distance between the mass centers of the peptide and the target \(d_{cm(P-T)}\).

The program exploits ProDy Python package [35] and Gnuplot functionalities [36]. In its current implementation, pepSuMD is interfaced with the ACEMD [26] engine and supports AMBER and CHARMM force fields. Differently from the previous SuMD code, here more input parameters are user-editable, such as: the MD timestep (here, 2 fs), the number of MD steps within a pepSuMD step (here, 300000), an eventual substructure of the peptide used for mass centers computation (here, the entire peptide).

Three simulations were carried out for each system starting from the same initial geometry. The more significant replica for each test case is described in the Results section.

4.6 Analysis of pepSuMD Trajectories

All the trajectories generated by pepSuMD were analyzed by an in-house script written in tcl and python, that makes use of Numpy [37] and ProDy modules [35].
The single pepSuMD step trajectories were stridden (by a user defined value, here 10), superposed on the first frame Cα carbon atoms of the target protein, wrapped and merged. The following analyses were then performed on the whole trajectories. The peptide RMSD of Cα carbon atoms was computed with respect to the reference PDB structure, after superposing the target protein structure to its corresponding PDB one. The RMSD values were plotted over time and reported on the upper-right side of the Movies S1, S2, and S3 and in Figures 1B, 2B, and 3B. The RMSD graphic depicts the peptide conformation variation along the trajectory in comparison to the experimental conformation within the target binding site.

A peptide-target interaction energy estimation during the recognition process was calculated using an MMGBSA protocol with Amber2014 [32], adopting for non-polar and polar solvation energy calculations LCPO [38] and GB°Bc model II [39], respectively. MMGBSA values were plotted over time and reported in the lower-left side of the Movies S1, S2, and S3.

The MMGBSA values were also arranged according to the distances between peptide and target mass centers \((d_{cm(P,T)})\) in the Interaction Energy Landscape plots (Figures 1C, 2C, and 3C). Here, the distances between mass centers are reported on the x-axis, while the MMGBSA values on the y-axis, and are rendered by a colorimetric scale going from blue to red for negative to positive values. These graphs allow evaluating the variation of the interaction energy profile at different peptide-target distances, helping to individuate metastable binding states during the binding process.

In order to have a rapid indication about the residues mostly involved in the binding process, for each target residue, the total number of contacts with the peptide was computed along the trajectory. A target residue within a distance of 4 Å from any peptide atoms was considered as in contact with the peptide. On the basis of these data, the Peptide-Target Recognition Maps (Figures 1D, 2D, and 3D) were constructed, resuming the chronological evolution of the molecular contacts in a quantitative manner: here, for each residue (x-axis) the total number of contacts is reported with respect to the simulation time (y-axis) and rendered by a colorimetric scale going from white to dark green from 0 to higher numbers.

The most contacted residues were selected both for protein target and peptide to compute the per-residue electrostatic interaction energy with the peptide and target, respectively. NAMD was used for post-processing computation of electrostatic interactions, using AMBER14SB force field. The Dynamic Electrostatic Interaction Energy plots (Figures 1E, 2E, and 3E (protein), F (peptide)) depict for each selected residue (x-axis) the evolution of the electrostatic interaction energy along time (y-axis), using a colorimetric scale going from blue to red for negative to positive values.
The cumulative electrostatic interactions were computed for the same target residues by summing the energy values frame by frame along the trajectory, and the resulting graphs were reported at the lower-right of Movies S1, S2, and S3.

Representations of the molecular structures were prepared with VMD [40].

**4.7 Data and software availability**

pepSuMD code was written in python language starting from the previously developed SuMD code. The algorithm is described at the beginning of the Results section. An in-house script written in python and tcl was used to automate analyses of the pepSuMD trajectories. An explanation of the analyses is reported in the Method Details section. All software used are reported in the Method Details section, together with the Key Resources Table.

<table>
<thead>
<tr>
<th><strong>Key Resources Table</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Software and Algorithms</strong></td>
</tr>
<tr>
<td>ACEMD 3212u2</td>
</tr>
<tr>
<td>MOE suite</td>
</tr>
<tr>
<td>Gaussian 09</td>
</tr>
<tr>
<td>Ambertools2014</td>
</tr>
<tr>
<td>pepSuMD</td>
</tr>
</tbody>
</table>
References:


New Trends in Inspecting GPCR-ligand Recognition Process:
the Contribution of the Molecular Modeling Section (MMS)
at the University of Padova

Antonella Ciancetta, Alberto Cuzzolin, Giuseppe Deganutti, Mattia Sturlese, Veronica Salmaso, Andrea Cristiani, Davide Sabbadin, Stefano Moro


Abstract

In this review, we present a survey of the recent advances carried out by our research groups in the field of ligand-GPCRs recognition process simulations recently implemented at the Molecular Modeling Section (MMS) of the University of Padova. We briefly describe a platform of tools we have tuned to aid the identification of novel GPCRs binders and the better understanding of their binding mechanisms, based on two extensively used computational techniques such as molecular docking and MD simulations. The developed methodologies encompass: (i) the selection of suitable protocols for docking studies, (ii) the exploration of the dynamical evolution of ligand-protein interaction networks, (iii) the detailed investigation of the role of water molecules upon ligand binding, and (iv) a glance at the way the ligand might go through prior reaching the binding site.

2. Introduction

Today, it is largely recognized that G protein-coupled receptors (GPCRs) represent the largest family of surface receptors with more than 800 members in humans.[1] They respond to different extracellular stimuli ranging from small molecules to lipids, peptides, proteins, and even light.[2] The binding event triggers the activation of cytoplasmic heterotrimeric GTP binding proteins (G proteins) and mediates the signal transduction through the modulation of several downstream effectors. The participation of GPCRs in numerous physio-pathological processes entails a potential role for their modulation by agonist, antagonists and inverse agonists in the treatment of several diseases, including cardiovascular and mental disorders,[3] cancer,[4] and viral infections.[5] Nowadays, about more than 50 % of the drugs in clinical use targets a GPCR.[6]

According to the GRAFS classification,[7] human GPCRs are commonly grouped into five main classes: Glutamate (Class C), Rhodopsin (Class A), Adhesion (Class B), Secretin (Class B), and Frizzled/Taste2 (Class F). From a structural point of view, all members share a common architecture represented by seven membrane-
spanning helices connected by three intracellular and three extracellular loops with the N-Term domain exposed toward the extracellular side.

The insertion into the cell membrane along with receptors dynamism have hampered for long time the structural determination of GPCRs by X-ray crystallography. To overcome these limitations, several techniques have been developed: the use of fusion proteins such as T4 lysozyme or apocytochrome,[8],[9] complexation with antibody fragments,[10] and the receptor thermostabilization through systematic scanning mutagenesis.[11] The advances in protein engineering and crystallography have represented a breakthrough for the research focused on GPCRs and yielded numerous X-ray structures.[12] The availability of ligand-bound three-dimensional structures provides invaluable insights to understand GPCRs function and pharmacology and enables the application of structure-based drug design approaches to aid the discovery of novel candidates with improved pharmacological profiles.[13] In particular, molecular dynamics (MD) simulations have become a helpful complement for the study of GPCRs biophysics and molecular pharmacology, by enriching our understanding of, among other aspects, ligand-receptor interaction and ligand-subtype selectivity.[14],[15]

In addition, the recent exploitation of the commodity, graphics processing units (GPUs), a technology firstly designed to improve video game performances, in the molecular modeling field represents an important step forward for the simulation of GPCRs in explicit lipid-water environments within a reasonable computation time.[16] In this paper, we briefly survey the recent advances carried out by our research groups in the field of ligand-GPCRs recognition process simulations.[17] Following the description of the tools we have developed to aid the identification of novel binders of GPCRs binders and the better understanding of their binding mechanisms, we will discuss their use in a case study: the comparison between ZM 241385 (4-(2-(7-Amino-2-(furan-2-yl)triazolo[2,3-a][1,5-a][1,3],5-triazin-5-yl-amino)ethyl) phenol) and caffeine, a strong and weak human Adenosine 2A Receptor (hA2A AR) antagonists, respectively.

2. Methods

2.1 Docking Protocols Validation: the “Quality Descriptors”

The availability of ligand-bound crystal structures enables to perform docking simulations to rationalize structure-activity relationships of known binders or to conduct virtual screening campaigns to identify novel candidates. It is highly recommended to assess the performances of a docking protocol in reproducing the available experimental data prior to applying it. This procedure is best known as benchmark study. We have recently developed a pipeline that allows a fast graphical evaluation of different docking protocols, based on two newly defined quality descriptors: the “Protocol Score” and the “Interaction Energy Map” (IEM).[18],[19]
The “Protocol Score” is a RMSD based descriptor that assigns a 0–3 score to each docking protocol according to the following criteria: (i) if the protocol returns either a RMSD\textsubscript{ave} value lower than the crystal structure resolution (R) or generates at least 10 (out of 20) conformations having RMSD<R, a score 1 is assigned; (ii) if a protocol satisfies both the above mentioned requirements, a score 2 is assigned; (iii) if a protocol satisfies none of the above mentioned requirements, a score 0 is assigned. Moreover, a score 3 is conferred to the best protocols, i.e. those returning at the same time the lowest RMSD\textsubscript{ave} value and the highest number of conformers with RMSD<R. The scores are then converted in a color code and the data visualized as a colored map (Figure 1A): protocols corresponding to white and light green spots are not suitable for the system under consideration, dark green spots highlight good protocols, whereas blue spots identify the best among the tested ones.

Fig. 1 Schematic representation of the developed tools: A) Protocol Score; B) Interaction Energy Maps (IEMs); C) RMSD weighted Dynamic Scoring Function (wDSF); D) Water Fluid Dynamics (WFD) maps; E) Differential WFD maps; F) Ligand-receptor interaction energy landscape from supervised MD (SuMD) simulations. The figures were adapted from the original papers.[18],[21],[27],[35]
The IEMs are based on the analysis of ligand-protein interactions and are derived as follows. Firstly, per residue electrostatic and hydrophobic contributions to the interaction energy (denoted \( IE_{ele} \) and \( IE_{hyd} \), respectively) are computed for residues surrounding the binding site or known to play a role in the binding. The analysis is performed for both the crystallographic binding modes and the docking poses. These pieces of information are then graphically transferred into heat-like maps reporting the key residues involved in the binding with the considered ligands along with a color code reflecting the quantitative estimate of the occurring interactions (the more intense the color, the stronger the interaction). The comparison is therefore based on the quality of the interactions – in terms of number of established interactions and their relative strength – among the X-Ray binding mode and the generated docking poses (Figure 1B).

The main advantage of the proposed pipeline resides in the full automation of the benchmark procedure: the user is provided with pre-compiled input files for several docking programs, thus minimizing the required expertise to carry out the benchmark study. To this aim, the results are presented as easy to interpret colored maps enabling a fast graphical inspection of large amount of data. The results are analyzed on the basis of the above described quality descriptors.

### 2.2 Binding Modes Inspection: the Dynamic Scoring Function (DSF)

The docking approach suffers from several limitations.[20] Although is a valuable method to get insights on the final stage of ligand-protein recognition, it lacks the description of two fundamental aspects that might play a significant role in ligand binding: water molecules mediated interactions and protein flexibility. To complete the description provided by the docking method with such contributions, we have recently developed the “dynamic scoring function” (DSF), an approach that enables to follow the dynamical evolution of a docking pose in a realistic environment, \( i.e. \) the solvated membrane embedded ligand-protein complex.[21] The DSF provides a dynamic estimate of both the ligand position and the strength of the interaction network while accounting for the interplay of water molecules and protein side-chains flexibility.

The procedure envisages the dynamic selection of residues within a range of 4.5 Å from the ligand during the MD simulation, starting from a previously obtained docking pose. The DSF is the cumulative sum of electrostatic and hydrophobic contributions to ligand-protein interaction (\( DSF_{ele} \) and \( DSF_{hyd} \), respectively) and is calculated at frames extracted every 100 ps as follows:

\[
DSF_{ele} = \sum_{t=0}^{n} IE_{ele}
\]

\[
DSF_{hyd} = \sum_{t=0}^{n} IE_{hyd}
\]

The DSF value corrected for the ligand fluctuation (RMSD) with respect to the starting position yields the weighted DSF (wDSF), a number that highlights differences between stable and unstable poses. The
corresponding weighted electrostatic and hydrophobic DSFs (denoted as \(wDSF_{ele}\) and \(DSF_{hyd}\), respectively) are therefore obtained as reported below:

\[
\begin{align*}
\text{\(wDSF_{ele}\)} &= \frac{\sum_{t=0}^{n} F_{ele}}{RMSD} \\
\text{\(wDSF_{hyd}\)} &= \frac{\sum_{t=0}^{n} F_{hyd}}{RMSD}
\end{align*}
\]

The DSFs can be computed during the MD simulations or performed as a post-processing procedure, so that in principle any trajectory that has been previously produced can be re-analyzed with this approach. It can be regarded as an alternative to conventional scoring functions, as it is able to take into account both the complex flexibility in the membrane environment as well as water-driven interactions. The resulting graphs (Figure 1C) obtained by plotting the DSFs against the simulation time enable a graphical comparison of the relative stability of docking poses. This representation can help in detecting and validating the feasibility of alternative binding conformations proposed by the docking algorithm. We have recently exploited this feature to support an apparently less plausible binding mode of a series of 5-alkylaminopyrazolo[4,3-e]1,2,4-triazolo[1,5-c]pyrimidine at the hA3 AR.[22] Moreover, we tested the applicability of our approach by taking part in the community-wide 2013 GPCR Dock Assessment.[23] Among the proposed targets, we focused on the 5HT2B/ergotamine complex, whose X-Ray structure has been released after the predictions were submitted. Therefore we tested the applicability of our tool to homology models of a system on which our laboratory did not hold expertise. We submitted several alternative ligand-protein complexes suggested by the docking protocol to membrane MD simulations and selected the best final poses according to the outcomes of the DSFs analysis. Our predictions ranked 8th among 254, suggesting the portability of our approach to homology models as well as to other GPCRs.[23]

2.3 A Closer Look at Water Molecules: Water Fluid Dynamics (WFD) Maps

It is generally recognized that water molecules contribute to protein-ligand binding in at least two ways: they either stabilize the complex by forming hydrogen bond networks,[24] or are replaced by the ligand once the complex is formed.[25],[26] It is therefore crucial in a drug design process to be able to distinguish between water molecules that mediate protein-ligand interactions and those that can be targeted for being displaced. To this aim, we have very recently tuned a tool that inspects the time-dependent variation of fluid dynamics properties of water molecules as a consequence of the binding event by means of MD simulations.[27] Our approach detects structural water molecules inside the orthosteric binding site of the receptor and collects these pieces of information in a bi-dimensional graph, that we called water fluid dynamics (WFD) map. Unlike other existing MD based methodologies,[28],[29] our approach is aimed at localizing protein “hot-spots” – i.e. regions where water molecules playing a key role in ligand binding mostly reside – rather than
estimating their binding affinity. The WFD maps have been therefore mainly conceived as qualitative tool to drive ligand design to avoid substituents disrupting key water molecules’ networks.

The WFD maps are derived as follows: residues within a range of 5 Å from the ligand are selected and a box surrounding the binding site is created and split into a three-dimensional grid. During the MD simulations the diffusion of water molecules in each grid cell is followed. The data are acquired by saving the MD trajectories at regular intervals (every 10 ps) and by projecting the averaged position of water molecules showing a RMSF value below 1.4 Å into a bi-dimensional grid. The overlap of these grids yields a map (Figure 1D) with cell colored according to the residence time of water molecules on a 0–100 % scale. White zones (0 %) are occupied by water molecules with a residence time equivalent to bulk, whereas blue regions (100 %) are occupied by trapped water molecules showing the maximum residence time of the considered trajectory. The maps allow a fast graphical identification of water distribution inside the orthosteric binding pocket. Moreover, “differential” WFD maps representing by a color code the enrichment or displacement of water molecules as a consequence of ligand binding (Figure 1E) are derived by comparing the WFD maps of the receptor in the apo and bound states.

2.4 Exploring the Ligand-Receptor Recognition Process: the Supervised Molecular Dynamics (SuMD) Approach

One of the most challenging tasks for ligand-GPCRs modeling is the prediction of the recognition pathway, an event which knowledge would ease the development of drug candidates with better pharmacodynamic profiles. Unfortunately, the recognition of a ligand by a receptor is a process hard to simulate as it requires classical MD experiments in a long microsecond time scale.[14],[30],[31] To overcome this technical limitation, enhanced sampling methods that facilitate the crossing of energy barriers through the introduction of biased potentials have been developed.[32],[33] Another approach,[34] induces ligand unbinding by applying external forces to the system, thus requiring knowledge of the ligand-receptor complex final state. Within this framework, we have recently proposed an alternative strategy – the “supervised molecular dynamics” (SuMD)[35] – that enables to follow the ligand-GPCR approaching path by considerably reducing the simulation time scale and without introducing bias. SuMD performs standard simulations in which the distance between the center of masses of the ligand atoms and the receptor binding site is monitored by a tabu-like algorithm. If the location of the binding site is unknown, several simulations are run by setting the centers of previously detected cavities. An arbitrary number of distance points is collected “on the flight” and fitted into a linear function f(x)=mx. The tabu-like algorithm is applied to increase the probability to produce ligand-receptor binding events as follows: If the slope (m) is negative, the ligand-receptor distance is likely to be shortened and a classic MD simulation is restarted from the last set of
coordinates. Otherwise, the simulation is restored from the original set of coordinates and random velocities are reassigned to each atom. 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 1F). This approach can be exploited to analyze binding events to both orthosteric and allosteric sites and to assist the design of site-directed mutagenesis experiments in order to infer the role of specific residues on the molecular recognition process.

3. Application to Drug Design

To explain the applicability of the described tools, we discuss here as case study the comparison between ZM 241385 and caffeine, a strong and a weak hA2A AR binder with pK\textsubscript{D} values 9.18±0.23 and 5.31±0.44, respectively.[36] Among the hA2A AR available crystal structures we have selected the two co-crystallized with the ligands of interest identified by the following PDB IDs: 3EML and 3RFM.[36],[37] The starting point of the study is the evaluation of the reproducibility of the X-Ray binding modes through docking calculations. To accomplish this task we compare the IEMs computed for the best performing docking protocols. We then proceed by evaluating the dynamic evolution of alternative binding modes proposed by the docking algorithm, thus imaging the common case where X-Ray structures are not available for comparison. As anticipated, MD simulations allow taking into account the flexibility of the receptor and the role of water molecules in the binding. A more careful inspection of water dynamics is then performed by deriving differential WFD maps from the computed trajectories of a selected docking pose for each structure. Finally, we move outside the receptor and try to reproduce the binding pathways from the extracellular side through SuMD experiments.

3.1 Assessing the Reproducibility of a Binding Mode: IEMs Comparison

We start our case study by assessing the performance of a previously selected docking algorithm through IEMs inspection.[18] Figure 2 displays the comparison between the computed IEMs for the two considered structures. As shown, the binding mode of ZM 241385 (3EML, Figure 2A) encompasses a tight interaction network that is correctly reproduced by the majority of the 20 generated poses. On the other hand, the caffeine binding mode (3RFM, Figure 2B) is more challenging to be reproduced and implies a lower number of less intense interactions with the binding site residues. The comparison of IEMs therefore helps evaluating in a fast graphical fashion both the docking protocols performances and the reproducibility of X-Ray observed binding modes. Interaction patterns without interruptions are clue of binding modes easy to reproduce and indicate good protocol performances, whereas discontinuous patterns suggest binding modes challenging to be predicted and unsatisfactory protocol performances.
Fig. 2 Comparison of IEMs for two hA_2A AR ligands: ZM 241385 (A) and caffeine (B). While ZM 241385 establishes a strong interaction network conserved among the 20 generated poses, caffeine finds lower number and less intense interactions. IEele values: kcal Å^{-1} mol^{-1}, IEhyd values: arbitrary units. The figures were adapted from the original papers.[18]

3.2 Following the Dynamics of Ligand-receptor Interactions: wDSFs Profiles

When the X-ray structures of the ligand-protein complex of interest are not available, usually a modeler is asked to select among several feasible binding modes suggested by the docking protocol slightly differing for the assigned scores. How to recognize the solution best approaching the “real” binding mode? Figure 3 displays the exercise we have conducted to address this issue: in order to identify as many different as possible binding modes, we forced the docking protocol to return ten poses that differed in terms of RMSD for at least 1.75 Å.[27] Nevertheless, the protocol assigned to the generated conformations scores differing at most for ten units. We subjected each docking pose to MD simulation and evaluated the wDSFs. Figure 3A–B displays the results for the two considered structures: the different values of both the cumulative electrostatic and the hydrophobic contributions reflect the different affinities of the two binders. In particular, ZM 241385 exhibits higher absolute values for both contribution types consistently with its higher affinity for the receptor. Moreover, for both structures, the wDSFs trends enable to graphically recognize the
pose that best reproduces the X-ray observed binding mode, i.e. the one showing the slope with the highest absolute value.

![Fig. 3 wDSFs comparison: A) ZM 241385 wDSF\textsubscript{hyd} (top) and wDSF\textsubscript{ele} (down); B) caffeine wDSF\textsubscript{hyd} (top) and wDSF\textsubscript{ele} (down). I\textsubscript{ele} values: kcal Å\textsuperscript{-1} mol\textsuperscript{-1}, I\textsubscript{hyd} values: arbitrary units. For both ligands the boundle of poses subjected to MD are rendered coloring pose number 1 in red. The same color scheme is used in the plots to identify pose 1 among the others. The figures were adapted from the original paper.[21]]

3.3 What About the Role of Water Molecules? WFD Maps Inspection

A detailed inspection of the WFD maps of the two considered compounds further contributes to explain their different binding affinities to the hA\textsubscript{2A} AR. The WFD map for the ZM 241385 complex (Figure 4A) highlights the presence of water molecules bridging the aromatic scaffold to key residues in the binding site, namely Tyr9, Glu13, His278, Asn253, and Glu169.[27] The interactions with some of those residues were already detected from the docking pose, whereas other ones arose from MD simulations. The differential WFD map (Figure 4B) highlights that the ligand displaces water molecules close to Thr88 while binding. The WFD map corresponding to the caffeine complex (Figure 4C), instead, shows a high propensity of bulk water molecules to solvate the fragment-like compound.[27] This is a direct consequence of the lack of strong interactions with the residues of the binding site detected during the MD simulation and aid explaining the lower affinity of the compound.
Fig. 4 WFD maps comparison: A) position of water molecules experimentally determined for ZM 241385 complex structure; B) differential WFD for ZM 241385 in comparison to the apo-state of hA2A AR; C) differential WFD for caffeine in comparison to the apo-state of hA2A AR. Receptors are viewed from the membrane side facing TM6 and TM7. Side chains of key residues are displayed as gray sticks. Hydrogen atoms are not displayed. The figures were adapted from the original paper.[27]

3.4 On the Extracellular Side of hA2A AR: the SuMD Approach

On its way to the orthosteric binding site, the ligand might interact with the so-called meta-binding sites,[38] which in some cases, may coincide with possible allosteric sites. The SuMD path we have computed for ZM 241385 highlights two major interaction sites: the second and third extracellular loop (EL2 and EL3, respectively, Figure 5A).[35] As depicted in the diagram in Figure 5B, although a higher interaction (less favorable) energy is associated to these meta-binding sites, they seem to play a role in tuning the correct orientation of the ligand scaffold while approaching the orthosteric site. The EL3 also takes part in the caffeine recognition pathway (Figure 5C),[35] which, however, lacks strong interactions with the orthosteric site (Figure 5D). The SuMD simulations thus recognize the critical role of the hA2A AR extracellular loops in the ligand recognition process, role that has been postulated in the past by using site-directed mutagenesis.[39],[40] We have recently applied the SuMD approach to interpret the binding of two challenging ligands: (i) the natural agonist and a (ii) imidazoquinolinamine derivative acting as positive modulator (LUF6000). The binding of the natural agonist adenosine at the hA2 AR revealed a possible energetically stable meta-binding site.[41] The SuMD simulations suggested at least two possible mechanisms to explain the available experimental data for the positive allosteric modulation mediated by LUF6000 toward the hA3 AR.[42]
**4. Summary and Outlook**

Through this paper, we have surveyed the recent advances carried out by our research groups in modeling the ligand-GPCRs recognition process. The crystallographic revolution of the last decade, on one side, and the advent of graphics processing units (GPUs) in the molecular modeling field, on the other side, allowed us to tune several tools to assist the drug design procedure.

The proposed approaches enrich the pool of molecular modeling techniques currently available to disclose the factors influencing the ligand-GPCRs recognition process and exploit two computational methodologies extensively used by modelers such as molecular docking and membrane MD simulations. The majority of the methods herein presented are conceived as post-processing procedures, so that in principle any docking output or MD trajectory previously obtained can be rapidly re-analyzed using these tools. Moreover, the full automation of the procedures as well as the presentation of the results as easy to interpret colored maps are aimed at broadening their applicability within the scientific community encouraging non-expert users to approach them. A different philosophy is instead at the basis of the SuMD approach, which introduces a supervision of the MD trajectory through a tabu-like algorithm to speed up the computation time required to inspect the ligand-GPCRs recognition event. The comparison between ZM 241385 and caffeine, a strong and a weak hA2A AR antagonists, has been presented as case study to explain the usefulness and potentiality of our approaches.

---

**Published:** Mol Inform 2016 Sep;35(8-9):440–448.
As a future perspective we foresee to extend and improve the applicability of these computational tools to address other fascinating open questions in GPCRs field. We would like to summarize some of the hottest topics in the area: a) clarify at the molecular level the orthosteric and the allostERIC control mediated by different binders on GPCR functionality; b) elucidate the implication of phosphorylation and glycosylation in both ligand binding and receptor activation; c) understanding the physio-pathological meaning of monomer-oligomer (homo and/or hetero) receptor equilibrium; d) identification of novel second messengers involved in G protein-alternative signaling pathways; e) explore the possibility to perform high-throughput SuMD (HTSuMD) simulations for virtual screening applications as well as for real-time interpretations of mutagenesis data.

Concluding, we hope that these computational approaches carefully integrated with all other experimental GPCRs competencies will broaden our perspectives in several scientific areas from molecular pharmacology to drug discovery.
References


Conclusions and Future Perspectives
The present work has focused on the role of protein flexibility in binding processes and on the importance of considering this element in computer-aided drug design. Methods neglecting protein flexibility could be exploited more easily for high throughput virtual screening campaigns because of their speed. Otherwise, such an approximation could result both in the identification of false positives, since the kinetic component of binding is omitted, and false negative, since possible alternative conformations of the protein and binding pockets are not considered.

During this work, the flexibility issue has been afforded in mainly two strategies, one belonging to the field of traditional molecular docking methods, and the other to the field of molecular dynamics.

As regards molecular docking, a docking benchmark pipeline has been introduced, merging ligand-based and structure-based strategies to assess the best protein structure for each ligand of a database. In particular, each compound is associated to an ensemble of protein structures by similarity with the co-crystallized ligands. Then, a cross-docking job is used to select the best protein of the ensemble for that ligand. Subsequently, docking is performed for each compound of the database with its tailored protein structure. This method is based on the assumption that similar compounds bind similar conformations of the protein, either by inducing that conformation (induced-fit) or by selecting it (conformational selection). Thus, in our approach, protein flexibility is addressed by associating to each ligand the protein that, having memory of a similar compound, most likely will host it properly. The great advantage of the proposed method is that it is completely automatic, with a consequent convenience in terms of speed. Work is in progress to implement the ligand-similarity filter and the cross-docking engine into the previously developed DockBench software, with the idea to make the proposed benchmark strategy available to the scientific community.

In the field of molecular dynamics, the Supervised Molecular Dynamics tool has been developed. The SuMD algorithm can be considered a fully flexible docking method, which explores the approach of a ligand towards the target binding site along time. The difficulties related to the simulation of this event with classical molecular dynamics simulations are related to the long timescale of the process. SuMD has been observed to accelerate this event of at least 2 orders of magnitude, giving the possibility to do a flexible docking experiment to anyone is equipped with a low-cost GPU machine. The strength of this approach is the possibility to individuate metastable binding sites along the recognition pathway and to analyze the role of waters during the recognition.

The domain of applicability of SuMD is under investigation, but a precious implementation has been developed to study peptide-protein binding. At the moment, the test cases have included rigid peptides (well-structured peptides and cyclic peptidomimetics), while more flexible structures are still challenging. Currently SuMD is still in development, with the aim to broaden its application to a wider pool of test cases.