In silico modeling of protein-ligand binding

Alejandro Varela Rial

TESI DOCTORAL UPF / year 2021

THESIS SUPERVISOR
Dr. Gianni De Fabritiis
Department of Experimental and Health Sciences.
Universitat Pompeu Fabra

To those who embrace doubt and have the courage to challenge their most profound convictions.

Thanks

I want to thank all my colleagues at Acellera, Computational Science Laboratory and GPCR Drug Discovery group, for being patient and kind when teaching me the intricacies of this field, as well as for their support and friendship. I would certainly not have made it here without them. I also want to thank my supervisor, for giving me the opportunity, the trust, the guidance and the intellectual freedom to explore new ideas.

Moreover, I have to thank Marisa Castiñeira, for thinking out loud. I also need to thank Ana for her immense support over these years. Finally, I want to thank my family, they are an inspiration and a role model of humility, bravery and hard work.

The first principle is that you must not fool yourself and you are the easiest person to fool.

— Richard Feynman

Abstract

Protein-ligand binding prediction is one of the cornerstones of computational chemistry. A software capable of accurately modeling the interaction between proteins and small molecules would have a large impact on the pharmaceutical industry, as it could replace expensive and slow in-vitro experiments, reducing costs and time. In this work, we have developed methods to model the binding mode of a ligand to its target, to assess the feasibility of that prediction with machine learning methods and to understand what features of the protein-ligand complex these machine learning methods are using to guide their predictions.

Resum

La predicció del mode d'unió proteïna-lligand és una de les pedres angulars de la química computacional. Un programari capaç de modelar amb precisió la interacció entre aquestes dues parts tindria un gran impacte en la indústria farmacèutica, ja que podria substituir els costosos i lents experiments in vitro, reduint costos i temps. En aquest treball, hem desenvolupat mètodes per modelar el mode d'unió d'un lligand a la seva proteïna, per avaluar la viabilitat d'aquesta predicció amb mètodes d'aprenentatge automàtic i per entendre quines característiques del complex proteïna-lligand utilitzen aquests mètodes d'aprenentatge automàtic per guiar les seves prediccions.

Preface

The physiological changes that pharmaceutical agents elicit on the body are mediated by the binding of the agent – a small molecule – to a host's macromolecule, typically, a protein. Hence, a profound understanding of this binding event would allow the development and design of better drugs. This is the fundamental premise of structure-guided drug design.[1, 2] Ideally, we would like to understand the precise interactions that the ligand establishes with the protein, the binding pathway that if follows, the strength with which the ligand binds to the protein and the changes that the protein conformation undergoes upon ligand binding.[3, 4]

The complexity involved in the process of protein-ligand binding cannot be overstated; the desolvation of the ligand [5], the displacement of structural waters [6, 7], the interplay of enthalpy and entropy components in the energetics of binding [8, 9] or the flexibility of the ligand and the protein [10, 11] are only some of the factors that play a role in this process. However, despite this complexity, there are some simple rules that this process must obey and that can be easily modeled, like steric constraints or electrostatic complementarity.[12, 13] This fact, combined with the impressive improvements in hardware of the last decades [14], has opened the possibility to develop algorithms that can model the binding of a small molecule to a target, like molecular docking [15, 16, 17, 18] or molecular dynamics simulations [19].

An accurate and fast in-silico protocol that could model such phenomena and estimate the binding affinity of the system would be of enormous value to the pharmaceutical industry, by virtue of substituting expensive and relatively slow high-throughput screening hardware with much more scalable computational hardware, as well as offering the possibility to evaluate virtual, not yet existent molecules, on demand, without the need to synthesize or purchase them.[20, 21] Hence, predicting these properties has become one of the holy grail of computational chemistry.

Far from being a vague promise, current state-of-the-art in-silico proto-

cols have already proven successful and valuable at tasks such as identifying active compounds among large libraries of mainly inactive compounds (virtual screening),[22, 23, 24] or at predicting the binding mode of a ligand to its protein. [25, 26, 27, 28] Yet, limitations in these tools prevent further success. The objective of this thesis is to tackle some of these limitations and develop new software applications to improve the modeling of protein-ligand binding.

Contents

Chapter 1

INTRODUCTION

One of the key properties of a drug is its affinity to the target protein.[4] Such property leads the efforts in the early stages of drug discovery, where the sole objective is to identify compounds that can bind to the protein or macromolecule of choice.[29] Later on, these hits will be evaluated in terms of other properties, like ADMET profile, synthetic accessibility or intellectual property rights [30, 31, 32], but affinity is a condition that the compound must meet since the very beginning of the drug discovery process until the end. Under these premises, the amount of effort that has been devoted to predict this property with in-silico methods is not surprising. The literature on this topic is rich and the number and diversity of algorithms, protocols and methodologies that are centered in solving this problem is the best evidence supporting the importance of this task.[15, 16, 33, 34, 35, 36]

These algorithms usually follow one of two main paradigms: structure-based or ligand-based. Structure-based methods rely on the principle that the affinity of the ligand to its target can be approximated by accurately scoring the different interactions that the ligand establishes with the target, like hydrogen-bonds or pi-stacking contacts.[12, 37] This family of methods requires, of course, a structure where the interactions between protein and ligand are specified. This structure can be experimentally de-

termined, with methods such as NMR or X-ray [9] or predicted with software like molecular docking [38, 18] or molecular dynamics (MD) simulations [39, 40]. On the other hand, ligand-based methods rely on the principle that structurally similar compounds bind in similar ways,[34, 35] and that affinity can be estimated by leveraging this similarity, without the explicit use of the protein. Canonical examples of this paradigm are molecular fingerprints [34] and shape comparison methods [35]. Given that the methods presented in this work are all structure-based, we will focus on this paradigm.

1.1 In-silico binding mode prediction

Note: Parts of this section come from my review 3.3

Structure-based methods require a model of the protein-ligand complex to estimate its affinity. Unfortunately, although empirical methods exist that can provide high resolution structures, like X-ray, NMR or cryo-electron microscopy, [9, 41] they are expensive, time-consuming and lack the throughput necessary to provide structures for large numbers of compounds.[42] For that reason, in-silico methods have been developed to predict the binding mode of a ligand to a protein.

A binding mode predictor must succeed in two tasks: sampling a diverse set of possible binding modes (exploration), and evaluate them correctly based on their energetic profile (scoring). Depending on how extensive the sampling of binding modes is, and the number of terms and atoms involved in the scoring, the speed or throughput of each algorithm will vary. Hence, they can be classified as fast and slow.[43]

1.1.1 Fast: Molecular docking

Fast methods, referred as molecular docking, have become very popular because they can predict the binding mode of a compound and its affinity to the target in a matter of seconds or minutes, allowing them to screen large chemical libraries in a reasonable time. Some of the most popular docking programs include DOCK,[15] GOLD, [16] Glide,[17] AutoDock Vina,[38] and rDock,[18]. Docking programs predict the binding mode of small molecules in the binding site of a protein in terms of its location, orientation, and internal conformation. The protein is considered rigid or semi-rigid, allowing only the movement of selected groups to improve hydrogen bonding.[18, 16] Docking software uses a search algorithm to explore different ligand poses which are then evaluated by a scoring function. This function tries to approximate the free energy of binding by incorporating the enthalpic (scoring positively interactions like hydrogen bonds or salt bridges while penalizing clashes) and entropic terms, essentially penalizing highly flexible ligands based on their number of rotatable bonds [12] and by identifying hydrophobic contacts between protein and ligand that could lead the expulsion of waters from the cavity with the subsequent entropy increase. [44]

Because the exploration part of docking algorithms is typically limited to the ligand's degrees of freedom (location, orientation and dihedral angles), and the location dimension is sampled only in the vicinity of a given binding site, a large number of compounds can be evaluated in little time. Furthermore, regarding the scoring, interactions with waters are not considered explicitly, and only interactions between ligand atoms and nearby residues in the defined binding site are taken into account, saving more time. This allows for immense throughput [28], however, by neglecting protein flexibility, [10, 45], the role of structural waters [6, 7], coordinating metals [46], or desolvation effects [5], the accuracy of the predictions suffers.

Previously, we mentioned that ligand-based methods rely on the princi-

ple that structurally similar compounds bind in similar ways. A particular subfamily of docking algorithms, known as template docking or scaffold docking, exploits this principle to model the binding mode of a compound using a similar molecule, for which its binding mode to the protein is known, as a template or guide. One of the methods we describe in this thesis, SkeleDock (section 3.2) does precisely this. These methods can improve dramatically the pose prediction accuracy but they require the existence of a compound similar to the query for which the binding mode is known. Although such knowledge is not always available, the rise in popularity of fragment based drug discovery [47], and methods like crystal soaking [48], might increase the number of scenarios to which template docking might be applied.

1.1.2 Slow: Molecular Dynamics

Slow methods, based on molecular dynamics, offer a more realistic and detailed view of the binding event. MD simulates the dynamics of a system through time by means of Newton's equations of motion and a force field, which describes the interactions between the different atoms. In reality, these methods range from slow, microsecond long, all-atom MD simulations [39] through adaptive sampling [49] to faster coarse-grained simulations [50] and biased MD methods [40]. Most of these methods can provide valuable insights, like the binding and unbinding pathways that a ligand follows [51, 52, 53] or the high-energy barriers that it crosses [39]. Each conformation sampled by the MD represents a binding mode, and its feasibility can be directly derived from the forcefield, although more sophisticated approximations exist to evaluate the sampled binding mode, like metadynamics[54], dynamic undocking [55, 56] and Markov state models (MSM).[57]

Protocols leveraging MD simulations to predict binding mode and affinity have been successful [39, 52, 58, 59] and they can naturally account for challenging scenarios like water behaviour, protein flexibility or ion

coordination.[58] The major trade-offs of MD based protocols are the lack of high-throughput, the need for relatively expensive hardware (GPUs) and a steep learning curve.[43] However, some new and promising approaches have been developed to address some of these limitations [33, 49, 60], and if the improvements that we have seen in hardware in the last decades continue, these protocols could eventually meet the pace required by the industry. It is worth mentioning that until the appearance of GPUs, MD simulations were practically limited to the nanosecond-scale. [19]

1.2 Major limitations and challenges

Note: Parts of this section come from my review 3.3

We have now discussed the two main families of algorithms used to perform in-silico binding mode prediction: docking and molecular dynamics based protocols. In order to understand which tool might be better suited for a particular task, if any, it is important to be aware of both the limitations of the modeling tools and the underlying complexity of the reality we are trying to model. The two main challenges in pose prediction are protein flexibility and scoring or affinity prediction.

1.2.1 Protein flexibility

Several models exist that explain protein-ligand binding: lock-key paradigm, induced-fit and the more recent conformational selection theory.[3]

The lock-key paradigm is the oldest and arguably the simplest, as it assumes a rigid protein pocket to which a compound will bind by virtue of having a complementary shape to that of the pocket. Most molecular docking algorithms are based on this model and treat the protein as a rigid entity [61]. Hence, if the target of interest, or its pocket, is known to be very rigid or its flexibility is limited to a small degree of breathing [10],

a protein-rigid docking software or the use of a soft potential might be the best fit. Soft potentials [62] reduce or attenuate the penalty for steric clashes. Practically, this allows for minor clashes between the ligand and the protein, assuming that the protein will make room to accommodate the clashing atoms. Despite the fact that it can only account for small variations in protein structure, soft docking is an appealing way to account for protein flexibility. It is easy to implement, as it only requires manipulating the parameters of the scoring function, and does not reduce docking speed, which makes it practical in high-throughput applications. An equivalent approach is the modification of the VdW radii of the atoms in the protein, ligand, or both. [63].

The conformational selection paradigm incorporates protein flexibility and understands the protein as an ensemble of different conformations which are visited with different frequencies, including the one to which the ligand actually binds. This view can, to some extent, be easily incorporated into molecular docking by using several protein conformations, instead of just one. This approach is known as ensemble docking [64]. The ensemble of conformations can be obtained through side-chain rotamer sampling, MD simulations or from structural databases.[65, 66, 64]. The method comes, however, with its own limitations, namely, increased computational cost, the risk of docking ligands into conformations not visited by the protein in reality (artifacts), increased number of poses from which the right one will have to be selected, or issues with the generation of conformations and selection of the ensemble.[63, 60, 64]

Finally, induced fit postulates that, in some cases, it is the initial binding of the ligand that changes the protein conformation and allows the formation of the final protein-ligand complex [3, 42]. This last model is perhaps the most challenging one to account for and, although MD simulations might be the best fit for these scenarios [51], some docking software exist to tackle these use cases, too. [33, 63, 42]. For instance, the Induced Fit Docking (IDF) protocol described by Sherman *et al.* can capture induced fit effects in a computationally efficient manner.[63] In

this protocol, the ligand is first docked into a rigid pocket with a soft potential using Glide.[17] In addition to the use of a soft potential, up to three residues in the pocket can be mutated to alanine if their position is suspected to be ambiguous, allowing the ligand to occupy that space. In the second step, Prime software [67, 68, 69] explores different conformations of the protein structure in docked complexes by means of side chain sampling, followed by structure minimization. A second round of docking follows, where the ligand is redocked to the protein conformations sampled in the first round, this time using a hard potential. The final docked solutions are then scored accounting for the docking energy, the strain of the protein and the solvation terms.

Another protocol able to handle induced-fit effects is Protein Energy Landscape Exploration (PELE).[33] PELE is a search protocol which includes protein and ligand flexibility explicitly. Local perturbations of the ligand are introduced via displacements, rotations or changes in dihedral angles, followed by side chain sampling. The best conformation is minimized, which allows the movement of the backbone, and scored, leading to the acceptance or rejection of that conformation. Finally, a new round of local perturbation begins, starting a new cycle. This series of cycles result in a collection of highly redundant local minimas (a trajectory), however, large changes do occur from the start of the trajectory until its last frame. Although much slower than docking, PELE is still orders of magnitude faster than a typical MD run. Authors have reported success in reproducing entry and exit pathways of several protein-ligand complexes, and in sampling protein-ligand poses in close proximity to crystal structures for very challenging targets such as GPCRs.[70]

1.2.2 Accurate scoring

Although several binding modes might exist for a given protein-ligand combination, the one with the lowest energy will be the most populated one and, as a consequence, the one most likely to be responsible for the

therapeutic effect. The energy of the binding mode is determined by the balance between two components: enthalpy and entropy.[8, 9] The enthalpic component is determined by the interactions that the ligand creates with the protein upon binding to it, like hydrogen bonds or van der Waals contacts, but also by the interactions that it loses with the solvent (water). Hence, the interactions established with the protein must compensate the ones lost with the solvent in order for the enthalpic component to be a positive contributor for binding. The entropic component is related to disorder; when the ligand binds to the surface of the protein, the waters that were bound to the surface are displaced, usually leading to a more disordered system. However, the protein-ligand bound system has become more ordered, as they both lose rotation, translation and conformational freedom. [9] For a ligand binding process to occur, either the two components have to favour the bound state, or one has to be large enough to compensate the other.

The energy and, hence, the feasibility of a binding mode can then be estimated by approximating those two components. The enthalpic component is typically approximated by identifying the non covalent interactions that the ligand establishes with the protein and giving a weight to each, for instance, salt-bridges are more favourable than hydrogen-bonds or pi-stacking interactions.[71] The entropic component is perhaps more challenging, and its mostly accounted for by counting the number of rotatable bonds that the ligand has, as that influences the different number of microstates (rotamers) that the ligand can adopt [72] and by identifying hydrophobic contacts between the protein and the ligand, as such contacts would displace structural waters in the pocket, leading to a higher disorder. [44] These are, of course, simplifications that are far from perfectly modeling the complex scenario that we depicted in the previous paragraph, however, they are needed in order to obtain results in reasonable timescales.

As with pose prediction methods, there is a large spectrum of proposed algorithms that balance accuracy and throughput to score a pose or pre-

dict its energy or binding affinity. For instance, a Markov state model can be built from a series of MD simulations to estimate binding energy and kinetics by identifying the bound state and the bulk state and computing the transitions between the two. This has proven successful [73] but it is very computationally expensive and requires some degree of human supervision.

The scoring functions used by molecular docking algorithms are, in contrast, much faster. Essentially, these functions are constructed by defining a set of rules based on atom type pairs and distances; for instance, one can group all protein and ligand atoms by their pharmacophoric profiles (aromatic, hydrogen bond donor and acceptor, hydrophobic, negatively or positively charged) and create a set of functions to penalize unfavourable contacts (two negative charges together) and reward favorable ones (two aromatic groups close by). Weights for each contact can be derived via statistical methods, like linear regression [74] or more sophisticated, non-linear machine learning models [75] using binding affinity data, or by measuring the frequency of such contacts in crystal structures [76]. Alternatively, physics-based scoring functions approximate the binding energy with just a couple of terms, namely, the sum of the electrostatic and van der Waals interactions; although recently, some physics-based methods have included terms to approximate entropy and solvation effects to better estimate binding affinity.[77]

These scoring functions are orders of magnitude faster than those based on MD methods and affinity can be estimated rapidly for any structure allowing for immense throughput, which is the reason why they are used by docking algorithms. Modern docking software like VINA [38] and rDock [18] can evaluate libraries of hundreds of thousands of compounds in a matter of days with a modest number of CPUs using such scoring functions. This speed has allowed the screening of ultra-large libraries, involving more than 170 million compounds.[28]

Although in terms of speed, docking scoring functions can be said to

be close to the requirements of the industry, in terms of accuracy, they are still far away. Current state-of-the-art scoring functions show poor correlation with real affinity data, they have little power to discriminate actives from non actives, and usually, cannot distinguish the true binding mode from other reasonable poses.[12] This should not come as a surprise considering the major simplifications that scoring functions assume. Nonetheless, successful prospective virtual screening campaigns have been reported using such functions [25, 26, 27, 28] which suggests that their performance might be dependent on some variable like the target class or family [78, 79] or whether the binding of compound is entalphy or entropy driven.[8] Hence, developing better scoring functions is an area of intense research and new approaches have been proposed which we discussed extensively in the review included in this thesis (section 3.3) and which we now summarize in the following sections.

Exploiting prior knowledge

Going back to the "similar compounds bind in similar ways" principle, one way to improve scoring accuracy is by evaluating the overlap of the proposed binding modes against the structure of a known binder. If the contacts that the reference binder establishes with the protein are mirrored or cloned by the query compound, it is reasonable to assume that it has a higher chance of binding with high affinity. Following the lock-key analogy, if a compound is able to push the same pins of the lock as the key, then it might be capable of opening the lock.

There are two main ways of evaluating this similarity: with a pharmacophore,[80, 81] and using protein-ligand fingerprints.[82, 83] Pharmacophores can be built by defining a series of locations in space where a given pharmacophoric feature is expected within a tolerance radius. If the query ligand places the right feature inside the sphere, a reward is given. On the other hand, protein-ligand interaction fingerprints encode the binding mode as a sequence of bits, where each bit describes the type of inter-

action, the atoms involved in the interaction and the role of participating atoms (i.e., acting as hydrogen bond donor or acceptor).[84] Fingerprints can then be constructed for the reference and query compounds and evaluate their similarity easily. Da and Kireev[82] reported that using a protein-ligand interaction fingerprint (named SPLIF) was the best method at discriminating active compounds from decoys in 10 targets of the DUD-E database.[85]

Pharmacophores and fingerprints benefit from existing knowledge, and, hence, require the existence of a structure of an active bound ligand, which is not always available. One of the articles presented in this thesis, LigVoxel (section 3.1) aims to obtain an accurate pharmacophore, or, more precisely, a pharmacofield, based on the protein structure alone, circumventing the need for an active-bound complex structure. LigVoxel returns 8 cubes or matrices of side size 16 Å. Each cube accounts for one pharmacophoric property (hydrophobic, aromatic, hydrogen bond donor, hydrogen bond acceptor, positively charged, negatively charged, metal and excluded-volume). Each voxel in this cube or matrix has a value ranging from 0 to 1, signifying whether or not the given pharmacophoric feature is expected at that location. This is what we refer to as pharmacofield.

Another way to tackle limitations in scoring functions is the addition of restraints. This can be achieved by adding new terms to the scoring function to reward a particular contact. Such an approach was used by Caporuscio *et al.* [86] to force an interaction between an aromatic nitrogen of bifonazole and a heme iron in the pocket of cytochrome P450, which stand-alone Glide was failing to capture appropriately. After the addition of the constraint, the binding mode was correctly predicted with an RMSD value of 1.02 Å. Adding restraints can compensate certain deficiencies in the scoring function itself, and it also can further limit the exploration of certain poses, reducing the chances of sampling unreasonable binding modes, which could end up being ranked as the best. In that sense, scaffold docking algorithms, like SkeleDock [87], constitute

convenient protocols to greatly limit the number of explored poses by restraining the atoms of the scaffold that query ligand and reference compound share to occupy the same place. As long as the binding site of the common scaffold is conserved, scaffold docking algorithms can greatly improve the accuracy of binding mode prediction.

Leveraging Molecular Dynamics

More precise -and expensive- computational methods can be used to better assess binding mode and affinity predictions. For instance, some methods evaluate the stability of a docked pose through metadynamics[54] or dynamic undocking.[55, 56] These protocols assume that compounds with great affinity are strongly anchored to the protein and will remain in the same pose after a series of short simulations, while poorly predicted poses will leave the pocket. Other methods are based on binding free energy estimations and have shown success improving hit rates in virtual screening campaigns.[88, 89, 90] Free energy methods went through major improvements over the last years,[91, 92, 93] with free energy perturbation (FEP) standing out. FEP methods compute the binding free energy of a ligand by performing an MD simulation with the simultaneous alchemical transformation from one ligand to a reference structure with known affinity.[94, 95, 96] These methods can reach affinity predictions accuracies that fall in the range of 1-2 kcal $\cdot$ mol^{-1}, [97, 98], which allows medicinal chemists to effectively shortlist the set of compounds for synthesis and testing.

Adopting computer vision

Recently, thanks to advancements in hardware and, more importantly, to the impressive growth of deposited structures of protein-ligand complexes in the Protein Data Bank [99], modern machine learning algorithms have entered the scene with great success.[100, 75, 101] Particularly, convolutional neural networks (CNN) have recently been applied to the scor-

ing problem due to their success in the field of computer vision. Here, the protein-ligand complex is converted into a discretized 3D space composed of voxels, where the value of each voxel depends on the proximity and type of the nearest atoms around it. The resulting matrix can then be passed to the CNN, either to train it, or to obtain a prediction. K_{DEEP} [75] was one of the first applications of CNNs to the scoring problem, obtaining a Pearson's correlation coefficient of 0.82 with real binding affinity data. Furthermore, it was also validated in the D3R Grand Challenge 4th edition, where it ranked first in two blind sub-challenges. One of the articles presented in this thesis, SkeleDock (section 3.2) describes the participation in this contest.

Despite achieving great success, machine-learning based methods have their own caveats. One such caveat is their black-box nature; while traditional scoring functions are based on human-engineered functions, CNNs develop their own functions during training, which are hard or impossible to interpret, requiring the researcher to trust the result at face value. The last article of this thesis, Glimpse (section 3.4), describes an application that tackles this problem and improves the explainability of CNNs.

1.3 In-silico modeling in current drug discovery

Note: Parts of this section come from my review 3.3

Despite these limitations, fast docking protocols have proven successful in a number of prospective studies [25, 26, 27, 28], and they have done so for the right reasons; that is, by predicting the right pose and assigning its compound a great affinity that ranked it amongst the best in the library [102]. However, one could argue that such success might not be directly attributable to docking itself, but rather to the additional steps taken by the practitioner, such as visual inspection, the addition of prior knowledge or

rescoring with alternative methods. This would mean that docking per-se is not that powerful, showcasing it as a simple ”pose generator” that requires further work to provide meaningful results. However, unsupervised or largely-unsupervised docking campaigns have been successful too.

In a recent article, Lolli and Caflisch used the docking software SEED to screen 350 fragments against the target BAZ2B.[103] Of these, 12 fragments were prioritized for testing. Soaking experiment resulted in 4 fragments bound in the binding pocket. Furthermore, the authors report that the binding mode predicted by their docking software was correct for three of the four fragments, two of them being practically identical according to the figure. This example is particularly impressive for two reasons: (1) Fragments are believed to be particularly challenging for docking algorithms [60] and (2) there was no use of prior knowledge nor visual inspection.

In another work, Hermann *et al.* discovered a substrate of an orphan enzyme Tm0936 by docking a library of metabolites into its catalytic site.[104] They performed docking using DOCK3.5.54 into an apo structure, which is more challenging than docking to a ligand-bound conformation. The molecule which was then confirmed to be the substrate (SAH) was ranked 6th among 4,207 molecules. Moreover, the crystal structure of the selected molecule agreed almost completely with the docking prediction, highlighting the success of such approach.

Another valuable example is the initiative of Continuous Evaluation of Ligand Protein Prediction (CELPP).[105] In this program, a set of fully automated docking protocols is challenged with predicting the binding modes of a new group of protein-ligand complexes on a weekly basis. The ligands are docked to a structure of a close homolog bound to a similar molecule, which makes it a difficult cross-docking experiment. Although it is not a typical virtual screening campaign, it offers valuable information regarding docking power for several reasons. First, the docking protocols are fully automated, which allows for a more fair comparison

among them and removes the impact of the human expert. Second, the evaluation is blind and prospective, therefore unintended biases do not affect the results. Finally, the challenge showcases the strengths and limitations of using docking as an out-of-the-box tool, without any manual parameter tweaking. Although the number of adopters is still low, some protocols have already docked over 3000 molecules. One of them, named *kauoh*, has a Q1 value of RMSD 1.81, meaning that a large fraction of the predicted poses were under the cutoff value of 2.0 Å.

Despite this success, it is true that classical docking algorithms, due to their need to be fast paced, will probably not be able to deal with particular challenging scenarios, like those where structural waters or coordination atoms play a key role, or those involving cryptic pockets [106] or major changes in pocket conformations [107]. This is where MD based methods have an edge, as they can naturally account for all these scenarios, with the additional benefit of providing valuable mechanistic and thermodynamics insights, like reconstructing binding pathways, shedding light into the desolvation process, or identifying high-energy barriers or metastable states. Some of these insights can not even be obtained by crystallography experiments.

The two paradigms, fast and slow, can easily be combined and used in different stages of the virtual screening campaign: while docking can filter the vast majority of the library, MD can provide more accurate predictions for a subset of it, re-scoring the best compounds or providing mechanistic insights into their binding pathways. [108]

Most drug discovery projects today rely to some extent in in-silico modeling [109, 110], typically combining different approaches and techniques (ligand based and structure based virtual screening, FEP, ADMET prediction, etc.) in parallel or in a hierarchical fashion [111]. Several companies performing in-silico modeling have appeared in the market, and those using AI to help drug discovery raised $1 billion in funding in 2018 [112], signaling a strong trust in computational methodologies.

With the advent of new protocols, improvements in hardware and the adoption of new technologies, like artificial intelligence, and the increase of structures and binding affinity data for more protein-ligand complexes, in-silico methods will keep improving and becoming progressively more valuable for drug discovery. The recent success of AlphaFold [113] in protein folding is perhaps the best proof of how powerful in-silico approaches can become.

Chapter 2

OBJECTIVES

The objective of this thesis is to address some of the limitations we have presented in the introduction. Namely:

1. Improve accuracy in protein-ligand binding mode prediction in scenarios where the pose of a similar compound is known (SkeleDock).

2. Validate the capacity of a pharmacofield generator (LigVoxel) to identify the right pose among a set of decoy poses.

3. Improve the interpretability of scoring functions based on CNNs (Glimpse).

Chapter 3

PUBLICATIONS

3.1 LigVoxel: inpainting binding pockets using 3D-convolutional neural networks

Skalic, M., Varela-Rial, A., Jiménez, J., Martínez-Rosell, G., de Fabritiis, G. LigVoxel: inpainting binding pockets using 3D-convolutional neural networks. Bioinformatics. 2019; 35(2):243-250.

Summary

In this article we described LigVoxel, a convolutional neural network autoencoder that predicts a pharmacofield for a given protein pocket. My contribution was limited to designing and executing experiments to evaluate the scoring capacity of the predicted pharmacofields, which we did by measuring their capacity to recover the right binding pose from a set of decoy poses.

Bioinformatics, 35(2), 2019, 243–250
doi: 10.1093/bioinformatics/bty583
Advance Access Publication Date: 6 July 2018
Original Paper

OXFORD

Structural bioinformatics

LigVoxel: inpainting binding pockets using 3D-convolutional neural networks

Miha Skalic[1], Alejandro Varela-Rial[2], José Jiménez[1], Gerard Martínez-Rosell[1] and Gianni De Fabritiis[1,3,*]

[1]Computational Science Laboratory, Universitat Pompeu Fabra, Barcelona Biomedical Research Park (PRBB), [2]Acellera, Barcelona Biomedical Research Park (PRBB), Doctor Aiguader 88, 08003 Barcelona, Spain and [3]Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig Lluis Companys 23, 08010 Barcelona, Spain

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on February 19, 2018; revised on May 23, 2018; editorial decision on July 3, 2018; accepted on July 4, 2018

Abstract

Motivation: Structure-based drug discovery methods exploit protein structural information to design small molecules binding to given protein pockets. This work proposes a purely data driven, structure-based approach for imaging ligands as spatial fields in target protein pockets. We use an end-to-end deep learning framework trained on experimental protein–ligand complexes with the intention of mimicking a chemist's intuition at manually placing atoms when designing a new compound. We show that these models can generate spatial images of ligand chemical properties like occupancy, aromaticity and donor–acceptor matching the protein pocket.

Results: The predicted fields considerably overlap with those of unseen ligands bound to the target pocket. Maximization of the overlap between the predicted fields and a given ligand on the Astex diverse set recovers the original ligand crystal poses in 70 out of 85 cases within a threshold of 2 Å RMSD. We expect that these models can be used for guiding structure-based drug discovery approaches.

Availability and implementation: LigVoxel is available as part of the PlayMolecule.org molecular web application suite.

Contact: gianni.defabritiis@upf.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Many methods in structure-based drug discovery focus on exploiting structural information of protein pockets in order to select and optimize possible ligand binders. A three dimensional spatial featurization of proteins and small molecules and a metric on such space is usually utilized to characterize the pocket or ligand. Shape similarity, for instance, is commonly used in several virtual screening pipelines as is the case of the volumetric aligned molecular shapes (VAMS) algorithm (Koes and Camacho, 2014) and the ROCS software (Hawkins *et al.*, 2007). Docking procedures such as PharmDock (Hu and Lill, 2014), use this type of information to refine poses. Another common application where three dimensional features are used is *de novo* drug-design, in order to build a binding site model with the intent of designing a novel ligand taking into account some constraints (Yuan *et al.*, 2011). Grid generation is also central in the elucidation of pharmacophores, such is the case of FLAPpharm (Cross *et al.*, 2012) and VolSite (Desaphy *et al.*, 2012), and their search (Ebalunode *et al.*, 2008).

Whether optimal features can be automatically learned from the data rather than imposed remains an open question. In the last few years, deep learning based approaches (Goodfellow *et al.*, 2016), have demonstrated state of the art performance on tasks such as 2-D image classification (Krizhevsky *et al.*, 2012), natural language processing (Goldberg, 2015), speech recognition (Graves *et al.*, 2013) and more. The promise of deep learning in computational biology, however, is still yet to be fully developed (Angermueller *et al.*,

2016), but is increasingly common. Proteins, for instance, can be seen from a computer vision point of view, as if they were three-dimensional images after featurization. Variants of this approach have already been applied in protein binding site prediction (Jiménez *et al.*, 2017), protein–ligand binding affinity prediction (Jiménez *et al.*, 2018), large-scale active-inactive classification of drug-like compounds (Ragoza *et al.*, 2017) and quality assessment of protein folds (Derevyanko *et al.*, 2018).

This work shows how end-to-end deep convolutional neural network (DCNNs) models (LeCun *et al.*, 1998) can generate ligand fields given the structure of a protein binding site. Unlike previous knowledge-based and hand-crafted rules approaches, the proposed model is purely data-driven with the intention to mimic a chemist's intuition at manually placing atoms when designing a new compound. Furthermore, one can adjust the properties of these generated fields by tuning several input parameters. We treat proteins and ligand structures from a computer vision perspective, by voxelizing them according to different physio-chemical properties, and show that the learned and predicted representation can be useful in several related applications by evaluating it against unseen co-crystal structures.

The proposed methods can be considered complementary to molecular interaction fields (MIF) methods such as GRID (Carosati *et al.*, 2004), where an interaction energy is calculated between a target molecule and a probe, which may be an atom or a group. However, there are key differences: MIFs create volumetric maps by iteratively placing the probe at each point and calculating the interaction energy with the macromolecule. In contrast, the method proposed here takes as input the whole area at once and produces all outputs simultaneously. This is done in an end-to-end fashion. Values representing protein shape are multiplied by model's weights generating latent feature maps. The operation is repeated sequentially, until an output feature map is generated. While MIFs predict interaction energies, the method described here gives probabilities of seeing atoms with a property at each point.

2 Approach

Both the protein cavity and the ligand are represented spatially via their voxelization into a discretized 1 Å cubic grid of side size 16 Å as in Jiménez *et al.* (2017). The influence of each atom to each voxel is determined by its atom type and their respective distance r:

$$n(r) = 1 - \exp\left(-(r_{\text{vdw}}/r)^{12}\right), \quad (1)$$

where r_{vdw} is the corresponding van der Waals radius of a particular atom. The pair correlation function in Eqn.1 assigns voxels in the featurization box that are close to atoms values close to 1, while voxels far away from the atoms will approach 0. Each of the voxels is featurized with information according to the physio-chemical nature of its neighbouring atoms depending on eight predefined rules defined in Table 1. Atom types were determined based on Autodock 4 (Morris *et al.*, 2009) definitions. Voxelization algorithm is available in Supplementary Material and all descriptor computations are available as part of the HTMD Python framework for molecular dynamics (Doerr *et al.*, 2016). Voxelization rules for defining these properties are equivalent for both protein and ligand. A schema on how our method works can be seen in Figure 1: Starting with a target protein pocket (A), and optionally setting up ligand design parameters such as the number of aromatic carbons or h-bond donors (B) we can then featurize the protein pocket using the detailed *channel* descriptors (C). This featurized pocket is then used as input for a trained 3d-convolutional neural network (D), which produces the desired spatial ligand properties (E). The work presented here focuses on generating four voxelized ligand properties: an approximation to the ligand shape (total excluded volume), aromatic rings and hydrogen-bond donors and acceptors.

3 Materials, methods and evaluation procedure

3.1 Datasets

Our model was both trained and evaluated on the 2013 release of the sc-PDB database (Desaphy *et al.*, 2015), which provides 9190 high-quality protein–ligand co-crystals extracted from the PDB (Berman *et al.*, 2000). For further evaluation, we also use the Astex diverse set, which includes 85 protein–ligand complexes considered to be of pharmaceutical interest (Hartshorn *et al.*, 2007) and has been previously used in several studies for docking software benchmarking (Davis and Baker, 2009; Ravindranath *et al.*, 2015).

In order to provide a fair benchmarking of the method, a cross-validation scheme based on protein sequence similarity was used. The goal was to test how well the proposed method generalizes to proteins that differ from the ones used in training. We extracted the latest available sequence similarity clusters from the PDB (Berman *et al.*, 2000) website using a 90% similarity cutoff and assigned each of the proteins available in the scPDB database to a particular cluster only if all of its chains belong to the same one, while proteins with chains belonging to different clusters are discarded. This procedure yielded a total of 8808 protein–ligand complexes belonging to 3305 different sequence similarity clusters that were then randomly grouped together into $k = 10$ splits with sizes ranging between 525 and 1230 complexes that we evaluate using k-fold cross-validation. In practice this procedure implies that k models are trained using all splits minus a single one, used for evaluation. All PDB IDs used in each split are available as Supplementary Material. In the case of the Astex diverse set, we make the sequence cluster difference between proteins in the scPDB database and the former for training, yielding a training set of size 7147.

Table 1. Rules for atom physical chemistry properties

Property	Rule
Hydrophobic	Aliphatic or aromatic C
Aromatic	Aromatic C
Hydrogen bond acceptor	Acceptor 1 H-bond or S Spherical N Acceptor 2 H-bonds or S Spherical O Acceptor 2 H-bonds S
Hydrogen bond donor	Donor 1 H-bond or Donor S Spherical H with either O or N partner
Positive ionizable	Gasteiger positive charge
Negative ionizable	Gasteiger negative charge
Metallic	Mg, Zn, Mn, Ca or Fe
Occupancy (Excluded volume)	All heavy atoms

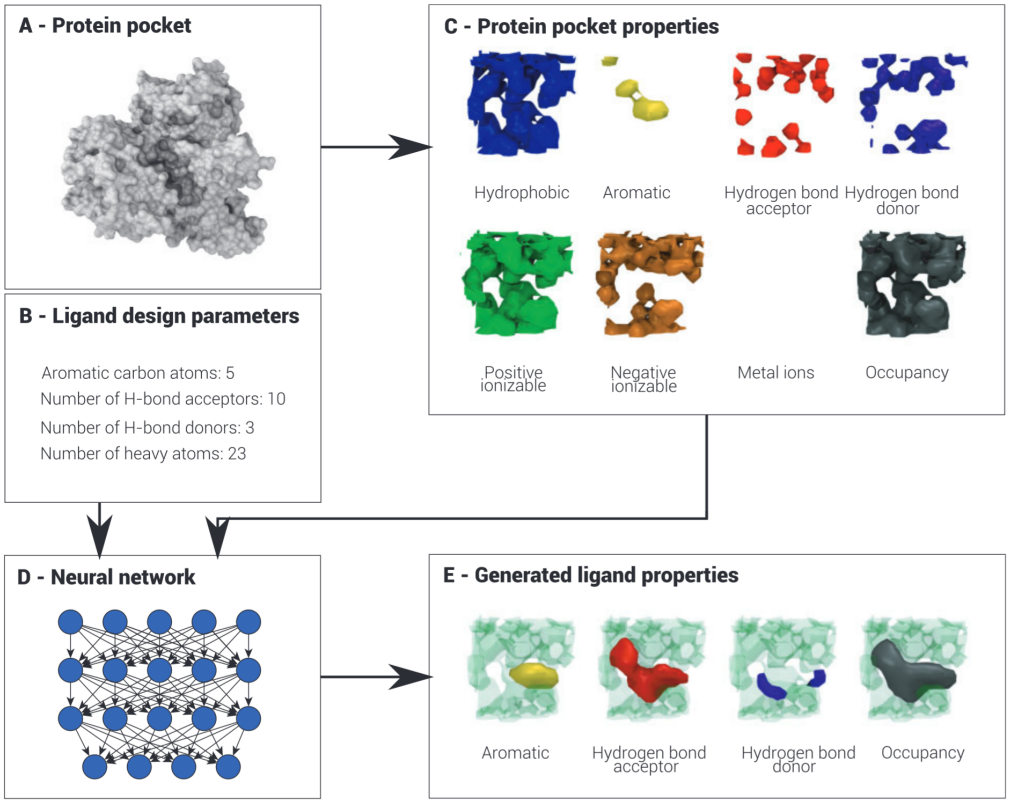


Fig. 1. Pipeline for generating ligand properties. We first voxelize the proteins (**A**), optionally providing several design parameters (**B**), according to eight different properties (**C**) then pass this input to a trained DCNN model (**D**), which will then predict the corresponding ligand properties (**E**)

3.2 Neural network architecture and learning

In order to train the proposed architecture, we aim to minimize the following two-part loss:

$$\mathcal{L}(y,\hat{y}) = \frac{1}{n}\sum_{i=1}^{n} -y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i) + \mathcal{D}_{\mathrm{KL}}(y|\hat{y}) \tag{2}$$

where

$$\mathcal{D}_{\mathrm{KL}}(y|\hat{y}) = \frac{\lambda}{n}\sum_{i=1}^{n}\left(\frac{1}{C}\sum_{c=1}^{C} s(y_{ic}) \log \frac{s(y_{ic})}{s(\hat{y}_{ic})}\right), \tag{3}$$

n corresponds to the total number of samples, $c \in C$ to a particular and total number of channels, respectively, and s to a sorting operation. The first term in the expression corresponds to the binary cross-entropy (De Boer *et al.*, 2005) loss function commonly used in classification tasks, and whose goal is to minimize incorrect label predictions, while penalizing extremely confident wrong ones. The second term is a discrete approximation to the Kullback–Leibler (KL) divergence (Kullback and Leibler, 1951), which in turn aims to preserve a good approximation between predicted and actual output distributions. Finally, λ is a user-set parameter that controls the balance of each term to the overall loss. Intuitively, if using only binary cross-entropy, when the model is unsure where to put a property point it would give a low probability across the whole volume. By adding KL loss this low value distribution is incentivized to give points with higher probabilities, proportional to number of atoms specified in the input.

Two different types of model were trained in this work, named conditional and unconditional. The rationale behind the distinction is that in conditional models an additional four atom counts input was used (hydrogen donor and acceptor, aromatic and total number of atoms), while in the unconditional one we choose not to provide this information. The former uses both terms of the loss, with the second (corresponding to the KL-divergence) forcing the model and its additional input to place a fixed number of ligand fields in its predicted output. By adjusting or providing different additional channels to the same pocket channel inputs the conditional model can therefore be fine-tuned. In contrast, the unconditional model only uses the first part of the loss (corresponding to binary cross-entropy). Our proposed model architecture is presented in Figure 2, and code snippet is available in Supplementary Material. Model consists of seven sequential and independent layers of convolution, activated by a rectified linear unit (ReLU) non-linearity except the last one, which instead uses a sigmoid function, outputting values ranging between 0 and 1. The output of this last layer represents the

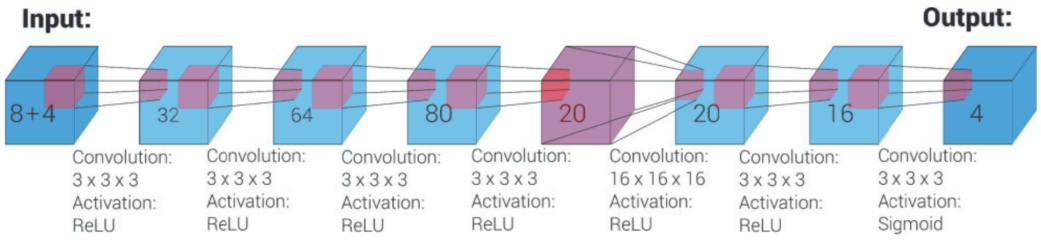


Fig. 2. Architecture of the proposed DCNN model for ligand property generation. The number inside each layer, pictured as blue boxes, represents channel count, while the inner boxes represent the size of the input (3 or 16 units) and output (1 unit) of convolutional kernels (Color version of this figure is available at *Bioinformatics* online.)

predicted representation. All convolutions in the model use a filter size of 3 units, as commonly used in image processing tasks, except for the one in the fifth layer with uses an extensive 16 units one, so as to capture long-distance relationships in the input. Zero-padding is used in all feature maps and the number of latent feature maps in each layer has been chosen so that the model could be trained on a single GPU with 8GB of VRAM. It is common in the literature to use the back-propagation algorithm and some form of stochastic gradient descent to train deep neural networks. For this study, we chose Adam (Kingma and Ba, 2014) with default parameters for momentum scheduling ($\beta_1 = 0.99, \beta_2 = 0.999$) as model optimizer. The neural network was then trained during 300 epochs using randomized batch size of 128 samples, introducing molecule rotations and a shifting of 0.5 Å from the ligand center during training so as to prevent overfitting and set $\lambda = 10$ after visual inspection of the trained models. Model learning and predictions were done in the Keras (Chollet, 2015) deep learning library (keras.io) and the Theano (Theano Development Team, 2016) tensor backend.

4 Results

4.1 Ligand design

The predictive model is capable of generating useful property fields that reasonably fill pocket space, while avoiding protein clashes. One can further control the proposed conditional model by providing different input parameters, in practice affecting the properties of the generated fields. Figure 3 shows how these changes modify the behavior of the algorithm by generating bigger grids as input values increase. Additional examples of varying property counts are depicted in Supplementary Figure S3. Although the model itself learns how to place ligand hydrogen-bond acceptors next to protein hydrogen-bond donors and viceversa, one common shortcoming we observed is that for properties where variability can be high, the model is not efficient at generating distinguishable clouds, typically predicting grids around the hydrogen-bond partner. This phenomenon can be partially explained by the fact that donors and acceptors are located close in space, but can be positioned in a wide variety of angles.

4.2 Evaluation procedure

To evaluate the quality of the predicted property fields, we analyzed the volumetric overlaps between crystal ligand features and predicted features. To that end, three different and complimentary tests were carried out, where we either measure or maximize the overlap between the predicted fields and the actual ligand representations. The overlap scores were used to evaluate: (i) sensibility of predictions, comparing it to random ones and values generated by MIFs, (ii) ability to discriminate between random and crystal poses, taking into the account rotation and translation and (iii), ability to discriminate crystal ligand conformers from generated ones. The goal of the proposed tests is not to prove the method's superiority compared to existing approaches, but rather to show and motivate that the architecture can flexibly be used alongside similar representations in the same or related tasks.

Test (i) was carried out using the scPDB database, while for (ii-iv) the Astex diverse set was used instead. Overlap is measured using a discretized version of the Jaccard Index:

$$J = \frac{\#|V_p \cap V_r|}{\#|V_p \cup V_r|}, \tag{4}$$

where V_p and V_r correspond to arrays of voxels of a particular channel of the predicted and crystal representations, respectively, after using a threshold of $P = 0.75$ to discretize both arrays. This metric naturally balances both correctness and size of the prediction. In our first test, Equation (4) was used to demonstrate that our method produces sensible ligand representations, using random predictions as baseline. Results for test (i) can be checked in Figure 4 and Supplementary Figure S1 (varying thresholds values) and S2 (separating folds). For each of the splits, we measure the Jaccard index between the predicted and computed values of the crystal structures and compare it with a random baseline. The conditional model, using both parts of the loss and atom counts from the ligand is significantly better at accurately predicting defined volumetric clouds than the naive unconditional model. In particular it can be appreciated that the latter exhibits higher variance than the former, suggesting that the second term of the loss is in fact helping preserve a match between the original and predicted distribution. Furthermore, it can also be seen that the model accurately predicts the total excluded volume channel, with Jaccard indexes ranging between 0.5 and 0.6 for most cases and around 0.2 for randomized baseline. Results for the aromatic channel are also satisfactory, again with indexes ranging 0.4 and 0.5 and for some splits even surpassing it (randomized baseline of 0.1). Hydrogen-bond donor and acceptor results, while certainly more modest than the others are still significantly better than the randomized baseline. Finally, MIFs were generated using EasyMIFs (Ghersi and Sanchez, 2009) a freely available software. Aromatic, carbonyl and hydroxyl probes were used to generate interaction fields and compared to aromatic, H-bond acceptor and H-bond donor ground truth voxels, respectively. As seen in Figure 4 the MIFs give a poor estimation of the ligand

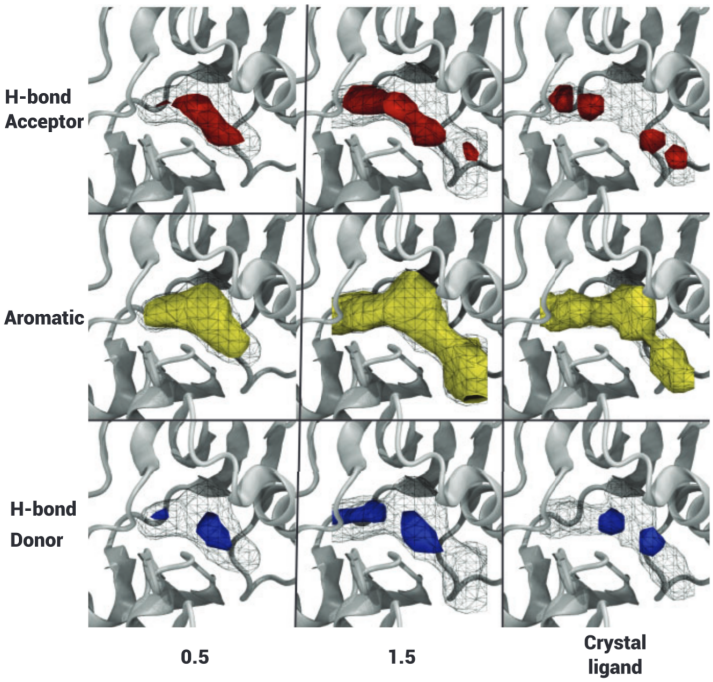


Fig. 3. Example of generated properties (PDB entry 1FPU). Ligand occupancy is displayed in wireframe while H-bond acceptors, aromatics and H-bond donors are in first, second and third row, respectively. The generated predictions shown in the column labeled 0.5 used half the atom counts of the cocrystallized ligand as the input. Column 1.5 follows the same logic, while the third shows actual cocrystallized ligand. As the atom count grows, the generated fields expand and are able to match more peripheral groups

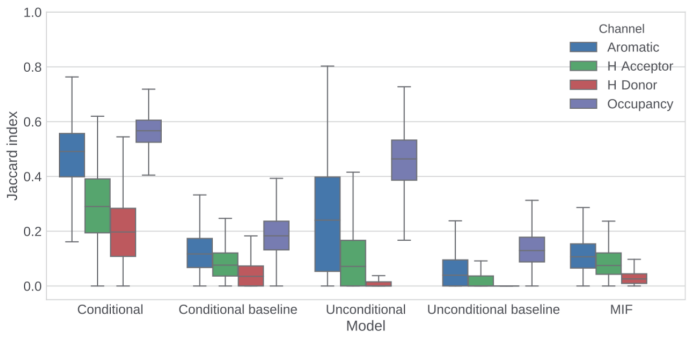


Fig. 4. Per complex results of predicted and actual physical chemistry values in the scPDB database using a Jaccard index, for the four channels of interest: aromatic, hydrogen-bond acceptor, hydrogen-bond donor and total excluded volume. The values are thresholded at 0.75 and the folds are pooled together. Per sample values in baseline groups are randomly shuffled within the sample and channel. Differences of all channels between groups conditional–conditional baseline, conditional–unconditional and unconditional–unconditional baseline are highly significant (*P*-value ≈0). Column MIF shows overlaps of the molecular interaction fields with the voxelization ground truth, at an energetic cut-off that maximizes the Jaccard index (Aromatic: $-5.10\,\mathrm{kJmol}^{-1}$, H-acceptor: $-6.12\,\mathrm{kJmol}^{-1}$, H-donor: $-6.33\,\mathrm{kJmol}^{-1}$)

fields. The Jaccard indices are comparable with the conditional model baseline.

In our second test, we measure the model's ability to recover the crystal pose by rotating the crystallized ligand in the pocket around its own geometrical center with a $30^{\circ}\Delta$ and measure the overlap between each rotated state and the predicted fields. All rotated states were generated starting from a random orientation of the ligand. Ideally, the intention is to see that rotated states close to the crystal pose are ranked significantly higher than those far apart when considering volumetric overlaps. Analysis were further expanded by

Table 2. Results for pose and conformer enrichment analyses (tests ii–iii)

	Top 1	Top 5
Rotations	70 (82%)	75 (88%)
Rotations and translations	64 (75%)	80 (94%)
Conformer enrichment	74 (93%)	76 (95%)

Note: For test ii we first only took rotations into account and then translations.

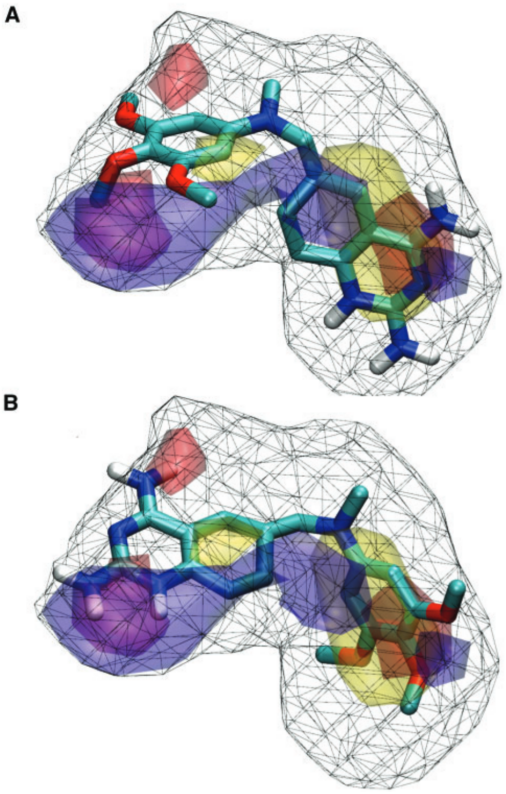


Fig. 5. Example of one wrong pose prediction. (**A**) Best ranked result using LigVoxel. (**B**) Native ligand pose. The excluded volume prediction is displayed in black wireframe, while aromatic and hydrogen bond acceptor and donor channels are displayed as clouds (Color version of this figure is available at *Bioinformatics* online.)

adding translations. In this last experiment the featurization window was slided ±3 Å in all three dimensions with a stride of 1 Å around the binding site.

In order to provide a single pose scoring function that does not depend on a threshold parameter, a second Jaccard-like index was defined using raw probabilities:

$$J_{\text{raw}} = \sum_{v \in V} \frac{V_p \otimes V_r}{V_p + V_r - (V_p \otimes V_r)}, \tag{5}$$

where V_p and V_r now represent the raw three dimensional subgrid with probabilities and property values, respectively, $\otimes$ is an element-wise multiplication and $v \in V$ is a voxel in a subarray. J_{raw} indexes

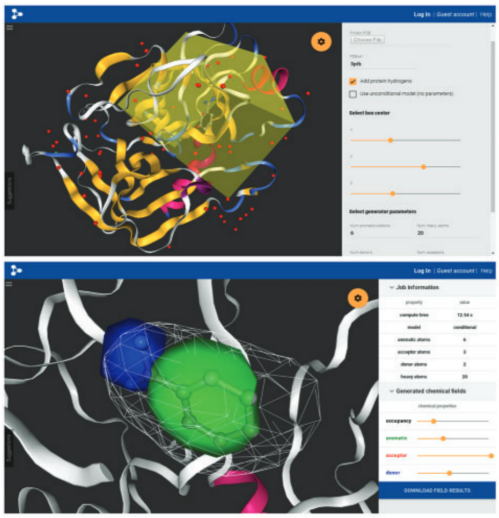


Fig. 6. Web application user interface. (Top) Pocket and parameter selection phase for Trypsin (PDB Id. 3PTB). (Bottom) Aromatic and H-bond donor iso-surface predictons with cocrystallized ligand superimposed

are then summed up: $J_s = \sum_i w_i J_{\text{raw},i}$, previously weighting each one by the prevalence of their corresponding property i:

$$w_i = \frac{\#\text{atoms with property } i}{\#\text{total atoms}}. \tag{6}$$

In translation tests, a clash penalty was introduced by multiplying J_s by a discount factor P function, representing the proportion of the ligand in clash with the protein:

$$J_c = J_s\Big(1 - \tanh(\phi P)\Big), \tag{7}$$

where we set $\phi = 2$ and

$$P = \frac{\#|V_{\text{prot,occ}} \cap V_{\text{lig,occ}}|}{\#|V_{\text{lig,occ}}|}, \tag{8}$$

with $V_{\text{prot,occ}}$ and $V_{\text{lig,occ}}$ representing protein and ligand occupancy voxels values, respectively, thresholded at $P = 0.75$. J_c can be considered as a discounted score of J_s, where the score is multiplied by a reward factor. The factor is 1 in case of no clashes and converges to 0 with bigger portion of the ligand volume clashing with the protein's.

We pick the best poses according to the detailed metric, taking into account possible ligand redundancies by discarding those poses with an RMSD below 2 Å with respect to any better scored pose. A pose was considered correct if the RMSD from its corresponding crystal was less than 2 Å. Results for test (ii) are presented in Table 2. After evaluation, an enrichment of scored poses below RMSD of 2 Å with respect to the crystallized ligand is observed: for the 85 protein–ligand pairs, and the 1728 total generated poses evaluated for each considering only rotations, those ranked first by the algorithm were under the 2 Å threshold in 70 pairs, while if we consider the 5 top ranked poses, the number of correct poses increases to 75. If translations are included, 64 and 81 are under the defined threshold for top 1 and top 5, respectively. This enrichment suggests that the proposed model and scoring function can effectively discriminate crystal poses. However, some ligand topologies represent

a harder challenge; Figure 5 shows one case (PDB Id. 1S3V, ligand TQD) of unsuccessful pose recovery. In this example, the shape of the molecule and its almost symmetrical distribution of aromatic and h-bond groups along the ligand allows it to accommodate inside the predicted excluded volume with the two orientations shown, while fulfilling the aromatic and h-bond prediction restraints. The placement of two consecutive aromatic rings near the bigger aromatic patch prediction suggests why this pose was scored higher than those closer to the crystal. Similar scenarios were found during our testing, suggesting that molecules displaying functional groups symmetry are harder to recover - but it remains unclear if these alternative poses are simply less likely, but still viable.

For our third test, we design a conformer enrichment procedure. To that end, 10 ligand conformers were generated using RDKit (Landrum, 2012). Only poses with at least 2 Å difference between them were kept. For each generated conformer including the crystal conformer the best scoring 30° Δ pose was selected. Ranks of these poses were compared. Results for tests (iii) can also be found in Table 2, where 5 ligands had to be discarded from the analyses as RDKit failed to successfully provide 10 conformers. The proposed model is able to recover the crystal ligand out of the pool of generated conformers in 74 and 76 out of 80 cases for top 1 and 5 ranks, respectively. When considering only 10 generated poses, however, we did not observe successful pose prediction. Only in two cases the method was able to generate poses with an RMSD <2. This low performance could be attributed to non-exhaustive conformer generation, large Δ angle or translation.

5 Implementation and availability

An implementation of LigVoxel can be further explored by the scientific community through a web-application, part of the playmolecule.org suite. An example sesssion is shown in Figure 6. Users can submit their own protein binding site in.pdb format, choose either the conditional or unconditional model for field predictions and retrieve results, that can either be visualized using NGL (Rose and Hildebrand, 2015) or downloaded locally for further analysis.

6 Discussion

In this work, we have presented a novel approach for the generation of ligand images filling protein pockets based on deep neural networks. Results shown here suggest that these predictions are responsive to the number of atoms selected as input, they significantly overlap with ligand features of previously unseen ligands, and they can be used to select poses and conformers close to the native ligand orientation and geometry. We believe that approaches similar to the one proposed here will be important for structure based design of protein binders. For instance, in *de novo* drug-design, these generated fields could be used as reference for incremental atom placement. Furthermore, in computational docking, the proposed fields can assist classical scoring functions by overlap or proximity optimization. The extent to which these models can be applied in drug discovery and design pipelines has yet to be demonstrated fully, but we believe that the initial validation performed in this work is promising for future applications.

Funding

The authors thank Acellera Ltd. for funding. G.D.F. acknowledges support from MINECO (BIO2017-82628-P) and FEDER. This project has received funding from the European Union's Horizon 2020 Research and innovation programme under Grant Agreement No. 675451 (CompBioMed project).

Conflict of Interest: none declared.

References

Angermueller,C. *et al.* (2016) Deep learning for computational biology. *Mol. Syst. Biol.*, **12**, 878–816.

Berman,H.M. *et al.* (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.

Carosati,E. *et al.* (2004) Hydrogen bonding interactions of covalently bonded fluorine atoms: from crystallographic data to a new angular function in the GRID force field. *J. Med. Chem.*, **47**, 5114–5125.

Chollet,F. (2015) Keras. https://github.com/fchollet/keras.

Cross,S. *et al.* (2012) GRID-based three-dimensional pharmacophores I: fLAPpharm, a novel approach for pharmacophore elucidation. *J. Chem. Inf. Model.*, **52**, 2587–2598.

Davis,I.W. and Baker,D. (2009) Rosettaligand docking with full ligand and receptor flexibility. *J. Mol. Biol.*, **385**, 381–392.

De Boer,P.T. *et al.* (2005) A tutorial on the cross-entropy method. *Ann. Oper. Res.*, **134**, 19–67.

Derevyanko,G. *et al.* (2018) Deep convolutional networks for quality assessment of protein folds. *ArXiv:1801.06252v1*.

Desaphy,J. *et al.* (2012) Comparison and druggability prediction of protein–ligand binding sites from pharmacophore-annotated cavity shapes. *J. Chem. Inf. Model.*, **52**, 2287–2299.

Desaphy,J. *et al.* (2015) Sc-PDB: a 3D-database of ligandable binding sites-10 years on. *Nucleic Acids Res.*, **43**, D399–D404.

Doerr,S. *et al.* (2016) HTMD: high-throughput molecular dynamics for molecular discovery. *J. Chem. Theory Comput.*, **12**, 1845–1852.

Ebalunode,J.O. *et al.* (2008) Novel approach to structure-based pharmacophore search using computational geometry and shape matching techniques. *J. Chem. Inf. Model.*, **48**, 889–901.

Ghersi,D. and Sanchez,R. *et al.* (2009) EasyMIFs and SiteHound: a toolkit for the identification of ligand-binding sites in protein structures. *Bioinformatics*, **25**, 3185–3186.

Goldberg,Y. (2015) A primer on neural network models for natural language processing. *arXiv Preprint arXiv: 1510.00726*, pp. 1–76.

Goodfellow. *et al.* (2016) *Deep learning*. MIT Press, Cambridge, MA. http://www.deeplearningbook.org.

Graves,A. *et al.* (2013) Speech recognition with deep recurrent neural networks. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 6645–6649.

Hartshorn,M.J. *et al.* (2007) Diverse, high-quality test set for the validation of protein–ligand docking performance. *J. Med. Chem.*, **50**, 726–741.

Hawkins,P.C. *et al.* (2007) Comparison of shape-matching and docking as virtual screening tools. *J. Med. Chem.*, **50**, 74–82.

Hu,B., and Lill,M.A. (2014) PharmDock: a pharmacophore-based docking program. *J. Cheminform.*, **6**, 14.

Jiménez,J. *et al.* (2017) DeepSite: protein-binding site predictor using 3D-convolutional neural networks. *Bioinformatics*, **33**, 3036–3042.

Jiménez,J. *et al.* (2018) Kdeep: protein–ligand absolute binding affinity prediction via 3d-convolutional neural networks. *J. Chem. Inf. Model.*, **58**, 287–296.

Kingma,D. and Ba,J. (2014) Adam: a method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, pp. 1–13.

Koes,D.R. and Camacho,C.J. (2014) Shape-based virtual screening with volumetric aligned molecular shapes. *J. Comput. Chem.*, **35**, 1824–1834.

Krizhevsky,A. *et al.* (2012) ImageNet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.*, **25**, 1097–1105.

Kullback,S. and Leibler,R.A. (1951) On information and sufficiency. *Ann. Math. Statist.*, **22**, 79–86.

Landrum,G. (2012) RDKit: Open-source cheminformatics. http://www.rdkit.org. [Online; accessed December-2017].

LeCun,Y. *et al*. (1998) Gradient-based learning applied to document recognition. *Proc. IEEE*, **86**, 2278–2323.

Morris,G.M. *et al*. (2009) AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. *J. Comput. Chem.*, **30**, 2785–2791.

Ragoza,M. *et al*. (2017) Protein–ligand scoring with convolutional neural networks. *J. Chem. Inf. Model.*, **57**, 942–957.

Ravindranath,P.A. *et al*. (2015) AutoDockFR: advances in protein–ligand docking with explicitly specified binding site flexibility. *PLoS Comput. Biol.*, **11**, e1004586.

Rose,A.S. and Hildebrand,P.W. (2015) NGL viewer: a web application for molecular visualization. *Nucleic Acids Res.*, **43**, W576–W579.

Theano Development Team (2016) Theano: a Python framework for fast computation of mathematical expressions. *arXiv e-Prints*, *arXiv:1605.02688v1*.

Yuan,Y. *et al*. (2011) LigBuilder 2: a practical de novo drug design approach. *J. Chem. Inf. Model.*, **51**, 1083–1091.

Supplementary Information

LigVoxel: Inpainting binding pockets using 3D-convolutional neural networks.

Miha Skalic[1], Alejandro Varela-Rial[2], José Jiménez[1], Gerard Martinez-Rosell[1], and Gianni de Fabritiis[1,3,*]

[1]Computational Biophysics Laboratory, Universitat Pompeu Fabra, Parc de Recerca Biomèdica de Barcelona, Carrer del Dr. Aiguader 88. Barcelona 08003, Spain.
[2]Acellera, Barcelona Biomedical Research Park (PRBB), Doctor Aiguader 88, 08003 Barcelona, Spain.
[3]Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig Lluis Companys 23, 08010 Barcelona, Spain

* Email: gianni.defabritiis@upf.edu

Contents

List of Figures

Voxelization algorithm

The voxelization algorithm is the same as described in *Jiménez et. al. 2017* and it is implemented in HTMD sotware (htmd.org).

Algorithm 1 Descriptor pseudo-code computation.

function OCCUPANCY(atomCoords, centerCoords, radii, channels)
  for each atom A in compound **do**
    $a \leftarrow \text{atomCoords}_A$
    $h \leftarrow \text{channels}_A$
    $r_{\text{vdw}} \leftarrow \text{radii}_A$
    for each center c in centerCoords **do**
      $r \leftarrow L_2\text{Dist}(c, a)$
      $x \leftarrow r_{\text{vdw}} r$
      $n \leftarrow 1 - \exp(-x^{12})$
      for each channel p in h **do**
        $O_{c,p} \leftarrow \max\{n, O_{c,p}\}$
      end for
    end for
  end for
end function

Neural network implementation

Listing 1: Python implementation of neural network.

```
# theano==0.9.0
# keras==1.2.2.

from keras.layers import Conv3D
from keras.models import Sequential

def retrieve_model():
    model = Sequential()
    model.add(Conv3D(32, (3, 3, 3), activation='relu', padding='same',
                     input_shape=(12, 16, 16, 16), name='protein'))
    model.add(Conv3D(64, (3, 3, 3), activation='relu', padding='same'))
    model.add(Conv3D(80, (3, 3, 3), activation='relu', padding='same'))
    model.add(Conv3D(20, (3, 3, 3), activation='relu', padding='same'))
    model.add(Conv3D(20, (16, 16, 16), activation='relu', padding='same'))
    model.add(Conv3D(16, (3, 3, 3), activation='relu', padding='same'))
    model.add(Conv3D(4, (3, 3, 3), activation='sigmoid', padding='same',
                     name='gen_lig'))
    return model
```

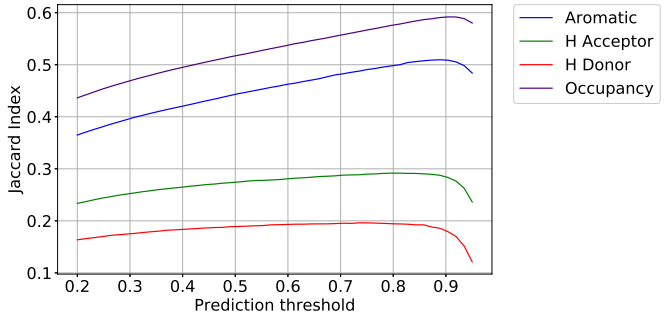


Figure S1: Median sample Jaccard index as a function of threshold cut-off. Ground truth values are at thresholded at 0.75

Complexes in each split

Supporting file `splits.txt` list PDB IDs of complexes used in the scPDB database as test in each split. Astex lists PDB ids that were not used during training in tests (ii-iv). PDB complex 1GRB was excluded due to .pdbqt conversion failure.

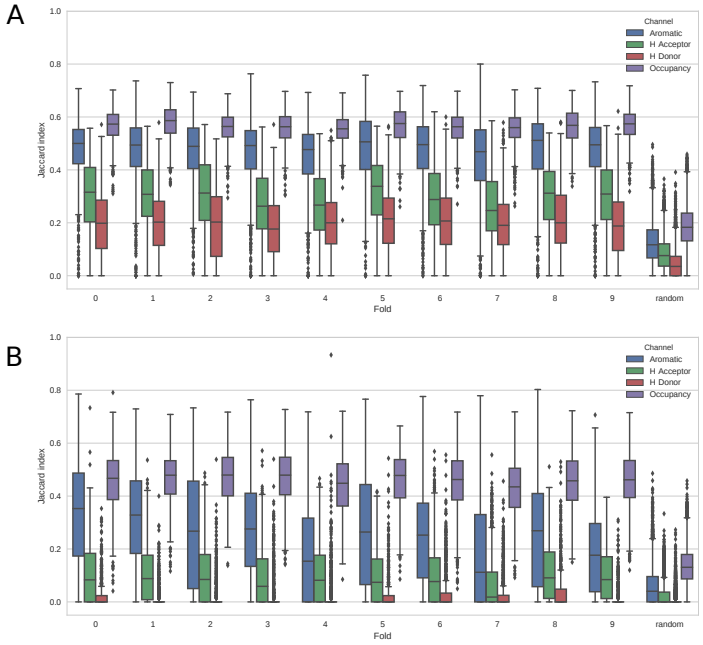


Figure S2: Per split 10-fold cross-validation results of predicted and actual physico-chemical voxel values in the scPDB database, for channels aromatic, hydrogen-bond acceptor, hydrogen-bond donor and total excluded volume, using a Jaccard index thresholded at 0.75. Results for the the (A) conditional and (B) unconditional model respectively. Random baseline includes samples from all folds.

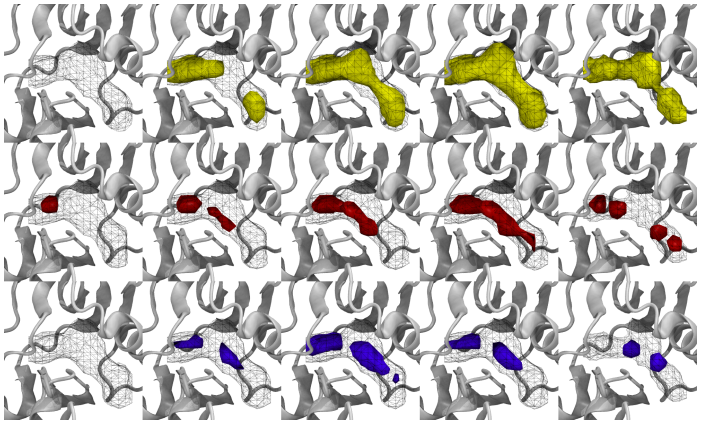


Figure S3: Variance in generated properties by changing input property count for PDB entry 1FPU. Ligand occupancy is displayed in black wireframe while aromatics, H-bond acceptors and donors are in yellow, red and violet, respectively. Each row shows 3 variational predictions, 4th prediction is generated using property count of crystal ligand followed by the crystal ligand itself. Other, not displayed, input properties are kept the same as in crystal structure. Displayed are predictions using property counts Aromatic: 0, 6, 12, 20; H-bond acceptor: 0, 1, 3, 5; H-bond donor: 0, 1, 5, 2.

3.2 SkeleDock: A Web Application for Scaffold Docking in PlayMolecule

Varela-Rial A., Majewski M., Cuzzolin A., Martínez-Rosell G., De Fabritiis G. SkeleDock: A Web Application for Scaffold Docking in PlayMolecule. Journal of Chemical Information and Modeling. 2020; 60(6):2673-2677.

Summary

In this article we described SkeleDock, a new scaffold docking algorithm that allows the modeling of the binding mode of a query ligand using the structure of a similar protein-ligand complex as a template. Besides describing the algorithm itself, we also present the results obtained in the 4th edition of the D3R Grand Challenge.

JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Application Note

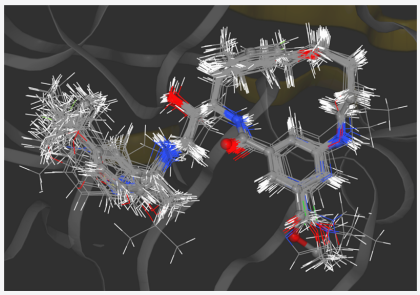
SkeleDock: A Web Application for Scaffold Docking in PlayMolecule

Alejandro Varela-Rial, Maciej Majewski, Alberto Cuzzolin, Gerard Martínez-Rosell, and Gianni De Fabritiis*

Cite This: J. Chem. Inf. Model. 2020, 60, 2673−2677 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: SkeleDock is a scaffold docking algorithm which uses the structure of a protein−ligand complex as a template to model the binding mode of a chemically similar system. This algorithm was evaluated in the D3R Grand Challenge 4 pose prediction challenge, where it achieved competitive performance. Furthermore, we show that if crystallized fragments of the target ligand are available then SkeleDock can outperform rDock docking software at predicting the binding mode. This Application Note also addresses the capacity of this algorithm to model macrocycles and deal with scaffold hopping. SkeleDock can be accessed at https://playmolecule.org/SkeleDock/.



INTRODUCTION

Predicting the binding mode of small molecules in a protein pocket is one of the main challenges in the field of computational chemistry. Accurate predictions can substantially reduce the costs of drug development and speed up the process.[1] Several software solutions exist that address this problem, including AutoDock Vina,[2] Glide,[3] Gold,[4] or rDock.[5] Typical docking protocols use the protein cavity and the query ligand to generate poses that are later evaluated with a generalized scoring function. However, structural knowledge about the target system is usually available, such as protein homologues with similar cocrystallized ligands. Hence, given that the binding mode of similar molecules is usually conserved,[6,7] it is reasonable to exploit this information to increase the accuracy of the prediction. Considering the growing amount of structural data available in the Protein Data Bank (PDB)[8] and the popularity of fragment-based drug discovery,[9] we expect this knowledge-rich scenario to become increasingly prevalent.

Docking algorithms which make use of such knowledge are usually referred to as similarity docking or scaffold docking.[10] Scaffold docking methods usually rely on maximum common substructure (MCS) approaches, such as fkcombu.[11] MCS methods try to find the largest common substructure (subgraph) between two molecules. When found, the conformation of that substructure in the query ligand can be modeled by simply mimicking the conformation of that same substructure in the template, while the position of the remaining atoms is decided by a general scoring function. However, due to the characteristics of the MCS methods, two almost identical molecules that only differ in minor modification can return disappointingly short subgraphs. A shorter MCS means that more atoms in the query ligand would have to be modeled without any reference by the docking software, which is not desirable. Additionally, such minimal mismatches can be of critical interest in medicinal chemistry, as they can constitute scaffold hops that can, potentially, improve the pharmacological properties of a compound or circumvent intellectual property.[12] Therefore, there is a need to maximize the use of structural information. We present here SkeleDock, a new scaffold docking algorithm that can overcome local mismatches.

FEATURES

Algorithm. The SkeleDock web application provides a user-friendly interface to perform scaffold docking, starting from files with the structure of the receptor (PDB), a template molecule (PDB), and a set SMILES representing query ligands (CSV). After submission, these files follow SkeleDock's algorithm, whose main steps are summarized in Figure 1. The algorithm begins by building a graph for the query and the template molecules. These two graphs are then compared to identify a common subgraph, that is, a continuous set of atoms whose element (node) and bonds (edges) are equivalent in both the query and the template molecules. Hence, if this step is successful, a mapping linking several atoms in the query

Received: February 10, 2020
Published: May 14, 2020

https://dx.doi.org/10.1021/acs.jcim.0c00143
J. Chem. Inf. Model. 2020, 60, 2673−2677

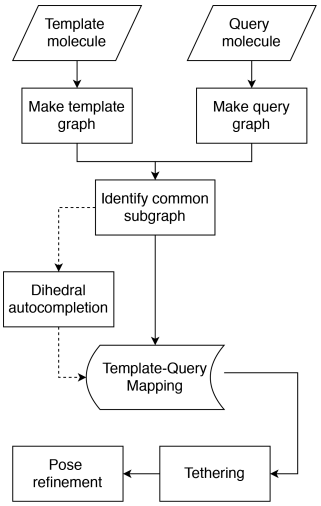


Figure 1. Main steps of SkeleDock algorithm. The dihedral autocompletion step is optional.

molecule to their template counterparts will be returned. In the following step, tethering, this mapping is used to change the position of the atoms of the query molecule. This is done by creating a force in each query atom that points toward the location of its template counterpart, effectively biasing the conformation of the query ligand toward that of the template. Finally, in order to find an appropriate location for those atoms in the query molecule for which no template equivalent was found, the *tethered template docking* protocol of rDock[5] is used. This protocol allows the user to constrain the degrees of freedom of the docking run (orientation, position, and dihedral angles of the ligand) based on the initial conformation of the provided molecule and a set of atom indexes. These indexes correspond to the atoms that the user wants to be fixed, in our case, those atoms for which we have found a template counterpart. If a given dihedral is composed by atoms whose indexes belong to this set, its dihedral angle will not be sampled at all or only within a user-defined range.

Autocompletion Step. As previously discussed, one limitation of methods based on MCS is its sensitivity to small changes: two molecules which are almost identical, except for some minor modifications, will return a smaller mapping, as the common subgraph shared by both is now smaller. Figure 2a depicts such a scenario. To avoid this problem, we added an optional step called dihedral autocompletion. As shown in Figure 2b, the mapping found in the graph comparison step has *stopped* just before the atom whose element differs between the query and template molecules, depicted as X in Figure 2a. However, this mismatching atom belongs to a dihedral (highlighted in a ball−stick representation) in which three consecutive atoms are already mapped to the template. We can then assume that the mismatching atom—the fourth atom of this dihedral—matches the fourth atom of the equivalent dihedral in the template. This is what we refer to as dihedral autocompletion. After each dihedral autocompletion cycle, a new nonmapped fourth atom appears, and this step is repeated recursively until no more atoms are available. If the template dihedral offers

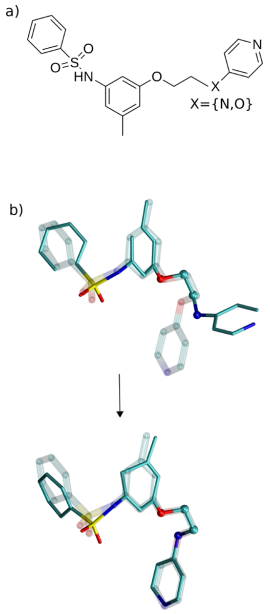


Figure 2. Dihedral autocompletion step. (a) Chemical structure of template and query molecules. The mismatching atom is depicted as X. (b) Overlap between query molecule (opaque licorice) and template molecule (transparent texture) before (top image) and after (bottom image) the autocompletion step. The semicompleted dihedral (atoms depicted with a ball) propagates to the right side, improving the overlap with the template. Template molecule is PDB code, 1UVT; resname, I48. Rings have to be broken to allow this step, but they are restored before the tethering. These conformations are not the final docked poses.

several possibilities for the fourth atom, all of them are explored and evaluated. This functionality is key to overcome local mismatches, which makes SkeleDock able to handle some minor scaffold hops. Some MCS methods can overcome simple mismatches as the one shown in Figure 2, as they could identify the two disconnected, common subgraphs. However, if these disconnected subgraphs are not highly similar, typical MCS methods could fail.

Application Options. Different options are available to change the behavior of the application. The *rDock refinement step* is enabled by default, but it might not be necessary if every query atom has a template equivalent. The *scaffold-hopping tolerant mode* enables or disables the dihedral autocompletion step. Users can also modify the magnitude of the force applied to each atom during tethering. Higher values result in a better alignment but might introduce some artifacts, like a change of chirality. The last option is *probe radius* that defines the radius of the spheres used to define the size of the docking cavity for rDock.[5] After execution, the best pose of each ligand can be displayed together with the protein and the template ligand (Figure 3). Results can be downloaded in a tar.gz file.

Time Performance. We assessed the efficiency of SkeleDock by docking congeneric series for two different targets: Cathepsin S (459 ligands, average of 46.6 heavy atoms) and BACE-1 (154 ligands, average of 38.4 heavy atoms). We used rDock as a baseline, and each test was run

https://dx.doi.org/10.1021/acs.jcim.0c00143
J. Chem. Inf. Model. 2020, 60, 2673−2677

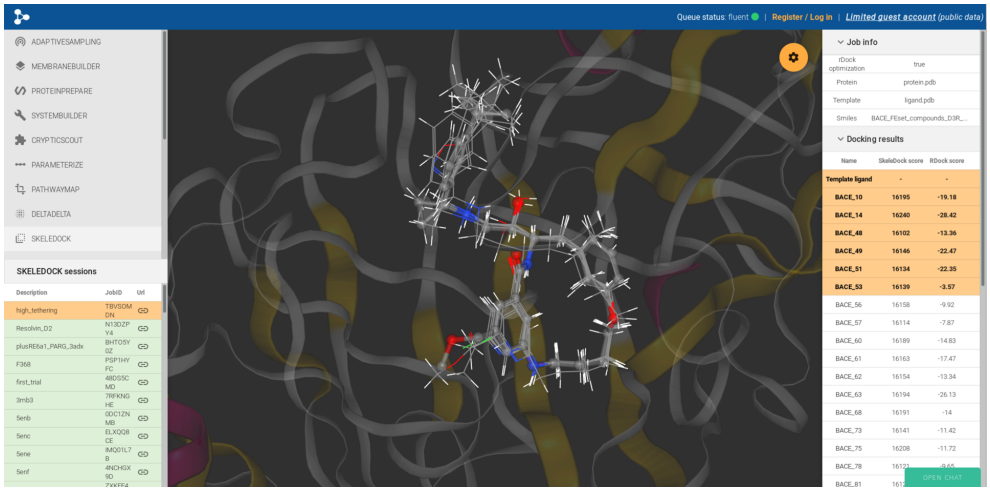


Figure 3. SkeleDock's graphical user interface. Docked molecules (line representation) are shown overlapped with the template used (ball and stick representation).

using 4 and 30 cores. SkeleDock is two to three times slower than rDock, but we believe that the increase in accuracy compensates for it. Table 1 sums the results of the time performance evaluation.

Table 1. Time Performance of SkeleDock and rDock for BACE-1 and CathepsinS Congeneric Series[a]

		BACE-1 (154)	CatS (459)
Method	#CPU	Yield [Lig/min]	Yield [Lig/min]
SkeleDock	4	15.2	13.9
	30	43.8	50.7
rDock	4	50.5	53.4
	30	102.0	127.5

[a]Number of simulated ligands is listed in parentheses.

VALIDATION

Fragment-Based Docking. We evaluated SkeleDock's ability to recover the native pose of a ligand using a fragment as a template. Due to the lack of crystal structures of protein with ligands and corresponding fragments, we decided to artificially generate fragments for complexes from the refined set of PDBbind (version 2018)[13] and use them as templates for SkeleDock. The ligands were fragmented by breaking a selected rotatable bond. We prepared three sets of fragments of increasing difficulty by excluding from the fragment 1, 3, or 5 rotatable bonds of the complete ligand. Deleting more atoms from the template increases the difficulty of predicting the right pose, as there is no reference for them.

We compared SkeleDock's performance with two MCS-based methods and with an unconstrained docking protocol. The MCS-based methods are two different settings of RDKit's[14] findMCS function: The first, where the element of the atoms must match (strict MCS), and the second, where the element and bond-order mismatches are allowed (agnostic MCS). The function returns a mapping (just as the graph comparison step of SkeleDock), which is then directly passed as an input to the tethering and pose refinement steps. Finally, for the unconstrained docking protocol, we used rDock with free rotation, translation, and dihedral angle exploration (free rDock). The performance of docking algorithms is evaluated by the number of correct predictions. By convention, poses are considered correct if their RMSD from their crystal pose is under 2.0 Å.[15] We report two levels of success: Top 1, where only the top pose was selected, and Top 5, where the best among the top five poses was selected (Figure 4). A full report of the docking results can be found in Table S1.

In terms of success rate, SkeleDock outperforms other approaches in all fragmentation scenarios (Figure 4). Strict

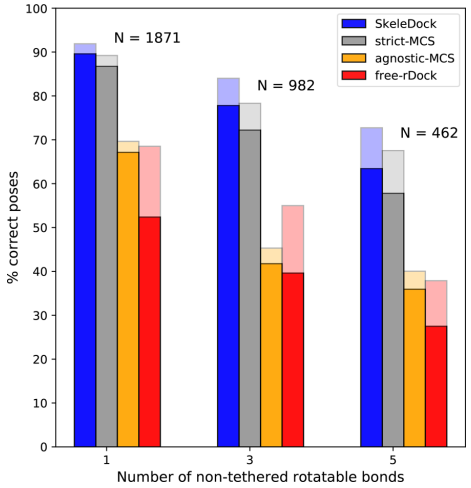


Figure 4. Self-docking performance of SkeleDock, strict MCS, agnostic MCS, and free rDock. Different shades correspond to the different success levels: Top 1, opaque; Top 5, transparent.

https://dx.doi.org/10.1021/acs.jcim.0c00143
J. Chem. Inf. Model. 2020, 60, 2673−2677

MCS is comparable to SkeleDock, and they both outperform agnostic MCS and free rDock. This result is expected because the less-strict nature of the agnostic MCS setting might find mappings which are feasible in terms of equivalence in other features (as ring−ring) but lead to wrong orientations of the ligand. These results suggest that when the binding mode of the query ligand and its fragments are conserved then biasing the prediction using SkeleDock or MCS approaches can substantially increase the success rate of binding mode prediction.

D3R Grand Challenge 4. In order to evaluate SkeleDock prospectively, we engaged in the D3R Grand Challenge 4 (GC4) pose prediction subchallenge. The D3R Grand Challenge is an international contest where participants complete different computational tasks of pharmaceutical interest.[16] In its fourth edition, the objective was to predict binding modes of 20 ligands of BACE-1 protein. As templates for SkeleDock, we used crystal structures of close homologues and their cocrystallized ligands from PDB (Table S2). At the time the challenge took place, the final rDock pose refinement step was not implemented in the protocol. Instead, we run a short molecular dynamics (MD) simulation to relax the poses (see MD Simulation in the SI for further details). To asses the performance of the final protocol, a retrospective analysis was run using SkeleDock's web application at PlayMolecule.

This subchallenge was particularly complicated for two reasons: (1) All ligands except one had a macrocycle. (2) Most of the ligands had a shortened MCS with their template due to certain atoms differing in elements or the presence of rings. Conformational changes in macrocycles involve the concerted rotation of several dihedrals, making them difficult to model.[17] The gold standard among docking practitioners is to first sample different conformations of a macrocycle and then dock each one independently. This was not needed in our case, as SkeleDock can simply use the macrocycle of the template to model the one in the query ligand. Regarding the shortened MCS, the autocompletion step of SkeleDock can handle these mismatches, leading to a greater mapping and overlap with the templates both in the macro and nonmacro fractions of the molecules, as can be seen in Figure 5. We actually compared the RMSD of the poses generated by SkeleDock and the two MCS methods described in *fragment-based docking*. Both the global RMSD and the macrocycle RMSD is lower in SkeleDock poses thanks to the bigger mapping with the template (Tables S3 and S4).

SkeleDock's submission (code: qqou3) finished among the top-performing participants, ranking ninth out of 74 according to median RMSD (1.02 Å) and 15th according to mean RMSD (1.33 Å). In the retrospective analysis, the SkeleDock web application performed slightly worse with a mean RMSD of 1.47 Å. Given that this test was run in a fully automated fashion and with no human supervision, the gap between the two results is understandable.

CONCLUSIONS

SkeleDock algorithm offers four main features: (1) docking of molecules based on their analogues or fragments, (2) an autocompletion step that can handle local mismatches and, hence, model minor scaffold hops, (3) the ability to model macrocycles without having to pregenerate ring conformations, and (4) a user-friendly GUI that enables efficient scaffold docking and results exploration. The protocol can be accessed at https://playmolecule.org/SkeleDock/.

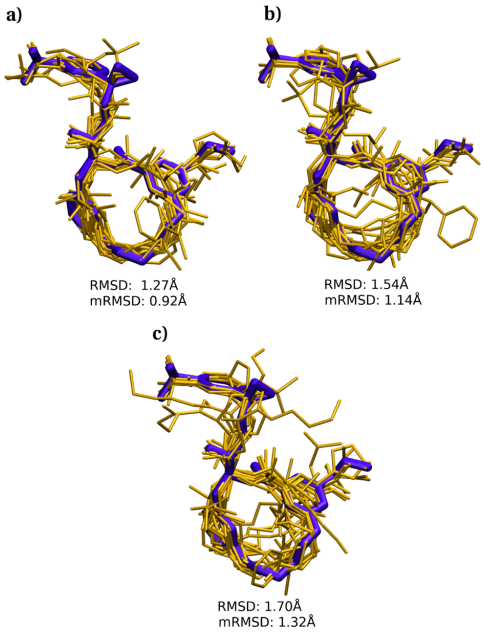


Figure 5. Overlap between predicted poses (gold) and template (violet) using three different methods: (a) SkeleDock, (b) Element Agnostic, and (c) Strict MCS. RMSD is the average RMSD value of the poses, and mRMSD is the mean value of the RMSD of the macrocycle atoms.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.0c00143.

Table S1: Detailed results of the retrospective validation analysis. (XLSX)

Table S2: PDB codes used as template for SkeleDock in the D3R GC4 pose prediction challenge. MD simulation: Description of the MD protocol used to refine the poses in D3R GC4 pose prediction challenge. Table S3: Mean RMSD obtained by the different methods used to model macrocycles. Table S4: Mean RMSD (computed for macrocycle atoms only) obtained by the different methods used to model macrocycles. (PDF)

AUTHOR INFORMATION

Corresponding Author

Gianni De Fabritiis − *Computational Science Laboratory, Universitat Pompeu Fabra, Barcelona, Spain; Acellera Labs, Barcelona, Spain; Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain;* orcid.org/0000-0003-3913-4877; Email: gianni.defabritiis@upf.edu

Authors

Alejandro Varela-Rial − *Acellera Labs, Barcelona, Spain; Computational Science Laboratory, Universitat Pompeu Fabra, Barcelona, Spain;* orcid.org/0000-0002-6918-1765

https://dx.doi.org/10.1021/acs.jcim.0c00143
J. Chem. Inf. Model. 2020, 60, 2673−2677

Maciej Majewski − *Computational Science Laboratory, Universitat Pompeu Fabra, Barcelona, Spain;* orcid.org/0000-0003-2605-8166
Alberto Cuzzolin − *Acellera Labs, Barcelona, Spain*
Gerard Martínez-Rosell − *Acellera Labs, Barcelona, Spain;* orcid.org/0000-0001-6277-6769

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.0c00143

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank Acellera Ltd. for funding and the D3R organizers for their efforts. G.D.F. acknowledges support from MINECO (BIO2017-82628-P) and FEDER. This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 823712 (CompBioMed2) and from the Industrial Doctorates Plan of the Secretariat of Universities and Research of the Department of Economy and Knowledge of the Generalitat of Catalonia.

■ REFERENCES

(1) Pinzi, L.; Rastelli, G. Molecular Docking: Shifting Paradigms in Drug Discovery. *Int. J. Mol. Sci.* **2019**, *20*, 4331.
(2) Trott, O.; Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **2010**, *31*, 455−61.
(3) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **2004**, *47*, 1739−1749.
(4) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking 1 1Edited by F. E. Cohen. *J. Mol. Biol.* **1997**, *267*, 727−748.
(5) Ruiz-Carmona, S.; Alvarez-Garcia, D.; Foloppe, N.; Garmendia-Doval, A. B.; Juhos, S.; Schmidtke, P.; Barril, X.; Hubbard, R. E.; Morley, S. D. rDock: A Fast, Versatile and Open Source Program for Docking Ligands to Proteins and Nucleic Acids. *PLoS Comput. Biol.* **2014**, *10*, No. e1003571.
(6) Drwal, M. N.; Jacquemard, C.; Perez, C.; Desaphy, J.; Kellenberger, E. Do Fragments and Crystallization Additives Bind Similarly to Drug-like Ligands? *J. Chem. Inf. Model.* **2017**, *57*, 1197−1209.
(7) Malhotra, S.; Karanicolas, J. When Does Chemical Elaboration Induce a Ligand To Change Its Binding Mode? *J. Med. Chem.* **2017**, *60*, 128−145.
(8) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(9) Lamoree, B.; Hubbard, R. E. Current perspectives in fragment-based lead discovery (FBLD). *Essays Biochem.* **2017**, *61*, 453−464.
(10) Fradera, X.; Mestres, J. Guided Docking Approaches to Structure-Based Design and Screening. *Curr. Top. Med. Chem. (Sharjah, United Arab Emirates)* **2004**, *4*, 687−700.
(11) Kawabata, T.; Nakamura, H. 3D Flexible Alignment Using 2D Maximum Common Substructure: Dependence of Prediction Accuracy on Target-Reference Chemical Similarity. *J. Chem. Inf. Model.* **2014**, *54*, 1850−1863.
(12) Hu, Y.; Stumpfe, D.; Bajorath, J. Recent Advances in Scaffold Hopping. *J. Med. Chem.* **2017**, *60*, 1238−1246.
(13) Wang, R.; Fang, X.; Lu, Y.; Yang, C.-Y.; Wang, S. The PDBbind Database: Methodologies and Updates. *J. Med. Chem.* **2005**, *48*, 4111−4119.
(14) Landrum, G. *RDKit: A Software Suite for Cheminformatics, Computational Chemistry, and Predictive Modeling.* http://www.rdkit.org (accessed 1 April 2020).
(15) Cross, J. B.; Thompson, D. C.; Rai, B. K.; Baber, J. C.; Fan, K. Y.; Hu, Y.; Humblet, C. Comparison of Several Molecular Docking Programs: Pose Prediction and Virtual Screening Accuracy. *J. Chem. Inf. Model.* **2009**, *49*, 1455−1474.
(16) Gaieb, Z.; Parks, C. D.; Chiu, M.; Yang, H.; Shao, C.; Walters, W. P.; Lambert, M. H.; Nevins, N.; Bembenek, S. D.; Ameriks, M. K.; Mirzadegan, T.; Burley, S. K.; Amaro, R. E.; Gilson, M. K. D3R Grand Challenge 3: blind prediction of protein-ligand poses and affinity rankings. *J. Comput.-Aided Mol. Des.* **2019**, *33*, 1−18.
(17) Allen, S. E.; Dokholyan, N. V.; Bowers, A. A. Dynamic Docking of Conformationally Constrained Macrocycles: Methods and Applications. *ACS Chem. Biol.* **2016**, *11*, 10−24.

https://dx.doi.org/10.1021/acs.jcim.0c00143
J. Chem. Inf. Model. 2020, 60, 2673−2677

Supplementary Information SkeleDock: A Web Application for Scaffold Docking in PlayMolecule

Alejandro Varela-Rial,[†,‡] Maciej Majewski,[‡] Alberto Cuzzolin,[†] Gerard Martínez-Rosell,[†] and Gianni De Fabritiis[*,‡,†,¶]

†*Acellera Labs, Doctor Trueta 183, Barcelona, Spain*
‡*Computational Science Laboratory, Universitat Pompeu Fabra, Barcelona Biomedical Research Park (PRBB), Barcelona, Spain*
¶*Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig Lluis Companys 23, Barcelona, Spain*

E-mail: gianni.defabritiis@upf.edu

Table S2: PDB codes used as template for SkeleDock in the D3R GC4 pose prediction challenge

PDB code used as template	D3RGC4 ligand identifier
3DV1	BACE19
3DV5	BACE2, BACE3, BACE4, BACE5
3K5C	BACE7, BACE8, BACE9, BACE10, BACE11, BACE12, BACE13, BACE14, BACE15, BACE16, BACE17
4DPF	BACE1
4DPI	BACE6
4GMI	BACE18
5MCO	BACE20

MD simulation. In the pose prediction subchallenge, SkeleDock's predicted poses underwent a short MD equilibration protocol to relax the system. MD simulations were prepared as follows: charges for the 20 molecules were set manually, then the Parameterize tool[1] was used to get the force field parameters of these molecules. Systems were build using AMBER force field[2] and prepared with HTMD[3] proteinPrepare protocol[4] at pH 5.1, as it was the pH at which the PDB structure 3DV1 (mentioned in the challenge description) was crystallized. A 10 Å padding was added to the system to then solvate using explicit TIP3 water. Heavy constraints were imposed in the heavy atoms of the backbone of the protein, softer constraints were applied to atoms in the side-chains and atoms in the ligand were left free to adopt a lower energy conformation. Under these constraints, 3 ns of equilibration at 300K with timestep 2 fs were run after 1000 steps of minimization with ACEMD.[5] Among the frames generated during the simulation, the first half was discarded, and from the second half, the frame whose conformation was closest to the starting system (SkeleDock's result) was selected as the submitted pose. The rationale behind this approach is that the poses will benefit from the mapping found by SkeleDock, while the most unfavourable interactions will be solved. No pose deviated more than 1.5Å with respect to the starting point.

Table S3: Mean RMSD obtained by the different methods used to model macrocycles. Units are in Angstroms. The error of SkeleDock is the standard deviation of the 5 replicas run. The other methods are deterministic.

BACE ID	**SkeleDock**	**Element-Agnostic**	**Strict-MCS**
BACE1	**1.89** ± 0.01	2.03	3.84
BACE2	**1.27** ± 0.08	5.09	1.64
BACE3	**1.42** ± 0.01	1.5	1.67
BACE4	**1.12** ± 0.03	1.31	2.27
BACE5	**1.5** ± 0.02	2.05	2.88
BACE6	**1.0** ± 0.07	1.34	2.15
BACE7	**1.52** ± 0.04	4.26	1.57
BACE8	0.88 ± 0.01	**0.78**	0.83
BACE9	**0.65** ± 0.02	0.7	0.67
BACE10	0.78 ± 0.02	0.79	**0.77**
BACE11	**0.81** ± 0.01	0.86	0.93
BACE12	0.76 ± 0.01	**0.64**	1.02
BACE13	1.81 ± 1.13	**1.0**	1.43
BACE14	0.69 ± 0.01	**0.64**	0.74
BACE15	0.73 ± 0.04	**0.71**	0.79
BACE16	**0.94** ± 0.02	0.98	1.89
BACE17	**0.76** ± 0.01	1.12	1.28
BACE18	**1.24** ± 0.01	1.26	1.57
BACE19	2.07 ± 0.01	**1.95**	4.58
BACE20	3.59 ± 0.02	1.82	**1.53**
Mean	**1.27**	1.54	1.70

Table S4: Mean RMSD (computed for macrocycle atoms only) obtained by the different methods used to model macrocycles. Units are in Angstroms. The error of SkeleDock is the standard deviation of the 5 replicas run. The other methods are deterministic.

BACE ID	**SkeleDock**	**Element-Agnostic**	**Strict-MCS**
BACE1	1.63 ± 0.02	**1.15**	2.11
BACE2	**1.15** ± 0.02	4.65	1.63
BACE3	**0.93** ± 0.01	1.35	1.5
BACE4	**1.09** ± 0.01	1.59	2.75
BACE5	**0.89** ± 0.04	0.98	3.59
BACE6	**1.11** ± 0.06	1.13	1.59
BACE7	**1.2** ± 0.06	2.31	1.31
BACE8	0.85 ± 0.01	**0.82**	0.83
BACE9	0.67 ± 0.01	0.68	**0.65**
BACE10	**0.92** ± 0.07	0.95	0.93
BACE11	**0.82** ± 0.01	0.84	0.91
BACE12	0.7 ± 0.01	**0.58**	0.78
BACE13	0.6 ± 0.25	0.46	**0.45**
BACE14	0.61 ± 0.03	**0.47**	0.58
BACE15	0.92 ± 0.01	**0.91**	0.94
BACE16	**0.78** ± 0.01	0.79	0.94
BACE17	0.58 ± 0.01	0.59	**0.55**
BACE18	**0.53** ± 0.01	0.54	1.33
BACE19	1.47 ± 0.01	**0.86**	1.67
BACE20	-	-	-
Mean	**0.92**	1.14	1.32

References

(1) Galvelis, R.; Doerr, S.; Damas, J. M.; Harvey, M. J.; De Fabritiis, G. A Scalable Molecular Force Field Parameterization Method Based on Density Functional Theory and Quantum-Level Machine Learning. *J. Chem. Inf. Model.* **2019**, *59*, 3485–3493.

(2) Salomon-Ferrer, R.; Case, D. A.; Walker, R. C. An overview of the Amber biomolecular simulation package. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2013**, *3*, 198–210.

(3) Doerr, S.; Harvey, M. J.; Noé, F.; De Fabritiis, G. HTMD: High-Throughput Molecular Dynamics for Molecular Discovery. *J. Chem. Theory Comput.* **2016**, *12*, 1845–1852.

(4) Martínez-Rosell, G.; Giorgino, T.; De Fabritiis, G. PlayMolecule ProteinPrepare: A Web Application for Protein Preparation for Molecular Dynamics Simulations. *J. Chem. Inf. Model.* **2017**, *57*, 1511–1516.

(5) Harvey, M. J.; Giupponi, G.; Fabritiis, G. D. ACEMD: Accelerating Biomolecular Dynamics in the Microsecond Time Scale. *J. Chem. Theory Comput.* **2009**, *5*, 1632–1639.

3.3 Structure based virtual screening: Fast and slow

Varela-Rial A., Majewski M., De Fabritiis G. Structure based virtual screening: Fast and slow. Wiley Interdisciplinary Reviews: Computational Molecular Science. 2021;e1544.

Summary

In this review we covered different modalities of structure-based virtual screening, discussing their advantages, limitations and their different speed profiles.

Received: 12 March 2021 | Revised: 20 April 2021 | Accepted: 21 April 2021

DOI: 10.1002/wcms.1544

ADVANCED REVIEW

WIREs COMPUTATIONAL MOLECULAR SCIENCE WILEY

Structure based virtual screening: Fast and slow

Alejandro Varela-Rial[1,2] | **Maciej Majewski**[2] | **Gianni De Fabritiis**[2,3]

[1]Acellera Labs, Barcelona, Spain

[2]Computational Science Laboratory, Universitat Pompeu Fabra, Barcelona Biomedical Research Park (PRBB), Barcelona, Spain

[3]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

Correspondence
Gianni De Fabritiis, Computational Science Laboratory, Universitat Pompeu Fabra, Barcelona Biomedical Research Park (PRBB), Barcelona, Spain.
Email: g.defabritiis@gmail.com

Funding information
Acellera Ltd; FEDER; H2020 Research Infrastructures, Grant/Award Number: 823712; Ministerio de Economía y Competitividad, Grant/Award Number: BIO2017-82628-P; Secretariat of Universities and Research of the Department of Economy and Knowledge of the Generalitat of Catalonia

Edited by: Modesto Orozco, Associate Editor

Abstract

For many decades virtual screening methods have provided a convenient and cost effective in silico solution in the early stages of drug discovery. In particular, molecular docking uses structural information to approximate protein–ligand recognition, providing valuable information for large chemical libraries at a fast pace with multiple success stories to validate the approach. Nevertheless, fast turnaround of results required assumptions and approximations which compromise the accuracy of these algorithms. On the other side of the spectrum, physical-based molecular simulations offer more precise and realistic models of protein–ligand binding at the cost of being slower and requiring more expensive computing infrastructure. Both fast and slow approaches are useful and solve different aspects of the same problem. Here, we aim to review these approaches focusing on their capabilities, context of usage and limitations, presenting multiple examples along the way.

This article is categorized under:

Molecular and Statistical Mechanics > Molecular Mechanics

Software > Molecular Modeling

Structure and Mechanism > Computational Biochemistry and Biophysics

KEYWORDS

molecular docking, molecular dynamics, virtual screening

1 | INTRODUCTION

Since Horvath reported the first virtual screening (VS) campaign in 1997,[1] a lot of progress has been made in the development of new methods[2–6] and their global adoption as fundamental techniques in the field of drug discovery. While high throughput screening requires large budgets, VS is relatively inexpensive, reducing the costs of the early stage drug discovery. This initially enabled small biotech companies and academic groups to enter the field. Since then in silico drug discovery evolved into a multimillion dollar market, making an indisputable impact on the pharmaceutical industry.[7,8] Despite their simplicity and many limitations, VS algorithms have led to multiple success stories such as identifying hit compounds in the low-medium micromolar range[9–11] and recovering true binding modes of pharmaceutically relevant complexes.[12–15]

VS methods can be classified into two groups based on their approach: ligand-based virtual screening (LBVS) and structure-based virtual screening (SBVS). LBVS methods rely on the principle that "similar ligands bind in similar fashion." Examples of this group are molecular fingerprints[5] and shape comparison methods,[6] which are used to evaluate a library of molecules against a reference molecule or a collection of known binders. SBVS methods take the advantage of structural data of the protein and can be further divided into two types which we refer to as fast and slow.

Fast methods, referred as molecular docking, are particularly popular because they can process large chemical libraries in a reasonable time. Some of the most popular docking programs include DOCK,[2] GOLD,[3] Glide,[16] AutoDock Vina,[17] and rDock.[18] Docking programs predict the binding mode of small molecules in the binding site of a protein in terms of its location, orientation, and internal conformation. The protein is considered rigid or semi-rigid, allowing only the movement of selected groups to improve hydrogen bonding.[3,18] Docking software uses a search algorithm to explore different ligand poses which are then evaluated by a scoring function. This function tries to approximate the free energy of binding by incorporating the enthalpic (scoring positively interactions like hydrogen bonds or salt bridges while penalizing clashes) and entropic terms, essentially penalizing highly flexible ligands based on their number of rotatable bonds.[19]

Slow methods, based on molecular dynamics (MD), offer a more realistic and detailed view of the binding event. MD simulates the dynamics of a system through time by means of Newton's equations of motion and a force field, which describes the interactions between the different atoms. In reality, these methods range from slow, microsecond long, all-atom MD simulations[20] through adaptive sampling[21] to faster coarse-grained simulations[22] and biased MD methods.[23] Most of these methods can provide valuable insights, like the binding and unbinding pathways that a ligand follows,[24–26] the high-energy barriers that it crosses,[20] or the discovery of allosteric pockets or cryptic sites.[27,28]

In this review we will explore some of the recent advances in SBVS methods, focusing on their strengths and weaknesses. Table 1 summarizes them. First, we will discuss fast molecular docking approaches and take a closer look at scoring functions and ways to address protein flexibility. The second part will cover slower molecular dynamics approaches, discussing in detail their impact on modeling pocket flexibility and their application in in silico binding assays.

2 | SEMI-RIGID FAST VIRTUAL SCREENING

Several models exist that explain protein–ligand binding: lock–key paradigm, induced-fit, and the more recent conformational selection theory.[42] Despite it being a complex phenomenon, there are very simple rules that this process obeys and that docking can easily model, steric constraints being one of them. The conformation that the ligand adopts inside

TABLE 1 Summary of the main strengths and weaknesses for the two groups of methods (fast and slow) studied in this review, together with proposed solutions to alleviate these problems

Group of methods	Strengths	Weaknesses	Solutions
Semi-rigid fast virtual screening	Fast Accessible Adaptable	Inaccurate scoring	Constraints[16,29]
			Template docking[30]
			Visual inspection[31]
			Pharmacophores[18]
			Protein–ligand fingerprints[32]
			Machine learning[33,34]
			MD based rescoring[35]
			Consensus scoring[36,37]
		Rigid receptor	Soft potentials[38]
			Ensemble docking[39]
			Protein–flexible algorithms[4,40]
Physics-based molecular simulations	Accuracy Modeling of binding pathways	Slow Computationally expensive	Adaptive sampling[21]
			Coarse graining[22]
			Replica exchange[41]
			Biased MD methods[23]

the pocket must not clash with the protein atoms, nor can the ligand atoms clash with each other. This can quickly discard plenty of poses and compounds, and can be easily modeled using the Van der Waals (VdW) radius of the atoms involved and a distance cutoff to detect clashes. The VdW potential alone might already give predictions below 2 Å in terms of root mean square deviation (RMSD) with respect to the crystal structure in some cases.[18] In addition to steric constraints, there are a number of well described interactions which also critically affect protein and ligand binding, like hydrogen bonds, ionic interactions, pi-stacking, or dispersion-type interactions.[19] Adding these interactions to the scoring function provides greater discriminatory ability as a pose or a compound establishing favorable interactions, for example, salt bridge or pi-stacking, can be prioritized over another molecule with none or weaker interactions. This is shown by many state-of-the-art docking tools.[43–45]

These tools have, however, limitations caused by the technical difficulties of modeling chemistry itself and the high-throughput requirements needed to evaluate millions of compounds in a reasonable amount of time. The most remarkable limitations are neglecting protein flexibility,[45,46] the role of structural waters,[47,48] and coordinating metals,[49] as well as poor modeling of the entropic component, largely driven by the loss of ligand flexibility inside the pocket[50] and desolvation effects,[51] among others.[52] The impact of ligand flexibility in the entropic term is usually only addressed by correcting the final score with the number of rotatable bonds,[53] while desolvation effects are accounted for by identifying hydrophobic contacts between protein and ligand that could lead the expulsion of waters from the cavity with the subsequent entropy increase.[54]

Furthermore, the success of docking methods is often tested with self-docking exercises, not in real-life VS campaigns. Self-docking is done by extracting the ligand from the structure of a complex and re-docking it into the binding site. This eliminates the problem of protein flexibility as the protein is already in the bound conformation. The much harder task, known as cross-docking, is when a ligand is docked to an apo protein or a protein bound to a different compound. In this case success rates at predicting the true binding mode are in the range of 20%–30%,[43,44] although Verdonk et al. reports 61%.[45] It must be noted that docking algorithms can usually sample the right pose (GOLD reports 91% in self-docking and 72% in cross-docking[45]), meaning the ranking of generated poses is one of the main problems of molecular docking..[55,56] The large gap between self-docking and cross-docking performance highlights the other great challenge of docking: protein flexibility. These two major limitations will be further discussed now, together with approaches on how to alleviate them. We will also discuss the current high-throughput capabilities of docking and strategies on how to deal with large libraries.

2.1 | Scoring functions

The ranking problem, that is, prioritizing the right pose over all the other sampled binding modes, showcases the limits of scoring functions. Scoring functions are responsible for evaluating the different binding modes proposed during docking by estimating their affinity to the target. Poor affinity prediction can therefore lead to incorrect discrimination of active compounds against inactive ones in VS campaigns,[19] as well as ineffective identification of the true binding mode among the proposed solutions. All these issues show the need for better scoring functions, that should finely differentiate the strengths of different hydrogen bond donors and acceptors, be robust enough to not be misled by small changes in distances between the two parties, and accurately account for desolvation effects and water-mediated interactions.[19,31,57]

New protocols have been developed to address such limitations. Typically, a docking campaign involves a preparation step and a post-docking analysis step. In these steps, practitioners can use their prior knowledge of the system to modify the parameters of the docking software or its results. We will now discuss some of these methods, the rationale behind them and how they can impact the results.

2.1.1 | Adding constraints

Prior to docking, a number of techniques can be used to influence the outcome of the docking campaign. One of the simplest is manipulating the partial charges of the atoms in the protein. This can help in prioritizing some interactions in the complex over others, acting effectively as a soft constraint. For instance, an interaction with atoms in the backbone of the protein might be desirable as it can decrease drug resistance.[58] In other cases, it might be known that establishing a hydrogen bond with a particular residue is key for a molecule to have an effect on the target. In the

campaign reported in Ref. 13, the partial charges of some residues were manually increased to improve the polar complementary between the ligands and the protein. Out of the total 66,661 compounds docked using DOCK,[2] the first 500 compounds (as ranked by the scoring function) included all the positive controls added. Visual inspection of these 500 compounds was performed and 16 compounds were selected based on complementary with the binding site and chemical diversity. One of the compounds (compound 3), ranked 97th, inhibited the target (β-lactamase) with a IC50 value of 140 μM and its predicted pose resembled closely the crystal structure (RMSD 0.9 Å). The study by Chen et al.[59] offers another successful example, where the partial charges of some protein atoms were increased. These atoms were known to be involved in hydrogen bonds with different known binders.

Another method to favor a given interaction is the use of explicit constraints,[29] which create an artificial force between two atoms. Such an approach was used by Caporuscio et al.[60] to force an interaction between an aromatic nitrogen of bifonazole and a heme iron in the pocket of cytochrome P450, which stand-alone Glide was failing to capture appropriately. After the addition of the constraint, the binding mode was correctly predicted with an RMSD value of 1.02 Å. Similarly, pharmacophores can be defined by a set of positions in space for which a pharmacophoric feature of a desired type is expected within a tolerance radius.[18] Interestingly, precise knowledge of where to expect each pharmacophore feature can be obtained by relatively fast MD methods like CrypticScout[27] and MixMd.[61] It can also be derived by a human expert or by software tools,[62] including machine learning methods like LigVoxel,[63] and LiGANN.[64] Adding constraints or pharmacophores can compensate inaccuracies in the scoring function, like a poor modeling of a particular interaction (as in the iron example by Caporuscio) and favor a given interaction known to be relevant for the target. However, this requires in the first place a good knowledge of the target which might not be available.

2.1.2 | Template docking

Extreme cases of constrained docking are template docking or scaffold docking.[65] These protocols guide the docking using a reference ligand, for which the binding mode to the target is known. First, the common scaffold between the reference ligand and the query molecule is identified. Then, the atoms of the scaffold in the query ligand are constrained to occupy the same location as their counterparts, determining the position and orientation of the pose. The remaining atoms can be modeled following the usual sampling and scoring steps. Similarly to LBVS methods, these docking protocols rely on the principle that similar ligands bind in a similar fashion, and can drastically improve performance over unconstrained docking.[30] Similarity driven docking greatly limits the number of solutions and reduces the role of the scoring function to a minimum. Although the need for a structure of a close protein–ligand complex reduces the usability of such protocols, the popularity of fragment based screening[66] could make this scenario rather common. In the 4th edition of D3R Grand Challenge,[67] the two targets proposed by the organizers had several homologs with bound ligands whose structures were available in the PDB.[68] For the target BACE-1, in the tasks of predicting the absolute free energy of binding and ranking a series of compounds based on their affinity, the most accurate protocol consisted of two steps, a scaffold docking protocol to generate the poses (SkeleDock)[30] and a convolutional neural network to predict their affinity (KDeep).[33]

2.1.3 | Visual inspection

A common practice in VS is the visual inspection of the top ranked compounds by an experienced researcher, who looks for the interactions that the docking software might have not accurately captured. In particular one assesses the impact of the pocket environment on hydrogen bonds, usually more favorable in hydrophobic pockets, the presence of unmatched hydrogen bond acceptors or donors which can negatively impact the binding affinity due to non-compensated desolvation and ligand and protein strain, among many others.[31] The structure of related protein–ligand complex can also be used to guide the visual inspection.[31] For instance, in the VS screening campaign against a target of tuberculosis (Mtb-ThyX) Luciani et al.[12] docked 2000 compounds with Autodock 4.2.[69] After the visual inspection, 13 compounds were selected based on the similarity of their binding mode with that of known binders, enforcing an interaction with a key arginine and with the cofactor FAD present in the pocket. One of the selected molecules (compound 7) was crystallized to a similar target (Tm-ThyX). The pose predicted by Autodock correctly captured most of the interactions present in the X-ray structure.

2.1.4 | Pharmacophore matching

Visual inspection is, of course, slow and only feasible for a small set of compounds. Some methods can imitate this process, making them faster and more scalable, like evaluating the overlap of the poses with a pharmacophore,[62,70] and using protein–ligand fingerprints.[32,71] Protein–ligand interaction fingerprints try to describe the binding mode as a sequence of bits, which can encode the information about the type of interaction (polar or not polar), the location of residues and atoms involved in the interaction and the role of participating atoms (hydrogen bond donors or acceptors).[72] The fingerprints can be used to prioritize compounds whose interaction fingerprints are similar to the ones of known binders. In a benchmarking study, Da and Kireev[32] reported that using a protein–ligand interaction fingerprint (SPLIF) was the best method at discriminating active compounds from decoys in 10 targets of DUD-E database.[73]

2.1.5 | Machine learning scoring

Machine learning methods, whether based on random forest[74] or modern convolutional neural networks (CNN),[33,75] allow for an automatic training of data-driven scoring functions. They benefit from fast developments in the field of machine learning as well as large and growing experimental databases.[68] These methods have pushed the state-of-the-art forward, offering greater accuracy than conventional methods. Interestingly, they can be fine-tuned by training on a tailored dataset containing only ligands for a particular target. The resulting scoring function is more aware of the specifics of the target. Such an approach is used by DeltaDelta,[34] which is trained on a congeneric series of ligands docked to the target and offers more accurate results than more computationally expensive methods like FEP. Finally, the possibility of easily adding new examples to the training set allows the model to be up-to-date with the new structures that are deposited in private or public databases. The main drawback is the lack of interpretability of most ML models which makes it impossible to deduce the rationale behind the prediction.[76] Some promising work, however, has tried to address this problem by using methods such as masking (evaluating how the deletion of some atoms affects the prediction) or by studying the gradients of the prediction with respect to the atoms used as input.[77] Another problem is over-fitting to the training set, which can result in a poor prediction.

2.1.6 | Molecular dynamics rescoring

Another way to improve docking predictions is to re-score the top poses with more precise computational methods. Some approaches assess the stability of a docked pose through metadynamics[78] or dynamic undocking.[79,80] These methods follow the rationale that stable, well-docked poses are strongly anchored to the protein and will remain docked after a series of short simulations, while poor poses will leave the pocket. Other methods are based on binding free energy estimations and have been successful at improving hit rates of VS.[81–83] In particular MM-GBSA was developed to more accurately assess the relative binding free energy from MD simulations.[84–86] Over the years free energy methods went through major improvements,[87–89] with free energy perturbation (FEP) standing out. FEP methods allow to compute the binding free energy of a ligand by performing a MD with the simultaneous alchemical transformation from one ligand to a reference structure with known affinity.[90–92] These methods can reach affinity predictions with accuracies of 1–2 kcal/mol,[93,94] allowing medicinal chemists to shortlist the set of compounds for synthesis and testing. Further improvements were made by applying machine learning to FEP. As demonstrated by Rufa et al., the researchers managed to reduce the error of absolute binding free energies from 0.97 to 0.47 kcal/mol for a congeneric ligand series for non-receptor tyrosine kinase TYK2 by correcting the conventional MM simulation with a neural network potential.[95]

2.1.7 | Consensus scoring

Finally, another method to compensate the defects of scoring functions is to use several different scoring functions at once and look for a consensus among them. This protocol has proven valuable in multiple studies[36,37] but it requires having access to multiple docking packages, some level of expertise on each of them, ensuring file compatibility between packages and more computing time. Furthermore, the selection of which programs to use can be problematic: if their scoring functions suffer from the same problems (for being trained in similar datasets), their errors can add up instead of compensate each other.

2.2 | Protein flexibility and docking

Protein flexibility is particularly challenging because it is hard to know how a ligand might impact the conformation of the protein pocket, that is, which side chains might flip or rearrange as a result of interacting with it (induced fit), or which of all the possible protein conformations the ligand will bind to (conformation selection).[96] Even subtle changes in the volume of the pocket, like pocket breathing[46] can directly impact the docking results as it influences the steric constraints in the binding pocket. In extreme cases, large changes in the backbone of the protein take place, leading to the opening or closing of entire subpockets.[97] Figure 1 shows an example of a tyrosine kinase adopting two drastically different pocket conformations.[98] Cryptic sites are not usually apparent on the crystallographic structures of the target,[99] but can become visible in the presence of some molecules that either trigger their opening or stabilize that conformation.[27] Therefore, docking a ligand into such pockets might be simply impossible without sampling its opened conformation first. We will now discuss some methods that try to address protein flexibility in docking and its limitations.

2.2.1 | Soft potentials

The volume and shape of the binding pocket might change upon binding. One approach to deal with these minor changes is the use of soft potentials,[38] which reduce or attenuate the penalty for clashes. Practically, this allows for minor clashes between the ligand and the protein, assuming that the protein will make room to accommodate the clashing atoms. Despite the fact that it can only account for small variations in protein structure, soft docking is an appealing way to account for protein flexibility. It is easy to implement, as it only requires manipulating the parameters of the scoring function, and does not reduce docking speed, which makes it practical in high-throughput applications. An equivalent approach is the modification of the VdW radii of the atoms in the protein, ligand, or both.[40]

2.2.2 | Ensemble docking

Ensemble docking[39] is another solution that accounts for protein flexibility. In this method, ligands are docked to an ensemble of conformations of the target protein. It is convenient and powerful, because it incorporates multiple protein conformations in the docking pipeline, allowing the docking protocol to explore both minor and major protein pocket

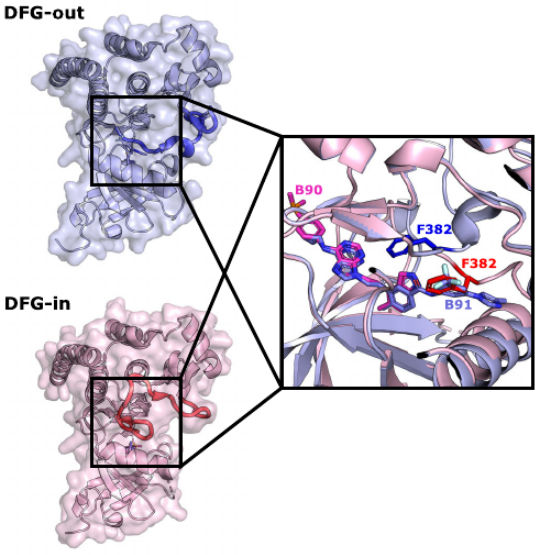


FIGURE 1 An example of a large change in the conformation of a pocket. Abl tyrosine kinase can adopt two dramatically different pocket conformations named DFG-in and DFG-out. In the DFG-out conformation (top left panel, DFG motif is shown in bright blue cartoon; PDBid: 3KFA), an extra allosteric pocket next to the main cavity becomes accessible, and the ligand B91 (purple licorice) can fit into it, while the DFG-in conformation (bottom left panel, DFG motif in red cartoon; PDBid: 3KF4) does not. On the right panel, zoomed in overlapped structures show that the purple ligand (B91) clashes with phenylalanine (F382) in the DFG motif in the DFG-in conformation[98]

rearrangements. Modeling large changes that involve both side chains and backbone is challenging for any docking algorithm or even MD protocols; hence, in some cases, ensemble docking might be the only possibility to approach the conformational diversity of the target. The ensemble of conformations can be obtained through side-chain rotamer sampling, MD simulations and from structural databases.[39,100,101] The method comes, however, with its own limitations, namely, increased computational cost, the risk of docking ligands into conformations not visited by the protein in reality (artifacts), increased number of poses from which the right one will have to be selected, or issues with the generation of conformations and selection of the ensemble.[39,40,102]

2.2.3 | Explicit flexibility in docking

Flexibility of the pocket can also be accounted for explicitly. Induced fit docking (IDF) protocol described by Sherman et al. can capture induced fit effects in a computationally efficient manner.[40] In this protocol, the ligand is first docked into a rigid pocket with a soft potential using Glide.[16] In addition to the use of a soft potential, up to three residues in the pocket can be mutated to alanine if their position is suspected to be ambiguous, allowing the ligand to occupy that space. In the second step, Prime software[103–105] explores different conformations of the protein structure in docked complexes by means of side chain sampling, followed by structure minimization. A second round of docking follows, where the ligand is redocked to the protein conformations sampled in the first round, this time using a hard potential. The final docked solutions are then scored accounting for the docking energy, the strain of the protein and the solvation terms.

Another protocol able to handle induced-fit effects is protein energy landscape exploration (PELE).[4] PELE is a search protocol which includes protein and ligand flexibility explicitly. Local perturbations of the ligand are introduced via displacements, rotations or changes in dihedral angles, followed by side chain sampling. The best conformation is minimized, which allows the movement of the backbone, and scored, leading to the acceptance or rejection of that conformation. Finally, a new round of local perturbation begins, starting a new cycle. This series of cycles result in a collection of highly redundant local minimas (a trajectory), however, large changes do occur from the start of the trajectory until its last frame. Although much slower than docking, PELE is still orders of magnitude faster than a typical MD run. Authors have reported success in reproducing entry and exit pathways of several protein–ligand complexes, and in sampling protein–ligand poses in close proximity to crystal structures for very challenging targets such as GPCRs.[106] Providing the entry pathways and identifying possible meta-stable transition states adds more value than just predicting the binding mode.[102]

2.3 | High-throughput methods for large chemical libraries

Docking as a high-throughput tool became very fast and efficient, with less than a minute needed to dock a compound against a receptor on a single CPU core. This has recently enabled the exploration of ultra-large libraries in the order of hundred million compounds.[15] However, a large computational infrastructure is required to exhaustively sample the ever-growing space of commercially available compounds in a reasonable amount of time, effectively limiting most docking campaigns to much smaller libraries.

Some protocols have been developed to alleviate this problem. One such protocol consists in the use of a series of subsequent, hierarchical filters.[107] The first filters in the hierarchy are usually very fast but also relatively inaccurate. Their main objective is to quickly discard the least promising compounds in the library and avoid modeling them with more expensive methods. The reduced set of molecules is then promoted and undergoes a series of more accurate and computationally demanding steps. For instance, Ashutosh Kumar et al. used a hierarchical approach where a library of compounds was first docked using a fast, rigid protein approach. Then, the top-ranking compounds from that campaign were re-docked again using a protein flexible methodology. Finally, an MD protocol was used to further refine the ranking.[108]

Other approaches use active learning to train a machine learning model, on the fly, while the docking campaign is running.[109–111] The model is trained on a sample of the molecules which have already been docked, predicts which of the remaining ligands will not dock favorably and removes them from the docking library. These models can be simple random forests or more complex, deep neural networks,[109] and work with molecular fingerprints or other descriptors. The rationale behind these approaches is that, after docking a subset of the library, the model can learn which ligand

features lead to good docking scores and which ones lead to bad docking scores. For instance, if the pocket is highly hydrophilic, the model could learn to discard ligands which are very hydrophobic. The recent study by Graff et al. reported that by evaluating only the 2.4% of a library of 100 million compounds, 87.9% of the top 50,000 ligands (as identified by the original VS protocol) were recovered by the active learning approach.[110] Such result indicates that there is a great potential in active learning protocols to drastically reduce the amount of computations needed to screen a library of compounds.

These two approaches, hierarchical and active learning, can actually be combined, as exemplified in the study by Konze et al.[111] In the protocol described in the study, a library of compounds is first generated via retrosynthesis and then docked against the target. The best compounds undergo an expensive FEP simulation to better evaluate their binding affinity. Here, active learning is applied to predict the FEP score that the remaining compounds will obtain. Subsequent FEP simulations are only executed for the most promising molecules.

Both methodologies, active learning and hierarchical docking, can dramatically reduce the computational costs of evaluating a library of compounds, allowing for broader sampling of the chemical space, which is key to find new chemical scaffolds.

3 | PHYSICS-BASED MOLECULAR SIMULATIONS IN VIRTUAL SCREENING

The second group of algorithms is based on molecular dynamics methods. All atom molecular simulations allow to address some of the major problems of docking: protein flexibility, structural waters, and even coordinating atoms.[112] Unfortunately, the entry barrier in terms of expertise required for running MD simulations is probably higher than that for docking. Furthermore, MD algorithms are slower and they need large computing resources. However, as we will see, methods like coarse-graining and adaptive sampling strategies can alleviate this problem, and even short, nanosecond-long MD simulations can provide useful insights in protein flexibility.

3.1 | Pocket discovery

Cryptic sites represent a challenge for docking algorithms, but have great pharmaceutical significance because they can act as allosteric pockets, might be the only druggable site in a target and can be involved in protein–protein interactions.[113–115] MD protocols have proven successful at identifying these pockets,[27,61,113,116–118] and probably are the only in silico methods that can sample their opened conformations.[119] These approaches range from nanosecond[27] to microsecond-long simulations complemented with Markov State Model analysis.[113] This long spectrum of timescales is aligned with the kinetics of the pocket opening. While some cryptic pockets only need some fast rearrangement of the side-chains, others require larger and slower changes in the protein conformation.[27,118]

Mixed solvent molecular dynamics (MSMD) protocols, like pyMDMix,[120] mixMD,[61] and CrypticScout,[27] simulate a protein in a solution of water and a co-solvent, which is typically a small organic molecule. The co-solvent molecules interact with the surface of the protein over the simulation and can trigger the opening of hidden pockets (Figure 2). Mapping the occupancy and residence time of the probe leads to identification of binding hotspots on a protein surface that can discriminate binding pockets. The study by Martinez-Rosell et al. showed that benzene binding hotspots identified by CrypticScout corresponded to experimentally determined cryptic pockets in a set of 18 different systems.[27] The hotspots can be translated to a minimal pharmacophore and used as restrains in VS.

The use of longer time scales coupled with Markov state models (MSM) was studied by Bowman and Geissler to identify allosteric cryptic pockets in β-lactamase.[113] In this study, they run hundreds of simulations of 500 nanoseconds each, leading to a total aggregated time of 100 μs. The data were then analyzed by MSM, where kinetic and structural information is used to cluster the simulations into a set of discrete states, representing local minimas. Equilibrium populations of each state and transition probabilities between them can then be estimated from the trajectories. After building a MSM for the simulations of the β-lactamase, a pocket detector algorithm (LIGSITE[121]) was applied to representatives of each of the states. This approach was able to correctly identify the known cryptic site as a transient pocket. More interestingly, in a follow up study, Hart et al. used the open conformations of the cryptic pocket sampled by the MD to run a docking campaign.[119] They identified two modulators of this enzyme, one activator with EC50 values of 63 ± 9 μM and one inhibitor with EC50 of 57 ± 3 μM. This is a clear and successful example

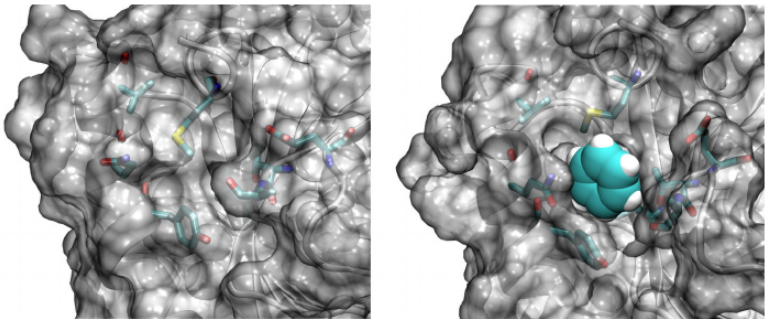


FIGURE 2 Opening of a small cryptic pocket in GTPase KRas with CrypticScout. The binding of the benzene probe disrupts the contacts between three residues (Gln, Met, Tyr) and opens the cavity. (PDB code: 4L8G[123])

of MD methods complementing docking. More information about MSM analysis can be found in the review by Pande et al.[122]

Another method, developed by Oleinikovas et al., called Sampling Water Interfaces through Scaled Hamiltonians (SWISH), makes the nonbonded interactions between water molecules and apolar protein residues more favorable.[118] This approach makes the water molecules more ligand-like and hydrophobic, and in consequence allows them to open cryptic pockets. The effectiveness of this protocol was further improved by the incorporation of fragments into the solution. The authors hypothesize that this additional success is a result of the fragments stabilizing and further opening the cryptic pockets exposed by the waters.

Finally, MD simulations can be used to study other relevant conformational changes. As we have discussed before, some kinases can switch between two states: DFG-in and DFG-out, which dramatically affect the conformation of the main and allosteric pockets (see Figure 1). Shan et al. run microsecond long MD simulations to explore the transition between active and inactive states of a kinase,[124] demonstrating the capacity of MD to sample drastically different protein conformations for the most challenging targets. Similarly, Lovera et al. used MD simulations to sample the active states of a GPCR starting from an inactive, crystal structure.[125]

3.2 | In-silico binding assays

In addition to opening hidden pockets and sampling the conformational landscape, MD can also simulate the entire binding process of a ligand to the target of interest in what we call in silico binding assay (ISBA). This approach has been successful at recovering the native binding mode of the ligand,[20,25,112] and estimating kinetics and binding free energies.[126] Importantly, valuable metrics can be obtained from the simulations, like the population of the different binding modes, which can then be used to select one of them, alleviating the scoring problem we discussed in docking.[102]

In the pioneer work, Shan et al. reproduced the binding pathway of two molecules, the cancer drug dasatinib and the kinase inhibitor PP1, to the target Src kinase, with long, unbiased MD simulations.[24] In both cases, the simulations started with the unbound ligand and the resulting bound poses were "virtually identical" to those found by X-ray crystallography. Furthermore, the simulations correctly identified the locations of structural waters in the Src kinase–PP1 complex and shed light on the desolvation of the cavity upon ligand binding.

In another study from D. E. Shaw Research, Dror and coworkers used a similar approach to reconstruct the binding of several ligands of β2-adrenergic receptor (β2AR), as a result obtaining binding modes practically identical to the crystal ones.[20] In fact, authors discovered a common binding pathway for all the ligands and identified the two main barriers of the binding process: (a) the protein and ligand dehydration associated with the entry of the ligand into the extracellular vestibule and (b) going from the vestibule to the pocket trough a narrow passage. Buch et al., offered similar insights for the trypsin-benzamidine binding process.[25] They reconstructed the full pathway from unbound to bound states, revealed a series of metastable states through which benzamidine travels to reach the final pose and in result achieved a pose with a RMSD below 2 Å. The same system was further studied by Plattner and Noé, revealing a large

degree of plasticity for trypsin and discovering an extra S1* pocket.[127] Additionally, six different metastable apo structures were identified, which were in close accordance with the crystallographic structures of other serine proteases. In the work by Guixa-Gonzalez et al., the authors used MD simulations to show that cholesterol molecules from the extracellular leaflet can diffuse through helices TM5–6 into the orthosteric pocket of a GPCR, and occupy a key area the pocket.[128] This discovery was confirmed by further experimental (non in silico) validation. The study by Ferruz and coworkers offers a great example of the advantages of MD over docking.[28] They started by docking a ligand into dopamine D3 receptor. The predicted pose did not agree with mutagenesis experiments. Therefore, the authors followed up with an ISBA, starting with an unbound ligand. The molecule then diffused toward the binding site, leading to a binding mode that differed from the docked one. Furthermore, a cryptic pocket opened upon binding, allowing the ligand to bind deeper into the cavity. This new pose was in agreement with mutagenesis information.

Another study by Ferruz et al. showcases the capabilities of MD to naturally approach a challenging case of ion coordination.[112] They investigated the binding mode of the substrate of Myo-inositol monophosphatase (IMPase). It was known that the catalysis of this enzyme requires 3 Mg^{2+} ions acting as cofactors. However, the binding process by which the substrate enters the catalytic site in cooperation with these ions was unknown. To shed light on this process, the authors simulated the diffusion of individual magnesium ions to the cavity one at a time. Then, they explored the binding of the substrate to the complex with two (IMPase-II) and three magnesium ions (IMPase-III). The substrate was seen to bind to both IMPase-III and IMPase-II, in coordination with Mg^{2+} or alone. Furthermore, one of the poses obtained from the MSM analysis showed great agreement with the crystallographic structure. This shows the capabilities of MD protocols to accurately describe complex scenarios where several bodies (protein, ligand, and cofactor) cooperate.

The previous articles demonstrate the capabilities of MD to predict binding modes and binding pathways. The study by Martinez-Rosell and coworkers tests ISBA in a VS campaign for a small library of compounds. They screened 129 fragments against the target CXCL12, where they predicted their binding modes, as well as kinetic rates (kon and koff) and ligand efficiencies.[126] Eight of the fragments had an estimated ligand efficiency higher than 0.3. The predicted binding modes agreed with the available structural data of that target, in terms of placement of pharmacophoric features, but no experimental validation could be performed.

3.3 | Improving sampling

Some of these ISBA experiments used an adaptive sampling protocol, which can decrease the amount of time needed for the simulation to converge by an order of magnitude.[21] This protocol uses multiple, parallel short simulations of around 80 ns long instead of a single, long one. These initial simulations are used to build a MSM which clusters the sampled conformational space. Subsequent short simulations are launched starting from conformations belonging to underpopulated clusters. This allows the protocol to avoid re-sampling of already explored areas, reaching convergence faster and, hence, reducing the computational costs. Another method that helps to increase the performance of MD is the use of coarse-grained potentials, where several atoms are represented and parameterized as a single entity. Such an approach was used in the recent study, where Souza et al. reported success in predicting binding modes for several targets of pharmaceutical interest, including a GPCR.[22]

Another method, known as replica exchange, can greatly accelerate the sampling of the energetic landscape. This method begins with multiple, short simulations (replicas) of the same system at different temperatures. After a predefined number of steps, the conformations from different replicas are exchanged based on the Metropolis criterion.[41] Simulating at different temperatures helps exploring different wells of the energy landscape, quickly overcoming large barriers and avoiding getting stuck in local minima.

One more way to accelerate ISBA is biased MD. In these protocols, the simulation is biased by an appropriate collective variable (CV).[129] For most of these methods, prior knowledge of the system is necessary to define the CV. However, the protocol proposed by Spitaleri et al. offers a generalized bias that could work on any protein–ligand system.[23] Such bias creates an artificial electrostatic interaction between the atoms of the ligand and the atoms of the binding site. The combination of the regular forces and artificial ones helps to identify a path leading to the binding site, which the ligand is "invited to follow." Following such path, instead of diffusing around the simulation box, avoids wasting time in the transition phase. When the ligand is close to the binding site, the strength of those artificial forces is decreased to allow a non-biased molecular recognition. The authors report a speed up of 2–3 orders of magnitude with the respect to unbiased MD. The method correctly predicted the binding modes of all six studied ligand, identifying all main

interactions, including a water mediated interactions, and resulted in poses with RMSD values below 2.5 Å with respect to the crystal structure. Furthermore, the predicted binding pathways of three protein–ligand complexes were similar to those reported by plain, unbiased MD simulations.

4 | DISCUSSION

Despite the limitations we have addressed in this review, classical, fast docking protocols have proven successful in a number of prospective studies,[12–15] and they have done so for the right reasons; that is, by predicting the right pose and assigning its compound a great affinity that ranked it among the best in the library.[44] However, one could argue that such success might not be directly attributable to docking itself, but rather to the additional steps taken by the practitioner, such as visual inspection, the addition of prior knowledge or rescoring with alternative methods. This would mean that docking per-se is not that powerful, showcasing it as a simple "pose generator" that requires further work to provide meaningful results. However, unsupervised or largely-unsupervised docking campaigns have been successful too.

In a recent article, Lolli and Caflisch used docking software SEED to screen 350 fragments against the target BAZ2B.[130] Of these, 12 fragments were prioritized for testing. Soaking experiment resulted in 4 fragments bound in the binding pocket. Furthermore, the authors report that the binding mode predicted by their docking software was correct for three of the four fragments, two of them being practically identical according to the figure. This example is particularly impressive for two reasons: (a) fragments are believed to be particularly challenging for docking algorithms,[102] (b) there was no use of prior knowledge nor visual inspection. In another work, Hermann et al. discovered a substrate of an orphan enzyme Tm0936 by docking a library of metabolites into its catalytic site.[131] They performed docking using DOCK3.5.54 into an apo structure, which is more challenging than docking to a ligand-bound conformation. The molecule which was then confirmed to be the substrate (SAH) was ranked 6th among 4207 molecules. Moreover, the crystal structure of selected molecule agreed almost completely with the docking prediction, highlighting the success of such approach. Another valuable example is the initiative of Continuous Evaluation of Ligand Protein Prediction (CELPP).[132] In this program, a set of fully automated docking protocols is challenged with predicting the binding modes of a new group of protein–ligand complexes on a weekly basis. The ligands are docked to a structure of a close homolog bound to a similar molecule, which makes it a difficult cross-docking experiment. Although it is not a typical VS campaign, it offers valuable information regarding docking power for several reasons. First, the docking protocols are fully automated, which allows for a more fair comparison among them and removes the impact of the human expert. Second, the evaluation is blind and prospective, therefore unintended biases do not affect the results. Finally, the challenge showcases the strengths and limitations of using docking as an out-of-the-box tool, without any manual parameter tweaking. Although the number of adopters is still low, some protocols have already docked over 3000 molecules. One of them, named *kauoh*, has a Q1 value of RMSD 1.81, meaning that a large fraction of the predicted poses were under the cutoff value of 2.0 Å. Figure 3 shows the current status of the contest.

Despite this success, it is true that classical docking algorithms, due to their need to be fast paced, will probably not be able to deal with particular challenging scenarios, like those where structural waters or coordination atoms play a key role, or those involving cryptic pockets or major changes in pocket conformations. This is where MD based methods have an edge, as they can naturally account for all these scenarios, with the additional benefit of providing valuable mechanistic and thermodynamics insights, like reconstructing binding pathways, shedding light into the desolvation process, or identifying high-energy barriers or metastable states. Some of these insights cannot even be obtained by crystallography experiments. Although the expertise required to run MD simulations is still high and their speed is still very limited, improvements in hardware and new methods like those we have discussed could drastically change this. It must be remembered that until the appearance of GPUs, MD simulations were practically limited to the nanosecond-scale.[133]

The two paradigms, fast and slow, can easily be combined, and we have seen examples of such in this review, like using MD simulations to rescore poses, or sampling the opened conformation of a cryptic pocket to then run a VS campaign. Furthermore, they can be used in different stages of the VS campaign: while docking can filter the vast majority of the library, MD can provide more accurate predictions for a subset of it, re-scoring the best compounds or providing mechanistic insights into their binding pathways.[129]

Most drug discovery projects today rely to some extent in in silico modeling,[134,135] typically combining different approaches and techniques (LBVS, SBVS, FEP, ADMET prediction, and so on) in parallel or in a hierarchical

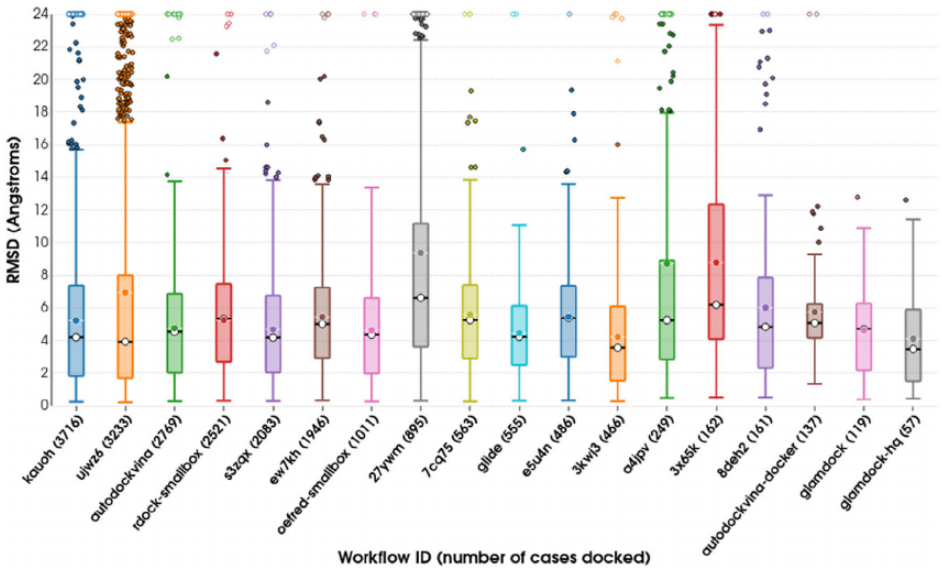


FIGURE 3 CELPP competition results as of February 25th, 2021. A set of automated docking protocols (X-axis) are challenged weekly to predict the binding mode of several ligands to a target. Y-axis reports the distribution of the RMSD values for the predicted poses. The number of attempted docking exercises for each protocol is reported next to its name, inside parenthesis

fashion.[108] Several companies performing in silico modeling have appeared in the market, and those using AI to help drug discovery raised $1 billion in funding in 2018,[136] signaling a strong trust in computational methodologies.

5 | CONCLUSIONS

From multiple success stories discussed in this review, it is apparent that computational methods can be powerful tools in drug discovery. Docking algorithms can screen large chemical libraries in a high-throughput manner. However, a series of approximations and poor parametrization of scoring function can lead to high false positive rates and inefficient selection of candidates. On the other hand, physics-based molecular simulations can overcome several of the docking issues, at the expense of speed and throughput. Both approaches are orthogonal in their features and speed profiles. Moreover they provide valuable information about different aspects of protein–ligand binding, and can complement each other in VS campaigns.[119] Although a huge variety of computational methods already exist, the field is very active and many more are still being developed. Better hardware and innovative approaches, especially from the field of machine learning, will keep improving the performance of VS algorithms.

ACKNOWLEDGMENTS

The authors thank Iain Maryanow for proofreading this manuscript. The authors thank Acellera Ltd. for funding. Gianni De Fabritiis acknowledges support from MINECO (BIO2017-82628-P) and FEDER. This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 823712 (CompBioMed2) and from the Industrial Doctorates Plan of the Secretariat of Universities and Research of the Department of Economy and Knowledge of the Generalitat of Catalonia.

CONFLICT OF INTEREST

Alejandro Varela-Rial is an employee at Acellera Labs, which offers virtual screening services. Maciej Majewski reports no conflict of interest. Gianni De Fabritiis is the co-founder and CEO of Acellera Labs, which offers virtual screening services.

AUTHOR CONTRIBUTIONS

Alejandro Varela-Rial: Writing-original draft. **Maciej Majewski:** Writing-review & editing. **Gianni De Fabritiis:** Supervision; writing-review & editing.

DATA AVAILABILITY STATEMENT

Data sharing not applicable - no new data generated.

ORCID

Alejandro Varela-Rial https://orcid.org/0000-0002-6918-1765
Maciej Majewski https://orcid.org/0000-0003-2605-8166
Gianni De Fabritiis https://orcid.org/0000-0003-3913-4877

RELATED WIREs ARTICLE

Outstanding challenges in protein-ligand docking and structure-based virtual screening

REFERENCES

1. Horvath D. A virtual screening approach applied to the search for trypanothione reductase inhibitors. *J Med Chem*. 1997;*40*:2412–23.
2. Ewing TJA, Makino S, Skillman AG, Kuntz ID. DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *J Comput Aided Mol Des*. 2001;*15*:411–28.
3. Verdonk ML, Cole JC, Hartshorn MJ, Murray CW, Taylor RD. Improved protein-ligand docking using GOLD. *Proteins*. 2003;*52*:609–23.
4. Borrelli KW, Vitalis A, Alcantara R, Guallar V. PELE: protein energy landscape exploration. A novel Monte Carlo based technique. *J Chem Theory Comput*. 2005;*1*:1304–11.
5. Muegge Ingo, Mukherjee Prasenjit. An overview of molecular fingerprint similarity search in virtual screening. *Expert Opinion on Drug Discovery*. 2016;*11*(2):137–148. http://dx.doi.org.sare.upf.edu/10.1517/17460441.2016.1117070.
6. Kumar Ashutosh, Zhang Kam Y. J. Advances in the Development of Shape Similarity Methods and Their Application in Drug Discovery. *Fron Chem*. 2018;*6*:315. http://dx.doi.org.sare.upf.edu/10.3389/fchem.2018.00315.
7. Lombardo F, Desai PV, Arimoto R, Desino KE, Fischer H, Keefer CE, et al. In silico absorption, distribution, metabolism, excretion, and pharmacokinetics (ADME-PK): utility and best practices. An industry perspective from the international consortium for innovation through quality in pharmaceutical development: Miniperspective. *J Med Chem*. 2017;*60*:9097–113.
8. Piñero J, Furlong LI, Sanz F. In silico models in drug development: where we are. *Curr Opin Pharmacol*. 2018;*42*:111–21.
9. Zhu T, Cao S, Su PC, Patel R, Shah D, Chokshi HB, et al. Hit identification and optimization in virtual screening: practical recommendations based on a critical literature analysis. *J Med Chem*. 2013;*56*:6560.
10. Al Olaby RR, Cocquerel L, Zemla A, Saas L, Dubuisson J, Vielmetter J, et al. Identification of a novel drug lead that inhibits HCV infection and cell-to-cell transmission by targeting the HCV E2 glycoprotein. *PLoS One*. 2014;*9*:e111333.
11. Perryman AL, Yu W, Wang X, Ekins S, Forli S, Li SG, et al. A virtual screen discovers novel, fragment-sized inhibitors of *Mycobacterium tuberculosis* InhA. *J Chem Inf Model*. 2015;*55*:645–59.
12. Luciani R, Saxena P, Surade S, Santucci M, Venturelli A, Borsari C, et al. Virtual screening and X-ray crystallography identify non-substrate analog inhibitors of flavin-dependent thymidylate synthase. *J Med Chem*. 2016;*59*:9269–75.
13. Babaoglu K, Simconov A, Irwin JJ, Nelson ME, Feng B, Thomas CJ, et al. Comprehensive mechanistic analysis of hits from high-throughput and docking screens against β-lactamase. *J Med Chem*. 2008;*51*:2502–11.
14. Pereira HM, Berdini V, Cleasby A, Garratt RC. Crystal structure of calf spleen purine nucleoside phosphorylase complexed to a novel purine analogue. *FEBS Lett*. 2007;*581*:5082–6.
15. Lyu J, Wang S, Balius TE, Singh I, Levit A, Moroz YS, et al. Ultra-large library docking for discovering new chemotypes. *Nature*. 2019;*566*:224–9.
16. Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, et al. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J Med Chem*. 2004;*47*:1739–49.
17. Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem*. 2009;*31*:455–461.
18. Ruiz-Carmona S, Alvarez-Garcia D, Foloppe N, Garmendia-Doval AB, Juhos S, Schmidtke P, et al. rDock: a fast, versatile and open source program for docking ligands to proteins and nucleic acids. *PLoS Comput Biol*. 2014;*10*:e1003571.
19. Pantsar T, Poso A. Binding affinity via docking: fact and fiction. *Molecules*. 2018;*23*:1899.
20. Dror RO, Pan AC, Arlow DH, Borhani DW, Maragakis P, Shan Y, et al. Pathway and mechanism of drug binding to G-protein-coupled receptors. *Proc Natl Acad Sci U S A*. 2011;*108*:13118–23.
21. Doerr S, De Fabritiis G. On-the-fly learning and sampling of ligand binding by high-throughput molecular simulations. *J Chem Theory Comput*. 2014;*10*:2064–9.
22. Souza PCT, Thallmair S, Conflitti P, Ramírez-Palacios C, Alessandri R, Raniolo S, et al. Protein–ligand binding with the coarse-grained martini model. *Nat Commun*. 2020;*11*:1–11.

23. Spitaleri A, Decherchi S, Cavalli A, Rocchia W. Fast dynamic docking guided by adaptive electrostatic bias: the MD-binding approach. *J Chem Theory Comput*. 2018;*14*:1727–36.
24. Shan Y, Kim ET, Eastwood MP, Dror RO, Seeliger MA, Shaw DE. How does a drug molecule find its target binding site? *J Am Chem Soc*. 2011;*133*:9181–3.
25. Buch I, Giorgino T, De Fabritiis G. Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. *Proc Natl Acad Sci U S A*. 2011;*108*:10184–9.
26. Zhu J, Lv Y, Han X, Xu D, Han W. Understanding the differences of the ligand binding/unbinding pathways between phosphorylated and non-phosphorylated ARH1 using molecular dynamics simulations. *Sci Rep*. 2017;*7*:1–14.
27. Martinez-Rosell G, Lovera S, Sands ZA, De Fabritiis G. PlayMolecule CrypticScout: predicting protein cryptic sites using mixed-solvent molecular simulations. *J Chem Inf Model*. 2020;*60*:2314–24.
28. Ferruz N, Doerr S, Vanase-Frawley MA, Zou Y, Chen X, Marr ES, et al. Dopamine D3 receptor antagonist reveals a cryptic pocket in aminergic GPCRs. *Sci Rep*. 2018;*8*:1–10.
29. Wang J, Dokholyan NV. MedusaDock 2.0: efficient and accurate protein-ligand docking with constraints. *J Chem Inf Model*. 2019;*59*:2509–15.
30. Varela-Rial A, Majewski M, Cuzzolin A, Martínez-Rosell G, De Fabritiis G. SkeleDock: a web application for scaffold docking in PlayMolecule. *J Chem Inf Model*. 2020;*60*:2673–7.
31. Fischer A, Smieško M, Sellner M, Lill MA. Decision making in structure-based drug discovery: visual inspection of docking results. *J Med Chem*. 2021;*64*:2489–2500.
32. Da C, Kireev D. Structural protein-ligand interaction fingerprints (SPLIF) for structure-based virtual screening: method and benchmark study. *J Chem Inf Model*. 2014;*54*:2555–61.
33. Jiménez J, Škalič M, Martínez-Rosell G, De Fabritiis G. KDEEP: protein-ligand absolute binding affinity prediction via 3D-convolutional neural networks. *J Chem Inf Model*. 2018;*58*:287–96.
34. Jiménez-Luna J, Pérez-Benito L, Martínez-Rosell G, Sciabola S, Torella R, Tresadern G, et al. DeltaDelta neural networks for lead optimization of small molecule potency. *Chem Sci*. 2019;*10*:10911–8.
35. Majewski M, Ruiz-Carmona S, Barril X. Dynamic undocking: a novel method for structure-based drug discovery. *Methods in molecular biology*. Volume 1824. Totowa, NJ: Humana Press Inc; 2018. p. 195–216.
36. Palacio-Rodríguez K, Lans I, Cavasotto CN, Cossio P. Exponential consensus ranking improves the outcome in docking and receptor ensemble docking. *Sci Rep*. 2019;*9*:1–14.
37. Charifson PS, Corkery JJ, Murcko MA, Walters WP. Consensus scoring: a method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J Med Chem*. 1999;*42*:5100–9.
38. Ferrari AM, Wei BQ, Costantino L, Shoichet BK. Soft docking and multiple receptor conformations in virtual screening. *J Med Chem*. 2004;*47*:5076–84.
39. Amaro RE, Baudry J, Chodera J, Demir Ö, JA MC, Miao Y, et al. Ensemble docking in drug discovery. *Biophys J*. 2018;*114*(10):2271–8.
40. Sherman W, Day T, Jacobson MP, Friesner RA, Farid R. Novel procedure for modeling ligand/receptor induced fit effects. *J Med Chem*. 2006;*49*:534–53.
41. Yoshida K, Yamaguchi T, Okamoto Y. Replica-exchange molecular dynamics simulation of small peptide in water and in ethanol. *Chem Phys Lett*. 2005;*412*:280–4.
42. Holyoak T. Molecular recognition: lock-and-key, induced fit, and conformational selection. *Encyclopedia of biophysics*. Berlin Heidelberg: Springer; 2013. p. 1584–8.
43. Cleves AE, Jain AN. Knowledge-guided docking: accurate prospective prediction of bound configurations of novel ligands using Surflex-dock. *J Comput Aided Mol Des*. 2015;*29*:485–509.
44. Kolb P, Irwin J. Docking screens: right for the right reasons? *Curr Top Med Chem*. 2009;*9*(9):755–70.
45. Verdonk ML, Mortenson PN, Hall RJ, Hartshorn MJ, Murray CW. Protein-ligand docking against non-native protein conformers. *J Chem Inf Model*. 2008;*48*:2214–25.
46. Stank A, Kokh DB, Fuller JC, Wade RC. Protein binding pocket dynamics. *Acc Chem Res*. 2016;*49*:809–15.
47. Roberts BC, Mancera RL. Ligand - protein docking with water molecules. *J Chem Inf Model*. 2008;*48*:397–408.
48. Michel J, Tirado-Rives J, Jorgensen WL. Prediction of the water content in protein binding sites. *J Phys Chem B*. 2009;*113*:13337–46.
49. Hu X, Balaz S, Shelver WH. A practical approach to docking of zinc metalloproteinase inhibitors. *J Mol Graph Model*. 2004;*22*:293–307.
50. Chang CEA, Chen W, Gilson MK. Ligand configurational entropy and protein binding. *Proc Natl Acad Sci U S A*. 2007;*104*:1534–9.
51. Klebe G, Böhm HJ. Energetic and entropic factors determining binding affinity in protein-ligand complexes. *J Recept Signal Transduct Res*. 1997;*17*:459–73.
52. Mohan V, Gibbs A, Cummings M, Jaeger E, DesJarlais R. Docking: successes and challenges. *Curr Pharm Des*. 2005;*11*:323–33.
53. Huang SY. Comprehensive assessment of flexible-ligand docking algorithms: current effectiveness and challenges. *Brief Bioinform*. 2018;*19*:982–94.
54. Huey R, Morris GM, Olson AJ, Goodsell DS. A semiempirical free energy force field with charge-based desolvation. *J Comput Chem*. 2007;*28*:1145–52.
55. Warren GL, Andrews CW, Capelli AM, Clarke B, LaLonde J, Lambert MH, et al. A critical assessment of docking programs and scoring functions. *J Med Chem*. 2006;*49*:5912–31.
56. Dixon JS. Evaluation of the CASP2 docking section. *Proteins Struct Funct Genet*. 1997;*29*:198–204.

57. Li Y, Han L, Liu Z, Wang R. Comparative assessment of scoring functions on an updated benchmark: 2. Evaluation methods and general results. *J Chem Inf Model*. 2014;*54*:1717–36.
58. Ghosh AK, Chapsal BD, Weber IT, Mitsuya H. Design of HIV protease inhibitors targeting protein backbone: an effective strategy for combating drug resistance. *Acc Chem Res*. 2008;*41*:78–86.
59. Chen Y, Shoichet BK. Molecular docking and ligand specificity in fragment-based inhibitor discovery. *Nat Chem Biol*. 2009;*5*:358–64.
60. Caporuscio F, Rastelli G, Imbriano C, Del Rio A. Structure-based design of potent aromatase inhibitors by high-throughput docking. *J Med Chem*. 2011;*54*:4006–17.
61. Ghanakota P, Carlson HA. Moving beyond active-site detection: MixMD applied to allosteric systems. *J Phys Chem B*. 2016;*120*:8685–95.
62. Hu B, Lill MA. Exploring the potential of protein-based pharmacophore models in ligand pose prediction and ranking. *J Chem Inf Model*. 2013;*53*:1179–90.
63. Skalic M, Varela-Rial A, Jiménez J, Martínez-Rosell G, De Fabritiis G. LigVoxel: inpainting binding pockets using 3D-convolutional neural networks. *Bioinformatics*. 2019;*35*:243–50.
64. Skalic M, Sabbadin D, Sattarov B, Sciabola S, De Fabritiis G. From target to drug: generative modeling for the multimodal structure-based ligand design. *Mol Pharm*. 2019;*16*:4282–91.
65. Fradera X, Mestres J. Guided docking approaches to structure-based design and screening. *Curr Top Med Chem*. 2005;*4*:687–700.
66. Lamoree B, Hubbard RE. Current perspectives in fragment-based lead discovery (FBLD). *Essays Biochem*. 2017;*61*(5):453–64.
67. Parks CD, Gaieb Z, Chiu M, Yang H, Shao C, Walters WP, et al. D3R grand challenge 4: blind prediction of protein–ligand poses, affinity rankings, and relative binding free energies. *J Comput Aided Mol Des*. 2020;*34*:99–119.
68. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The protein data bank. *Nucleic Acids Res*. 2000;*28*:235–42.
69. Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK, et al. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J Comput Chem*. 1998;*19*:1639–62.
70. Peach ML, Nicklaus MC. Combining docking with pharmacophore filtering for improved virtual screening. *J Cheminform*. 2009;*1*:6.
71. Deng Z, Chuaqui C, Singh J. Structural interaction fingerprint (SIFt): a novel method for analyzing three-dimensional protein-ligand binding interactions. *J Med Chem*. 2004;*47*:337–44.
72. Jasper JB, Humbeck L, Brinkjost T, Koch O. A novel interaction fingerprint derived from per atom score contributions: exhaustive evaluation of interaction fingerprint performance in docking based virtual screening. *J Cheminform*. 2018;*10*:15.
73. Mysinger MM, Carchia M, Irwin JJ, Shoichet BK. Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J Med Chem*. 2012;*55*:6582–94.
74. Ballester PJ, Mitchell JBO. A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. *Bioinformatics*. 2010;*26*:1169–75.
75. Ragoza M, Hochuli J, Idrobo E, Sunseri J, Koes DR. Protein-ligand scoring with convolutional neural networks. *J Chem Inf Model*. 2017;*57*:942–57.
76. Jiménez-Luna J, Grisoni F, Schneider G. Drug discovery with explainable artificial intelligence. *Nat Res*. 2020;*2*:573–584.
77. Hochuli J, Helbling A, Skaist T, Ragoza M, Koes DR. Visualizing convolutional neural network protein-ligand scoring. *J Mol Graph Model*. 2018;*84*:96–108.
78. Clark AJ, Tiwary P, Borrelli K, Feng S, Miller EB, Abel R, et al. Prediction of protein–ligand binding poses via a combination of induced fit docking and metadynamics simulations. *J Chem Theory Comput*. 2016;*12*:2990–8.
79. Ruiz-Carmona S, Schmidtke P, Luque FJ, Baker L, Matassova N, Davis B, et al. Dynamic undocking and the quasi-bound state as tools for drug discovery. *Nat Chem*. 2017;*9*:201–6.
80. Majewski M, Barril X. Structural stability predicts the binding mode of protein-ligand complexes. *J Chem Inf Model*. 2020;*60*:1644–51.
81. Hou T, Wang J, Li Y, Wang W. Assessing the performance of the molecular mechanics/Poisson Boltzmann surface area and molecular mechanics/generalized born surface area methods. II. The accuracy of ranking poses generated from docking. *J Comput Chem*. 2011;*32*:866–77.
82. Genheden S, Ryde U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opin Drug Discov*. 2015;*10*:449–61.
83. Pu C, Yan G, Shi J, Li R. Assessing the performance of docking scoring function, FEP, MM-GBSA, and QM/MM-GBSA approaches on a series of PLK1 inhibitors. *MedChemComm*. 2017;*8*:1452–8.
84. Zhang X, Perez-Sanchez HC, Lightstone F. A comprehensive docking and MM/GBSA rescoring study of ligand recognition upon binding antithrombin. *Curr Top Med Chem*. 2017;*17*:1631–9.
85. Guimarães CR, Cardozo M. MM-GB/SA rescoring of docking poses in structure-based lead optimization. *J Chem Inf Model*. 2008;*48*:958–70.
86. Kollman PA, Massova I, Reyes C, Kuhn B, Huo S, Chong L, et al. Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc Chem Res*. 2000;*33*:889–97.
87. Brandsdal BO, Österberg F, Almlöf M, Feierberg I, Luzhkov VB, Åqvist J. Free energy calculations and ligand binding. *Adv Protein Chem*. 2003;*66*:123–58.
88. Parenti MD, Rastelli G. Advances and applications of binding affinity prediction methods in drug discovery. *Biotechnol Adv*. 2012;*30*:244–50.

89. Limongelli V, Bonomi M, Parrinello M. Funnel metadynamics as accurate binding free-energy method. *Proc Natl Acad Sci*. 2013;*110*:6358–63.
90. Jorgensen WL, Thomas LL. Perspective on free-energy perturbation calculations for chemical equilibria. *J Chem Theory Comput*. 2008;*4*:869–76.
91. De Vivo M, Masetti M, Bottegoni G, Cavalli A. Role of molecular dynamics and related methods in drug discovery. *J Med Chem*. 2016;*59*:4035–61.
92. Cournia Z, Allen B, Sherman W. Relative binding free energy calculations in drug discovery: recent advances and practical considerations. *J Chem Inf Model*. 2017;*57*:2911–37.
93. Schindler CEM, Baumann H, Blum A, Böse D, Buchstaller HP, Burgdorf L, et al. Large-scale assessment of binding free energy calculations in active drug discovery projects. *J Chem Inf Model*. 2020;*60*:5457–74.
94. Sherborne B, Shanmugasundaram V, Cheng AC, Christ CD, DesJarlais RL, Duca JS, et al. Collaborating to improve the use of free-energy and other quantitative methods in drug discovery. *J Comput Aided Mol Des*. 2016;*30*:1139–41.
95. Rufa DA, Macdonald HEB, Fass J, Wieder M, Grinaway PB, Roitberg AE, et al. Towards chemical accuracy for alchemical free energy calculations with hybrid physics-based machine learning/molecular mechanics potentials. *BioRxiv*. 2020.
96. B-Rao C, Subramanian J, Sharma SD. Managing protein flexibility in docking and its applications. *Drug Discov Today*. 2009;*14*:394–400.
97. Vijayan RSK, He P, Modi V, Duong-Ly KC, Ma H, Peterson JR, et al. Conformational analysis of the DFG-out kinase motif and biochemical profiling of structurally validated type II inhibitors. *J Med Chem*. 2015;*58*:466–79.
98. Zhou T, Commodore L, Huang WS, Wang Y, Sawyer TK, Shakespeare WC, et al. Structural analysis of DFG-in and DFG-out dual Src-Abl inhibitors sharing a common vinyl purine template. *Chem Biol Drug Des*. 2010;*75*:18–28.
99. Vajda S, Beglov D, Wakefield AE, Egbert M, Whitty A. Cryptic binding sites on proteins: definition, detection, and druggability. *Curr Opin Chem Biol*. 2018;*44*:1–8.
100. Zavodszky MI. Side-chain flexibility in protein-ligand binding: the minimal rotation hypothesis. *Protein Sci*. 2005;*14*:1104–14.
101. Cozzini P, Kellogg GE, Spyrakis F, Abraham DJ, Costantino G, Emerson A, et al. Target flexibility: an emerging consideration in drug discovery and design. *J Med Chem*. 2008;*51*:6237–55.
102. Gioia D, Bertazzo M, Recanatini M, Masetti M, Cavalli A. Dynamic docking: a paradigm shift in computational drug discovery. *Molecules*. 2017;*22*:2029.
103. Jacobson MP, Friesner RA, Xiang Z, Honig B. On the role of the crystal environment in determining protein side-chain conformations. *J Mol Biol*. 2002;*320*:597–608.
104. Jacobson MP, Pincus DL, Rapp CS, TJF D, Honig B, Shaw DE, et al. A hierarchical approach to all-atom protein loop prediction. *Proteins*. 2004;*55*:351–67.
105. *Schrödinger Release 2021-1*. New York, NY: Prime, Schrödinger, LLC; 2021.
106. Lecina D, Gilabert JF, Guallar V. Adaptive simulations, towards interactive protein-ligand modeling. *Sci Rep*. 2017;*7*:1–11.
107. Kumar A, Zhang KYJ. Hierarchical virtual screening approaches in small molecule drug discovery. *Methods*. 2015;*71*:26–37.
108. Kumar A, Ito A, Hirohama M, Yoshida M, Zhang KYJ. Identification of sumoylation activating enzyme 1 inhibitors by structure-based virtual screening. *J Chem Inf Model*. 2013;*53*:809–20.
109. Gentile F, Agrawal V, Hsing M, Ton AT, Ban F, Norinder U, et al. Deep docking: a deep learning platform for augmentation of structure based drug discovery. *ACS Cent Sci*. 2020;*6*:939–49.
110. Graff DE, Shakhnovich EI, Coley CW. Accelerating high-throughput virtual screening through molecular pool-based active learning. *arXiv*. 2020.
111. Konze KD, Bos PH, Dahlgren MK, Leswing K, Tubert-Brohman I, Bortolato A, et al. Reaction-based enumeration, active learning, and free energy calculations to rapidly explore synthetically tractable chemical space and optimize potency of cyclin-dependent kinase 2 inhibitors. *J Chem Inf Model*. 2019;*59*:3782–93.
112. Ferruz N, Tresadern G, Pineda-Lucena A, De Fabritiis G. Multibody cofactor and substrate molecular recognition in the myo-inositol monophosphatase enzyme. *Sci Rep*. 2016;*6*:1–10.
113. Bowman GR, Geissler PL. Equilibrium fluctuations of a single folded protein reveal a multitude of potential cryptic allosteric sites. *Proc Natl Acad Sci U S A*. 2012;*109*:11681–6.
114. Durrant JD, McCammon JA. Molecular dynamics simulations and drug discovery. *BMC Biol*. 2011;*9*:71.
115. Wells JA, McClendon CL. Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. *Nature*. 2007;*450*:1001–9.
116. Bowman GR, Bolin ER, Hart KM, Maguire BC, Marqusee S. Discovery of multiple hidden allosteric sites by combining Markov state models and experiments. *Proc Natl Acad Sci U S A*. 2015;*112*:2734–9.
117. Cruz MA, Frederick TE, Singh S, Vithani N, Zimmerman MI, Porter JR, et al. Discovery of a cryptic allosteric site in Ebola's 'undruggable' VP35 protein using simulations and experiments. *BioRxiv*. 2020.
118. Oleinikovas V, Saladino G, Cossins BP, Gervasio FL. Understanding cryptic pocket formation in protein targets by enhanced sampling simulations. *J Am Chem Soc*. 2016;*138*:14257–63.
119. Hart KM, Moeder KE, Ho CMW, Zimmerman MI, Frederick TE, Bowman GR. Designing small molecules to target cryptic pockets yields both positive and negative allosteric modulators. *PLoS One*. 2017;*12*:e0178678.
120. Seco J, Luque FJ, Barril X. Binding site detection and druggability index from first principles. *J Med Chem*. 2009;*52*:2363–71.
121. Hendlich M, Rippmann F, Barnickel G. LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *J Mol Graph Model*. 1997;*15*:359–63.

122. Pande VS, Beauchamp K, Bowman GR. Everything you wanted to know about Markov state models but were afraid to ask. *Methods*. 2010;*52*(1):99–105.
123. Ostrem JM, Peters U, Sos ML, Wells JA, Shokat KM. K-Ras(G12C) inhibitors allosterically control GTP affinity and effector interactions. *Nature*. 2013;*503*:548–51.
124. Shan Y, Arkhipov A, Kim ET, Pan AC, Shawa DE. Transitions to catalytically inactive conformations in EGFR kinase. *Proc Natl Acad Sci U S A*. 2013;*110*:7270–5.
125. Lovera S, Cuzzolin A, Kelm S, De Fabritiis G, Sands ZA. Reconstruction of apo A2A receptor activation pathways reveal ligand-competent intermediates and state-dependent cholesterol hotspots. *Sci Rep*. 2019;*9*:1–10.
126. Martinez-Rosell G, Harvey MJ, De Fabritiis G. Molecular-simulation-driven fragment screening for the discovery of new CXCL12 inhibitors. *J Chem Inf Model*. 2018;*58*:683–91.
127. Plattner N, Noé F. Protein conformational plasticity and complex ligand-binding kinetics explored by atomistic simulations and Markov models. *Nat Commun*. 2015;*6*:1–10.
128. Guixà-González R, Albasanz JL, Rodriguez-Espigares I, Pastor M, Sanz F, Martí-Solano M, et al. Membrane cholesterol access into a G-protein-coupled receptor. *Nat Commun*. 2017;*8*:1–12.
129. Gervasio FL, Laio A, Parrinello M. Flexible docking in solution using metadynamics. *J Am Chem Soc*. 2005;*127*:2600–7.
130. Lolli G, Caflisch A. High-throughput fragment docking into the BAZ2B Bromodomain: efficient in silico screening for X-ray crystallography. *ACS Chem Biol*. 2016;*11*:800–7.
131. Hermann JC, Marti-Arbona R, Fedorov AA, Fedorov E, Almo SC, Shoichet BK, et al. Structure-based activity prediction for an enzyme of unknown function. *Nature*. 2007;*448*:775–9.
132. Wagner JR, Churas CP, Liu S, Swift RV, Chiu M, Shao C, et al. Continuous evaluation of ligand protein predictions: a weekly community challenge for drug docking. *Structure*. 2019;*27*:1326–35.
133. Harvey MJ, Giupponi G, De Fabritiis G. ACEMD: accelerating biomolecular dynamics in the microsecond time scale. *J Chem Theory Comput*. 2009;5:1632–9.
134. Macalino SJY, Gosu V, Hong S, Choi S. Role of computer-aided drug design in modern drug discovery. *Arch Pharm Res*. 2015;*38*(9):1686–701.
135. Usha T, Shanmugarajan D, Goyal AK, Kumar CS, Middha SK. Recent updates on computer-aided drug discovery: time for a paradigm shift. *Curr Top Med Chem*. 2018;*17*:3296–307.
136. Freedman DH. Hunting for new drugs with AI. *Nature*. 2019;*576*:S49–53.

How to cite this article: Varela-Rial, A., Majewski, M., & De Fabritiis, G. (2021). Structure based virtual screening: Fast and slow. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, e1544. https://doi.org/10.1002/wcms.1544

3.4 PlayMolecule Glimpse: Understanding protein-ligand property predictions with interpretable neural networks

Varela-Rial, A., Maryanow I., Majewski M., Doerr S., Schapin N., Jiménez, J., de Fabritiis, G. PlayMolecule Glimpse: Understanding protein-ligand property predictions with interpretable neural networks. Journal of Chemical Information and Modeling. *Accepted for publication.*

Summary

In this article we presented Glimpse, an application to perform model interpretability using the Integrated Gradients method. In addition to describing the application, we also presented the results obtained applying this method to three different neural networks and draw conclusions regarding whether such networks were learning some structural biology principles or not.

PlayMolecule Glimpse: Understanding protein-ligand property predictions with interpretable neural networks

Alejandro Varela-Rial,[†,‡] Iain Maryanow,[‡] Maciej Majewski,[†] Stefan Doerr,[‡] Nikolai Schapin,[†,‡] José Jiménez-Luna,[†] and Gianni De Fabritiis[*,†,‡,¶]

†*Computational Science Laboratory, Universitat Pompeu Fabra, Barcelona Biomedical Research Park (PRBB), Carrer Dr. Aiguader 88, 08003, Barcelona, Spain*
‡*Acellera Labs, Doctor Trueta 183, 08005, Barcelona, Spain*
¶*Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig Lluis Companys 23, 08010 Barcelona, Spain*

E-mail: gianni.defabritiis@upf.edu

Abstract

Deep learning has been successfully applied to structure-based protein-ligand affinity prediction, yet the black box nature of these models raises some questions. In a previous study, we presented K_{DEEP}, a convolutional neural network that predicted the binding affinity of a given protein-ligand complex while reaching state-of-the-art performance. However, it was unclear what this model was learning. In this work we present a new application to visualize the contribution of each input atom to the prediction made by the convolutional neural network, aiding in the interpretability of such predictions. The results suggest that K_{DEEP} is able to learn meaningful chemistry signals from the data, but it has also exposed the inaccuracies of the current model, serving as a guideline for further optimization of our prediction tools.

Introduction

Machine-learning methods have been widely applied in the field of chemoinformatics, ranging from simple, regressor-based QSAR models[1–4] to more complex neural networks. These latter methods have been reported to increase performance in some critical tasks for drug discovery, such as toxicity assessment,[5,6] pharmacokinetics, physicochemical property prediction,[7–10] and protein-ligand binding affinity prediction[11–15]

In a previous work, we developed K_{DEEP} - a 3D convolutional neural network (CNN) that accepts as input a voxelized representation of a protein-ligand complex and outputs a prediction of binding affinity with state-of-the-art accuracy.[11] However, it was unclear whether K_{DEEP} was learning meaningful chemistry or just exploiting shortcuts such as the positive relationship between molecular weight and affinity.[16] Learning these shortcuts instead of the underlying nature of the problem is a topic of concern in the field.[17] It is then comprehensible for many machine learning methods to spark criticism regarding the difficulty to understand the rationale

behind their predictions. It has been questioned whether a pharmaceutical company would promote a given molecule into a portfolio based only on an opaque prediction made by a neural network, without any clear explanation to support it.[18] Providing such explanation would undoubtedly increase the value, trustworthiness and usability of machine learning models in drug discovery.

Recently, advances in model interpretability,[19,20] as well as the availability of software libraries such as Captum[21] and Alibi,[22] have allowed researchers to get a first glimpse of what features of the input are more influential towards predictions made by neural networks (*i.e.* feature attribution assignment). One natural approach to measure this influence is to look at the gradients of the output neuron with respect to the input. In fact, in a CNN trained to discriminate accurate from inaccurate binding poses and to predict binding affinity, visually inspecting these gradients can reveal in which direction the atoms should move to improve the score that the network assigns it,[23] providing some degree of interpretability.

However, backpropagating the prediction relative to the input layer can produce very low gradients in the vicinity of the input vector,[19] a process which is known as "gradient saturation". The Integrated Gradients (IG) feature attribution technique[19] helps to mitigate this problem, providing a better measure of how each input feature influences the prediction. Instead of evaluating the gradients at one particular input value (the image in a traditional 2D-CNN), gradients are computed for several variants of that image, ranging from a user-defined baseline (typically, an image with all its pixel-channel values set to zero) to the actual image. In each variant, the values of all its pixels are multiplied by a scalar α, ranging from the zeroed-out input to the original image. At low values of α, the resulting input vector is far from the usual input space the network has been exposed to during training, circumventing the gradient saturation issue.

In this article, we present an application to visualize the contribution of the input features for the prediction of K_{DEEP} and similar CNNs. In addition to describing the methodology used herein, we also showcase several relevant examples of attributions which match with structural biology knowledge. We analyse the prediction of three distinct models: a clash detector, a docking pose classifier and K_{DEEP}. The clash detector provides a baseline to which we compare the other models and allowed us to validate the implementation of this application. The docking pose classifier and K_{DEEP} models were evaluated to see if CNNs trained to perform chemically relevant tasks were learning meaningful chemistry. The application, called Glimpse, is available to use at `https://www.playmolecule.org/Glimpse/`.

Methods

Model Training

K_{DEEP} is a 3D CNN which accepts as input a grid of size $24\mathring{A}^3$. This grid is generated by mapping the atom positions of the ligand and its surrounding protein residues to the corresponding voxel and channel in the grid. K_{DEEP} uses 8 different channels: hydrophobic, aromatic, hydrogen bond donor, hydrogen bond acceptor, positive ionizable, negative ionizable, metals and excluded volume (occupancy) for both protein and ligand. This gives a total of 16 channels and a grid of dimensions $16 \times 24\mathring{A}^3$ (see Jiménez *et al.*[11,24] for more details). The network was trained on the latest version of the refined set of PDBbind ,[25] achieving a Pearson's correlation coefficient of 0.79 in the test set. Details on the training and evaluation of the different models can be found in SI.

We also trained a clash detector. The objective behind it is two-fold. First, there is a clear expectation in terms of what attributions should look

like: clashing regions or close contacts would appear highlighted, while the remaining voxels would be of little importance. We were able to validate the implementation of Glimpse by checking if the computed attributions matched this expectation. Second, the computed attributions of this simple model served as a reference point to which we compare the other models, both visually and quantitatively. This model was trained to discriminate regular protein-ligand poses from clashed poses and achieved 0.97 classification accuracy and 0.98 precision in a held-out validation set.

We trained a third model — a docking pose classifier — for two reasons: i) it is a challenging task, comparable to that of predicting binding affinity; ii) there is much more data available from which the model can learn. This model was trained on a large set of good (RMSD < 1 Å) and bad poses (RMSD > 3 Å), showing an accuracy and precision of 0.94 and 0.83 on the validation set.

It must be noted that the performance of these three models was evaluated on a random test and validation sets. In some cases protein-ligand complexes in these sets might be similar to those in the training set, either in terms of protein structure or ligand composition. This yields overoptimistic results. In fact, when trained and evaluated in more strict splits which ensured sequence and ligand dissimilarity, K_{DEEP} performance ranged from r=0.09 to r=0.7 (see SI for details).

The models analysed in this work have been uploaded to Glimpse with the names: "K_{DEEP}", "Pose classifier" and "Clash detector". Attributions for these models can be computed and visualized in the app for any valid protein-ligand complex.

Implementation

Integrated Gradients

The IG method works by computing gradients of the output neuron with respect to the input layer along an interpolated path from a given input baseline (x'_i) to the original input (x_i) taking α infinitesimal steps as in:[19]

$$\mathrm{IG}_i(x) ::= (x_i - x'_i) \int_{\alpha=0}^{1} \frac{\partial F\left(x' + \alpha\left(x - x'\right)\right)}{\partial x_i}\, d\alpha, \tag{1}$$

where F denotes the forward pass of the neural network.

This effectively circumvents the issue of low gradients (gradient saturation) in the vicinity of the input by averaging the gradients along a range of different input values. Gradient saturation can occur if a given input value leads to a neuron being activated in a region of the activation function which is very flat, for instance, the extremes of a sigmoid. Hence, using the gradients of the prediction with respect to a given input could assign an importance of zero to it, regardless of its real importance.

The computed attributions provide a value for each voxel representing their importance towards the prediction. Glimpse uses the IG implementation from the Captum library.[21] Attributions are computed by approximating the integral as a series of discrete steps along the interpolated path from the selected baseline (an input vector in which all voxels are set to zero) to the evaluated, voxel map corresponding to the original protein-ligand complex. In this implementation, we used 100 steps, as it was shown to offer a good balance between computational expense and attribution quality.

Graphical User Interface

Glimpse provides a web-based graphical user interface (GUI) that helps to trace attributions to voxel maps. An overview of the GUI is provided in Figure 1. The computed attributions for the input channels are displayed as mesh isosurfaces whose isovalue can be tuned with a slider. A detailed description of the input channels is provided in the Model Training section of this manuscript.

To offer a summary of the results, the interface displays by default only the most contributing regions. These regions are the result of identifying, for each of the 16 channels, the voxel with the highest absolute attribution value and the neighbours around it. In this summary, only the channels containing the best 5 voxels are displayed, for simplicity. Additionally, the user can display the attributions for the different input channels individually, and the raw attribution maps can be downloaded as a Gaussian cube format file (.cube) and explored in VMD[26] or other molecular visualization software.

Usage

Glimpse requires a protein-ligand complex structure, either experimentally determined or predicted by a docking software. The protein must be correctly protonated and provided as a .PDB file. PlayMolecule platform offers proteinPrepare[27] to protonate the protein. The ligands have to be provided as a valid .SDF file. If needed, Glimpse provides an option for protonation of these ligands. Only 100 ligands are allowed per job. Finally, users can select which model to use from a list, which by default is "K_{DEEP}". In terms of time, evaluating 10 protein-ligand complexes takes around 150 s. When inspecting the attributions, one would typically start looking at the visual summary, followed by an inspection of individual channels. It is worth paying particular attention to the voxels with the highest and lowest attribution values and checking if the nearby atoms are involved in an interaction. The occupancy channels offer a good overview of the whole picture and constitute a good place to start.

Analysis

For each model, we visually inspected the attributions computed for several protein-ligand complexes to evaluate how well they match with structural biology knowledge. We focused on interactions known to contribute towards binding free energy, e.g. hydrogen bonds, π-stacking. Another aspect we inspected was the reciprocity in the attributions, that is, whether the two parties involved in the protein-ligand interaction are reflected in their attribution values.

While visual inspection can provide valuable insights, it can also be misleading and prone to unintended biases. Therefore, we designed a quantitative analysis, in which we computed the IG for all the crystal structures (not clashed or docked) coming from PDBbind database. For each of the 16 channels, the voxel with the highest, absolute value of the IG values was identified. Then, we measured the distance between the top contributing voxels in the complementary channels. We evaluated the following pairs of channels: protein hydrophobic and ligand hydrophobic, protein aromatic and ligand aromatic (π-stacking), protein acceptor and ligand donor, protein donor and ligand acceptor (hydrogen bonds), protein occupancy and ligand occupancy (steric component). As a baseline, we took the distance between two randomly selected voxels from the appropriate channels, whose occupancy value was over 0.75, ensuring that an atom was nearby the voxel.

Results

As a leading example for the analysis we selected a complex of a molecular chaperone, heat shock protein 90 kDa (Hsp90) sourced from PDB (PDBid: 3D0B).[28] This well studied oncology target has been a subject of numerous structure based virtual screening campaigns.[29,30] In the analyzed example HSP90 forms a potent complex with an analogue of benzamide tetrahydro-4H-carbazol-4-one (SNX), with an affinity of 290 nM. Additionally, the complex possesses few features that facilitate tight binding, mainly π-stacking and a hydrogen

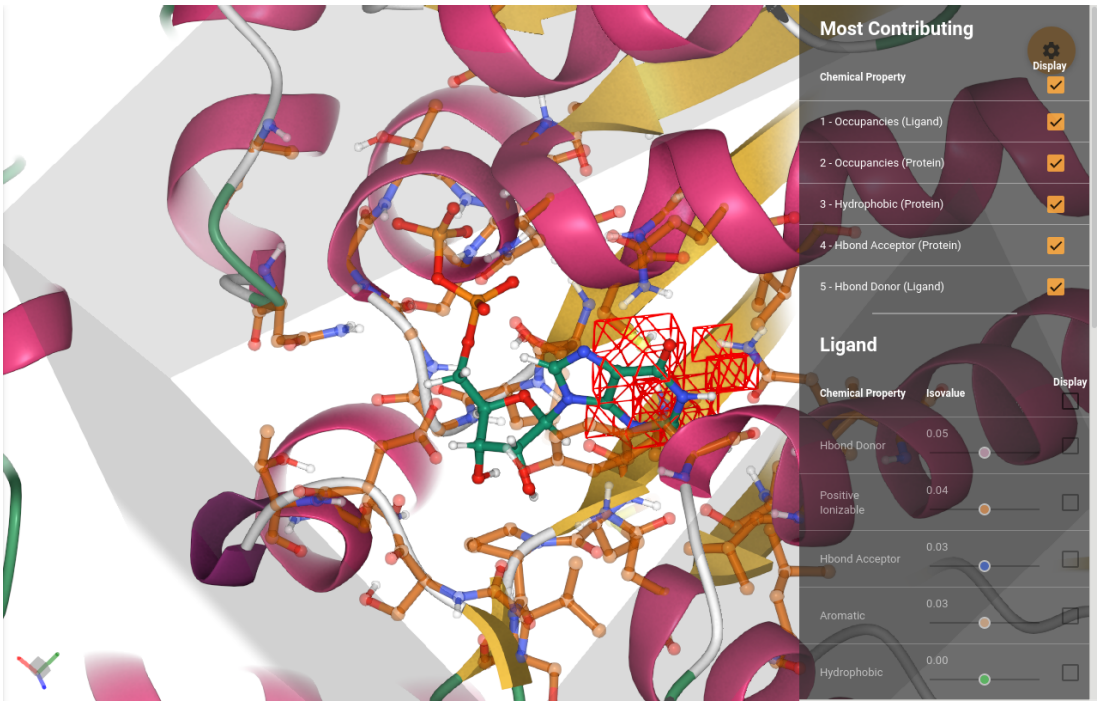


Figure 1: Main view of the graphical user interface. The protein-ligand complex is displayed, with the attributions of the most contributing voxels superimposed. The attributions for the different channels can be seen individually using the corresponding sliders in the menu on the right, which display isosurfaces at different isovalues. The full protein is shown in a cartoon representation, while residues in the binding site (defined by being within 5Å to the ligand) are shown in a transparent ball-stick representation (only heavy atoms and polar hydrogens). The all-atom representation of the ligand is shown in a bold ball-stick. The region of space seen by the model (voxelization cube) is delimited by a transparent, grey box.

bond with a conserved aspartate (D93) - a very frequent interaction among HSP90 inhibitors.[31] This helped us to relate the predictions to structural features of the complex.

Clash detector

As a sanity check, we started by evaluating the attributions of the simplest model, the clash detector. The visual inspection of multiple complexes revealed that, in all inspected cases, clashing regions or close contacts were highlighted while residues far apart from the ligand remained ignored. The clashes were clearly indicated by occupancy channels of both protein and ligand, showing reciprocity (Fig.2.1A). An example of HSP90 with a clashed pose clearly highlights clashed region between the ligand and a leucine (L92) in the pocket (Fig.2.1B). This confirmed that the protocol was working correctly and gave us a baseline for the analysis of the following models.

Docking pose classifier

The next model was trained to discriminate good and bad docking poses. The true binding mode depends on an enthalpic factor which is determined by formation of strong an stable interactions between the ligand and the protein, like hydrogen bonds or π-stacking.[32] Hence, we expect models

trained to perform such tasks to have learned these interactions.

We found several examples where the attributions correctly matched these expectations. Figure 2.2B shows a hydrogen bond being highlighted in the appropriate channels: hydrogen bond acceptor in the protein and donor in the ligand. An amide moiety in the ligand is establishing a hydrogen bond with the nearby aspartate (D93). It is indeed a key interaction frequently featured in HSP90 inhibitors. In addition to this hydrogen bond, a π-stacking interaction takes place between the aromatic ring of the ligand and a phenilalanine (F123) in the protein, as highlighted by the attributions for the occupancy (Figure 2.2A) and aromatic channels (not shown) of both the protein and the ligand.

After inspecting several examples, we saw that, as in the clash detector model, residues far from the ligand were ignored for the most part (Figure S7 shows one exception) and reciprocity between the ligand and the protein atoms was observed in the majority of inspected cases. However, the overall picture was less clear than in the clash detector model as can be seen by comparing the global view of the three models (sections 1A, 2A and 3A of Figure 2). While in the clash detector model high attributions are well focused on the clashing regions and close contacts, the other two models exhibit a more disperse view.

K_{DEEP}

Similarly to the previous model, K_{DEEP} is expected to predict binding affinity by detecting and correctly weighting the molecular interactions between protein and ligand. In the majority of inspected cases we saw reciprocity between the ligand and the protein atoms that formed interactions (Fig.2.3A), while protein residues distant from the ligand remained ignored (Figure S8 shows one exception). The overall attribution maps seem to be more disperse than for docking pose predictor.

For the example of HSP90, the predicted affinity value was 75 nM, reasonably close to the experimental value of 290 nM, making it a suitable example for attribution analysis. As in the case of pose classifier, the network correctly identified the key hydrogen bond with D93 (Fig.2.3B), as well as π-stacking between the aromatic ring of the ligand and F123 (Fig.2.3A). In this case, however, only phenylalanine ring is highlighted. Nonetheless, the attributions of the aromatic channels highlight the aromatic residues in close proximity to the ring system of the ligand, including that Phe, ignoring all the other aromatic residues in the box.

Quantitative analysis

The quantitative analysis confirmed, for the most part, the conclusions obtained by the visual inspection. Figure 3 shows the distribution of distances between the top voxels from protein occupancy and ligand occupancy channels. Figures S1 to S6 show the distance distribution for the remaining relevant combinations of channels. We can see that models have learned that ligand and protein atoms close to each other are important, which is exemplified by the different distributions being shifted towards the contact range (< 4 Å). This hints that the networks are learning relevant features of the complex: close contacts in the case of clash detector and interactions for the two remaining models. This is particularly clear for the clash detector, where the distance distribution is radically shifted towards the range under 3 Å. The pose classifier model follows, and shows a similar, shifted distribution, although not as clear as in the clash detector. In these two models, the cloud of points describes a 3-line pattern at 0, 1 and around 1.5 Å, showing that the most contributing voxels were in a very close proximity. In fact, the high number of examples observed at distance 0 Å indicates that, in a large fraction of complexes, the same voxel in the two relevant chan-

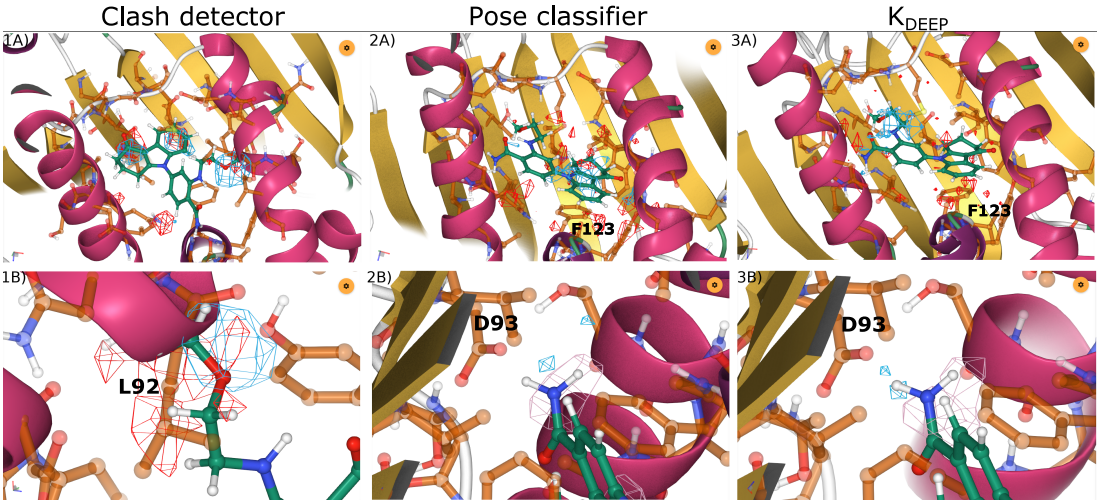


Figure 2: Comparison of computed attributions obtained for a complex of HSP90 with an analogue of benzamide tetrahydro-4H-carbazol-4-one (PDB code: 3D0B) by the three models: Clash detector (1A and 1B), Pose classifier (2A and 2B) and K_{DEEP} (3A and 3B). Pictures on the top row show the attributions for the protein and ligand occupancy channels, in red and blue, respectively. The bottom row focuses on particular interactions. 1B shows a clash between the ligand and the leucine and the attributions for the occupancy channels of protein and ligand (red and blue). 2B and 3B show the hydrogen bond between the benzamide moiety in the ligand and the aspartate (D93) residue in the protein. Attributions for the ligand donor channel are shown in pink, while for the protein acceptor channel are shown in blue.

nels was the most highlighted. The distance distribution for K_{DEEP} is slightly shifted towards higher values, but is still much better than the random baseline. This difference could be related to the fact that, during training, K_{DEEP} is only exposed to crystal poses, in which the distances between ligand and protein atoms should be uniform across examples. This is not obviously the case in the clash detector, nor in the pose classifier, as docking might generate poses which, in some areas, might be slightly too close or too far away from the surface, which might correlate with a bad pose. Hence, these two models would benefit more from paying attention to this low range of values, while K_{DEEP} might not. A similar scenario occurs to the other combinations of channels (Fig.S1-S6): The clash detector is usually the best, followed closely by the pose classifier and K_{DEEP}, which show similar distributions. The distributions of all three models are significantly different from the random baseline in all channel combinations studied according to a two-sided Mann-Whitney U Test (all combinations had a p-value lower than the significance threshold 0.0001). These results provide evidence supporting the hypothesis that K_{DEEP} and the pose classifier models have learned to focus on the interface between the ligand and the protein, as a trained chemist or biologist would do.

After this test, we checked if there was any correlation between the magnitude of the attribution of the two best voxels and the distance between them as a means to test if the quality of the explanation (reciprocity between the two parties) was correlated with the magnitude of the attribution. As in the prior experiment, we identified the voxels with the highest attribution values in the same pairs of channels and annotated the distance between them. We then evaluated the correlation be-

tween the sum of the attributions in those two voxels and their distance. For all models and combinations of channels studied, the Pearson's correlation coefficient ranged between -0.48 and -0.15 (Figure S9). Hence, the higher the attribution value of the two voxels, the more likely it is that those two voxels are in proximity.

Furthermore, for the K_{DEEP} model, we also evaluated if there was a correlation between the attribution values and the accuracy of the prediction (measured as an absolute difference between predicted and actual pK_D values). Neither the maximum attribution value across all channels nor the sum of the absolute attribution values correlated with the accuracy of the prediction in the K_{DEEP} model (Pearson's r was -0.05 and -0.02, respectively). There was not any strong correlation either between far away residues being highlighted and prediction accuracy (r=0.05, see SI for details).

Finally, we evaluated how sensitive the attributions were to minor changes in the input, namely: (1) rotations of the complex and (2) slight modifications of the protein-ligand pose. Ideally, the attributions should be consistent across different orientations and pose variants, hence, the same atoms should be highlighted in the different variations. All three models show a greater consistency than the expected by random, both for protein and ligand atoms. Although on average, the same atom was selected just around 2 times out of the 10 input variants (random baseline is close to 1.0), in all three models, in a great fraction of complexes, the same atom is picked more than 4 times, which is not the case in the random baseline (Figures S10-S13). These results show that the attributions tolerate some degree of input variability. Details on these experiments can be found in SI.

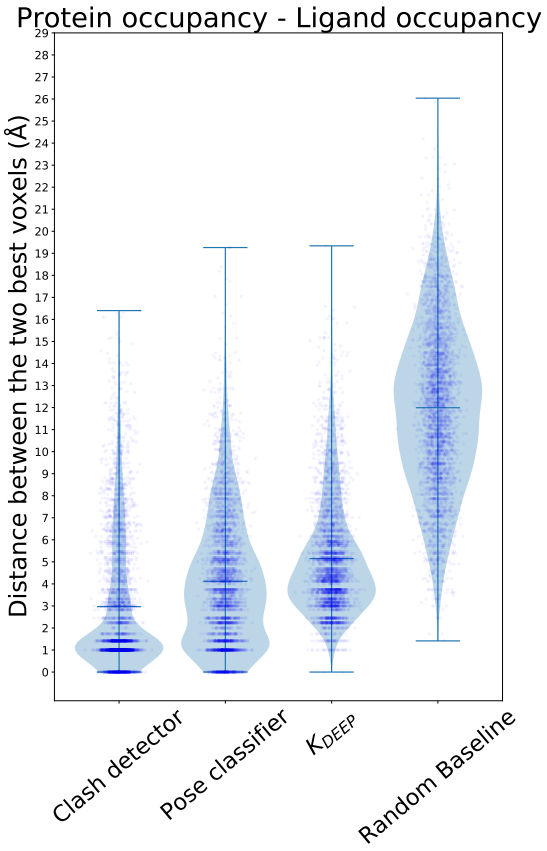


Figure 3: Distance distribution between the two voxels with highest, absolute value in protein and ligand occupancy channels.

Conclusion

The results indicate that the trained networks are able to learn meaningful chemical interactions. However, for the pose classifier model and K_{DEEP}, some cases were observed where the network had ignored strong interactions, highlighted residues far from the ligand (Fig.S7-S8) or highlighted ligand atoms whose interaction counterpart in the protein had low attributions. This can be the result of the difficulty of associating the occurrence of certain contacts or interactions with affinity or with the quality of the pose prediction, leading to shortcut learning. For instance, if all the complexes for

kinases in the training set have a pK_d of 5.0, the network might learn to identify this family by using a set of characteristic residues (which could be far from the ligand) and simply predict 5.0. In this sense, PDBbind is not an ideal training set, as the total number of examples contains few samples for deep-learning standards, pK_d values are distributed in a large range from 2 to 12, scarcely populated in both extremes. In the case of the clash detector model, we have more examples for each class and the two classes belong to very different distributions, and it is very easy to associate the occurrence of a given pattern in the input (a clash) to the correct class, discouraging shortcut learning, which manifests in the attributions of this model being much clear.

Furthermore, voxelization is limited to only 8 properties and excludes crystallographic waters. Given that the latter are known to mediate certain protein-ligand interactions (*e.g.* water bridges), a fraction of the variability in the binding affinity can only be explained by the presence of these waters. Additionally, hydrogen-bond donors and acceptors have diverse strengths (thiol being a weak donor and hydroxyl being a strong one), however, in the featurization they are grouped together in just two entities (donor and acceptor). The same reasoning applies for the positive and negative ionizable channels.

In this study, we have shown that Glimpse displays the capability to expose some of the flaws of the networks herein analysed, suggesting that it can act as an useful diagnostic tool for structure-based 3D-CNN models. We were also able to identify atoms or regions of the protein-ligand complex that play a bigger role on the predictions made by the networks, which is key to improve the usability of CNNs in computational chemistry.

Acknowledgement The authors thank Acellera Ltd. for funding. G.D.F. acknowledges support from MINECO (BIO2017-82628-P) and FEDER (European Regional Development Fund). This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 823712 (CompBioMed2) and from the Industrial Doctorates Plan of the Secretariat of Universities and Research of the Department of Economy and Knowledge of the Generalitat of Catalonia.

Data and Software Availability

Glimpse is available free of charge at `https://www.playmolecule.org/Glimpse/`. The three models studied in this article ("K_{DEEP}", "Pose classifier" and "Clash detector") can be found and used in the web interface. The protein-ligand complexes we used as input for generating the images are available to download in the "Examples" tab in the web interface. The databases used for training and validating the models (BindingMoad[33] and PDBbind[25]) are publicly available, as well as the docking software (rDock[34]).

Supporting Information Available

"Model training" and "Quantitative analysis" contain additional information for these sections. Figures S1 to S6: Distance distribution between the two voxels with highest, absolute attribution value for the different channel combinations studied. Figure S7 and S8: Examples of protein residues far from the ligand having high attribution values. Figures S9 Correlation between magnitude of the attributions of the two best voxels and distance between them. Figures S10 to S13: Attribution consistency distributions.

References

(1) Dudek, A. Z.; Arodz, T.; Gálvez, J. Computational methods in developing quantitative structure-activity relationships (QSAR): a review. *Comb. Chem. High Throughput Screening* **2006**, *9*, 213–228.

(2) Cherkasov, A.; Muratov, E. N.; Fourches, D.; Varnek, A.; Baskin, I. I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y. C.; Todeschini, R., et al. QSAR modeling: where have you been? Where are you going to? *J. Med. Chem.* **2014**, *57*, 4977–5010.

(3) Lo, Y.-C.; Rensi, S. E.; Torng, W.; Altman, R. B. Machine learning in chemoinformatics and drug discovery. *Drug discovery today* **2018**, *23*, 1538–1546.

(4) Neves, B. J.; Braga, R. C.; Melo-Filho, C. C.; Moreira-Filho, J. T.; Muratov, E. N.; Andrade, C. H. QSAR-Based Virtual Screening: Advances and Applications in Drug Discovery. *Front. Pharmacol.* **2018**, *9*, 1275.

(5) Zhang, L.; Zhang, H.; Ai, H.; Hu, H.; Li, S.; Zhao, J.; Liu, H. Applications of Machine Learning Methods in Drug Toxicity Prediction. *Curr. Top. Med. Chem. (Trivandrum, India)* **2018**, *18*, 987–997.

(6) Ma, H.; An, W.; Wang, Y.; Sun, H.; Huang, R.; Huang, J. Deep Graph Learning with Property Augmentation for Predicting Drug-Induced Liver Injury. *Chem. Res. Toxicol.* **2020**,

(7) Montanari, F.; Kuhnke, L.; Ter Laak, A.; Clevert, D.-A. Modeling physico-chemical ADMET endpoints with multitask graph convolutional networks. *Molecules* **2020**, *25*, 44.

(8) Peng, Y.; Lin, Y.; Jing, X.-Y.; Zhang, H.; Huang, Y.; Luo, G. S. Enhanced Graph Isomorphism Network for Molecular ADMET Properties Prediction. *IEEE Access* **2020**, *8*, 168344–168360.

(9) Feinberg, E. N.; Joshi, E.; Pande, V. S.; Cheng, A. C. Improvement in ADMET prediction with multitask deep featurization. *J. Med. Chem.* **2020**, *63*, 8835–8848.

(10) Skalic, M.; Varela-Rial, A.; Jiménez, J.; Martínez-Rosell, G.; De Fabritiis, G. LigVoxel: inpainting binding pockets using 3D-convolutional neural networks. *Bioinformatics* **2019**, *35*, 243–250.

(11) Jiménez, J.; Škalič, M.; Martínez-Rosell, G.; De Fabritiis, G. KDEEP: Protein-Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks. *J. Chem. Inf. Model.* **2018**, *58*, 287–296.

(12) Ragoza, M.; Hochuli, J.; Idrobo, E.; Sunseri, J.; Koes, D. R. Protein-Ligand Scoring with Convolutional Neural Networks. *J. Chem. Inf. Model.* **2017**, *57*, 942–957.

(13) Skalic, M.; Martínez-Rosell, G.; Jiménez, J.; De Fabritiis, G. PlayMolecule BindScope: large scale CNN-based virtual screening on the web. *Bioinformatics* **2019**, *35*, 1237–1238.

(14) Yang, S.; Lee, K. H.; Ryu, S. A comprehensive study on the prediction reliability of graph neural net-

works for virtual screening. *arXiv preprint arXiv:2003.07611* **2020**,

(15) Sakai, M.; Nagayasu, K.; Shibui, N.; Andoh, C.; Takayama, K.; Shirakawa, H.; Kaneko, S. Prediction of pharmacological activities from chemical structures with graph convolutional neural networks. *Sci. Rep.* **2021**, *11*, 1–14.

(16) Sieg, J.; Flachsenberg, F.; Rarey, M. In Need of Bias Control: Evaluating Chemical Data for Machine Learning in Structure-Based Virtual Screening. *J. Chem. Inf. Model.* **2019**, *59*, 947–961.

(17) DeGrave, A. J.; Janizek, J. D.; Lee, S.-I. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nat. Mach. Intell.* **2021**, 1–10.

(18) Vamathevan, J.; Clark, D.; Czodrowski, P.; Dunham, I.; Ferran, E.; Lee, G.; Li, B.; Madabhushi, A.; Shah, P.; Spitzer, M.; Zhao, S. Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.* **2019**, *18*, 463–477.

(19) Sundararajan, M.; Taly, A.; Yan, Q. Axiomatic Attribution for Deep Networks. 2017.

(20) Henderson, R.; Clevert, D.-A.; Montanari, F. Improving Molecular Graph Neural Network Explainability with Orthonormalization and Induced Sparsity. 2021.

(21) Kokhlikyan, N.; Miglani, V.; Martin, M.; Wang, E.; Alsallakh, B.; Reynolds, J.; Melnikov, A.; Kliushkina, N.; Araya, C.; Yan, S.; Reblitz-Richardson, O. Captum: A unified and generic model interpretability library for PyTorch. 2020; Accessed on 2021-06-17.

(22) Klaise, J.; Van Looveren, A.; Vacanti, G.; Coca, A. Alibi: Algorithms for monitoring and explaining machine learning models. 2019; `https://github.com/SeldonIO/alibi`, Accessed on 2021-09-08.

(23) Hochuli, J.; Helbling, A.; Skaist, T.; Ragoza, M.; Koes, D. R. Visualizing convolutional neural network protein-ligand scoring. *J. Mol. Graphics Modell.* **2018**, *84*, 96–108.

(24) Jiménez, J.; Doerr, S.; Martínez-Rosell, G.; Rose, A. S.; De Fabritiis, G. DeepSite: protein-binding site predictor using 3D-convolutional neural networks. *Bioinformatics* **2017**, *33*, 3036–3042.

(25) Liu, Z.; Li, Y.; Han, L.; Li, J.; Liu, J.; Zhao, Z.; Nie, W.; Liu, Y.; Wang, R. PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics* **2015**, *31*, 405–412.

(26) Humphrey, W.; Dalke, A.; Schulten, K. VMD – Visual Molecular Dynamics. *J. Mol. Graphics* **1996**, *14*, 33–38.

(27) Martínez-Rosell, G.; Giorgino, T.; De Fabritiis, G. PlayMolecule ProteinPrepare: a web application for protein preparation for molecular dynamics simulations. *Journal of chemical information and modeling* **2017**, *57*, 1511–1516.

(28) Barta, T. E.; Veal, J. M.; Rice, J. W.; Partridge, J. M.; Fadden, R. P.; Ma, W.; Jenks, M.; Geng, L.; Hanson, G. J.;

Huang, K. H., et al. Discovery of benzamide tetrahydro-4H-carbazol-4-ones as novel small molecule inhibitors of Hsp90. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 3517–3521.

(29) Erlanson, D. A. In *Fragment-Based Drug Discovery and X-Ray Crystallography*; Davies, T. G., Hyvönen, M., Eds.; Springer Berlin Heidelberg: Berlin, Heidelberg, 2012; pp 1–32.

(30) Ruiz-Carmona, S.; Schmidtke, P.; Luque, F. J.; Baker, L.; Matassova, N.; Davis, B.; Roughley, S.; Murray, J.; Hubbard, R.; Barril, X. Dynamic undocking and the quasi-bound state as tools for drug discovery. *Nat. Chem.* **2017**, *9*, 201.

(31) Hoxie, R. S.; Street, T. O. Hsp90 chaperones have an energetic hot-spot for binding inhibitors. *Protein Sci.* **2020**, *29*, 2101–2111.

(32) Majewski, M.; Barril, X. Structural Stability Predicts the Binding Mode of Protein–Ligand Complexes. *J. Chem. Inf. Model.* **2020**, *60*, 1644–1651.

(33) Hu, L.; Benson, M. L.; Smith, R. D.; Lerner, M. G.; Carlson, H. A. Binding MOAD (Mother of All Databases). *Proteins: Struct., Funct., Genet.* **2005**, *60*, 333–340.

(34) Ruiz-Carmona, S.; Alvarez-Garcia, D.; Foloppe, N.; Garmendia-Doval, A. B.; Juhos, S.; Schmidtke, P.; Barril, X.; Hubbard, R. E.; Morley, S. D. rDock: A Fast, Versatile and Open Source Program for Docking Ligands to Proteins and Nucleic Acids. *PLoS Comput. Biol.* **2014**, *10*, e1003571.

Supplementary Information
PlayMolecule Glimpse: Understanding protein-ligand property predictions with interpretable neural networks

Alejandro Varela-Rial,[†,‡] Iain Maryanow,[‡] Maciej Majewski,[†] Stefan Doerr,[‡] Nikolai Schapin,[†,‡] José Jiménez-Luna,[†] and Gianni De Fabritiis[*,†,‡,¶]

†*Computational Science Laboratory, Universitat Pompeu Fabra, Barcelona Biomedical Research Park (PRBB), Carrer Dr. Aiguader 88, 08003, Barcelona, Spain*

‡*Acellera Labs, Doctor Trueta 183, 08005, Barcelona, Spain*

¶*Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig Lluis Companys 23, 08010 Barcelona, Spain*

E-mail: gianni.defabritiis@upf.edu

Model Training

K_{DEEP}

K_{DEEP} was trained on the latest version of the refined set of PDBbind[1] comprising of 4749 protein-ligand complexes after filtering of duplicates and complexes that failed the preparation. A validation set was created for hyperparameter tuning and early stopping by taking a random sample of 10% of the codes in the refined set. As a test set, the core set of PDBbind was used, which is comprised of 272 complexes.

For all three models, during training and validation, the protein-ligand complex was rotated around the geometrical center of the ligand before generating the final grid, in order to augment the existing training set and compensate for the fact that CNNs are not rotationally invariant.[2]

Clash detector

To train the clash detector model, we used the complexes available in the refined set of the 2019 version of PDBbind. The clashed poses were artificially generated by randomly rotating the ligand on its own geometrical center while ensuring that at least one atom in the ligand was within a distance of 1.5Å to the protein. We used the same architecture as for K_{DEEP}, and the binary cross-entropy loss function.[3] We trained the model for 50 epochs with a batch size of 32 and a starting learning rate of 10^{-3}. This model achieved 0.97 classification accuracy and 0.98 precision in a held-out validation set, constituted by a randomly selected group of 10% of the protein-ligand complexes, for which clashed poses were also generated. Both the training and validation sets were constructed in a balanced way, so that half the examples were crystal poses and the other half were clashed poses.

Pose classifier

The pose classifier model was trained on BindingMoad database,[4] which contains $38,702$ protein-ligand complexes. Ten docked poses were generated for each complex using the rDock docking software,[5] which led to more than $310,110$ examples after removing failed jobs. This set was split into two classes, one featuring "good" poses (poses with RMSD below 1 Å) and "bad" poses (RMSD greater than 3 Å). Poses between 1 and 3 Å were discarded, similar to the work by,[2] to create a greater separation between the two distributions and ease the classification task. The final number of examples was $270,225$, constituting a much larger training set than the other two models. This model was trained with the same hyper-parameters, loss function and architecture as the previous one.

A validation set was created, composed by all the good and bad poses generated for a random selection of 10% of the protein-ligand complexes in the BindingMoad database, so that poses for the same protein-ligand complex cannot be found in both training and validation sets. Because most of the poses belonged to the "bad" category, a sampling correction was introduced in the training and a number of bad poses in the validation were removed to reach a 1:1 ratio, reaching a total of $17,478$ examples in the validation set.

Strict split: $\boldsymbol{K}_{\mathrm{DEEP}}$

For K_{DEEP}, we designed a more strict split, where the PDBbind refined set was clustered by sequence similarity using a 70% threshold. The three biggest clusters were selected for testing. A final filter was applied to these three clusters to discard complexes whose ligands had a fingerprint similarity greater than 0.6 with any ligand in any other cluster ensuring that these test sets were different both in terms of protein sequence and ligand composition from any other cluster. Finally, three different K_{DEEP} models were trained using one of the three clusters as test set (leaving one cluster out and training in all the others). Pearson's correlation coefficient in these three test sets was 0.70 (N=29), 0.28 (N=152) and 0.09 (N=81). Hence, predictive performance is lower than on the less strict split and it might be family-dependent.

Quantitative analysis

Correlation between far away residues and accuracy

We tried to measure if any correlation existed between the presence of far protein residues being highlighted and prediction accuracy. We summed the attributions of all protein channels for the voxels that were further than 8 Åfrom any ligand atom (bad attributions) and divided it by the sum of all protein attributions, obtaining a percentage of the attributions falling far from the ligand. The Pearson's correlation with the prediction error was just 0.05 meaning that the presence of far away residues being highlighted does not correlate well with

prediction accuracy.

Attribution consistency across rotations and pose variations

In order to check how sensitive attributions were to changes in the protein-ligand complex's orientation, attributions were computed for 10 different orientations for each system. Then the ligand and protein atoms closest to the voxel with the highest attribution in the occupancy channels were identified in each rotation. This allowed us to evaluate how consistent was the selection in comparison to a random baseline, where the ligand and protein atoms were selected randomly among those inside the $24\mathring{A}^3$ box. A similar experiment was performed to evaluate attribution's sensitivity to changes in the pose. Here, instead of 10 rotations, we applied minor rotations (up to 12^{o}) and displacements (sampled from a normal distribution with mean 0 and std 0.2Å) to the ligand alone, leading to 10 small variants of the same pose. Due to the computational cost of these experiments, both of them were performed in a random subset of 300 complexes sampled from PDBbind refined set.

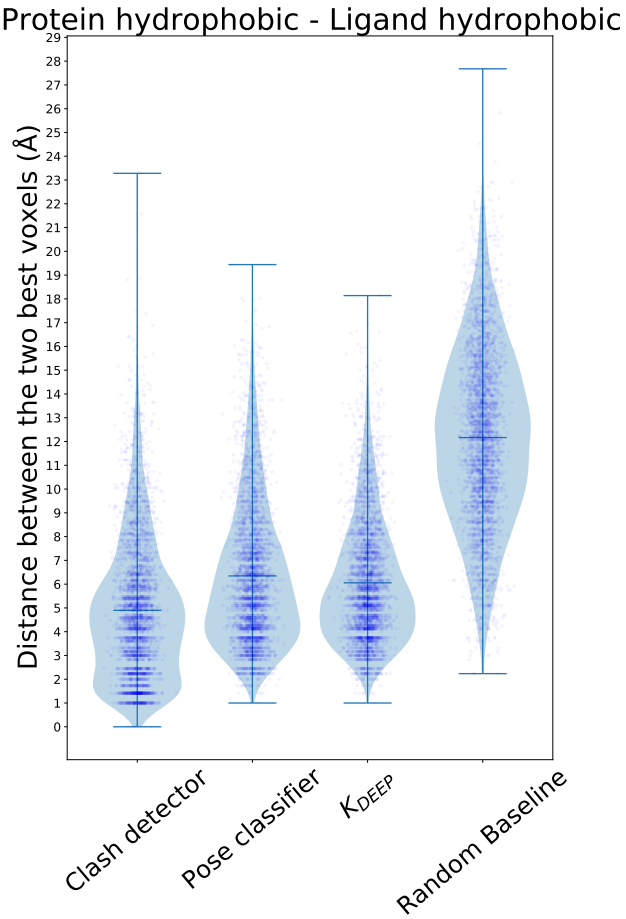


Figure S1: Distance distribution between the two voxels with highest, absolute attribution value in protein hydrophobic and ligand hydrophobic channels.

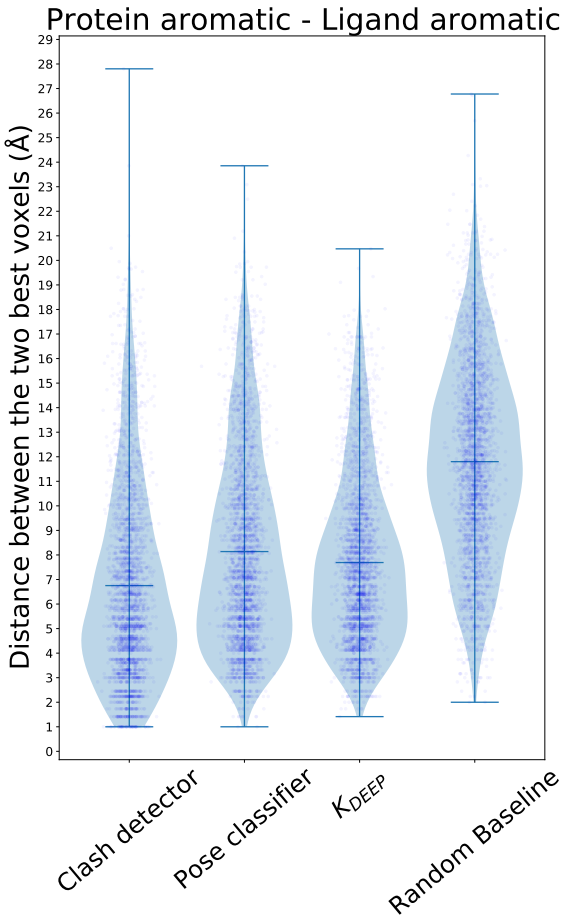


Figure S2: Distance distribution between the two voxels with highest, absolute attribution value in protein aromatic and ligand aromatic channels.

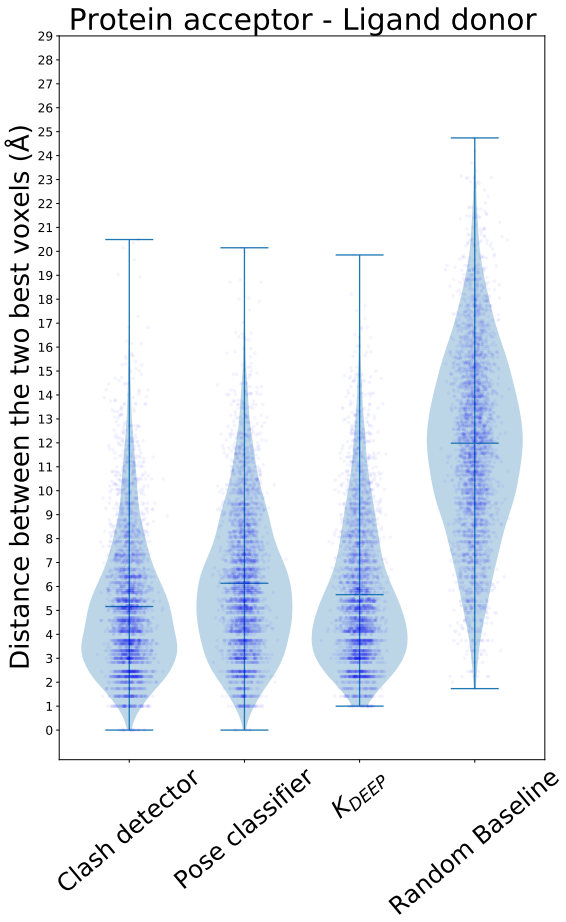


Figure S3: Distance distribution between the two voxels with highest, absolute attribution value in protein acceptor and ligand donor channels.

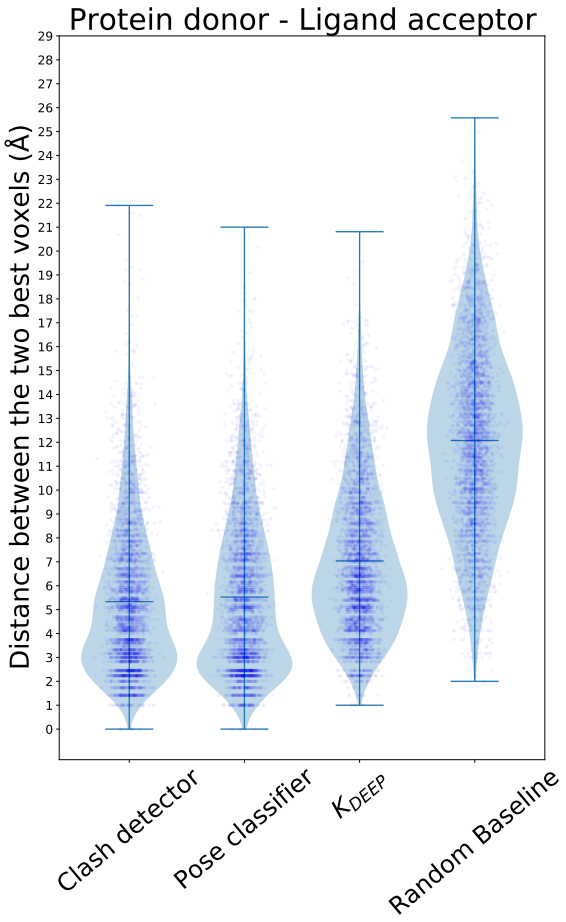


Figure S4: Distance distribution between the two voxels with highest, absolute attribution value in protein donor and ligand acceptor channels.

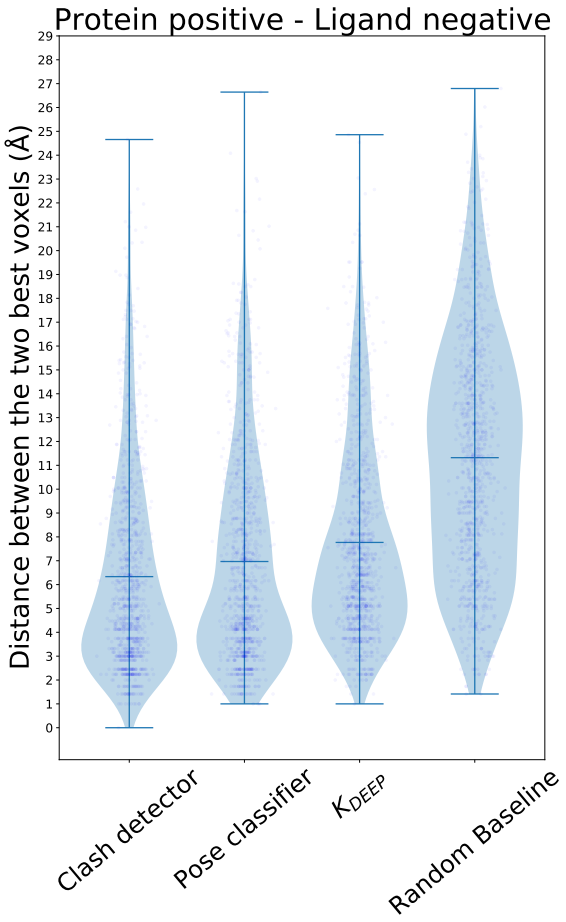


Figure S5: Distance distribution between the two voxels with highest, absolute attribution value in protein positive and ligand negative channels.

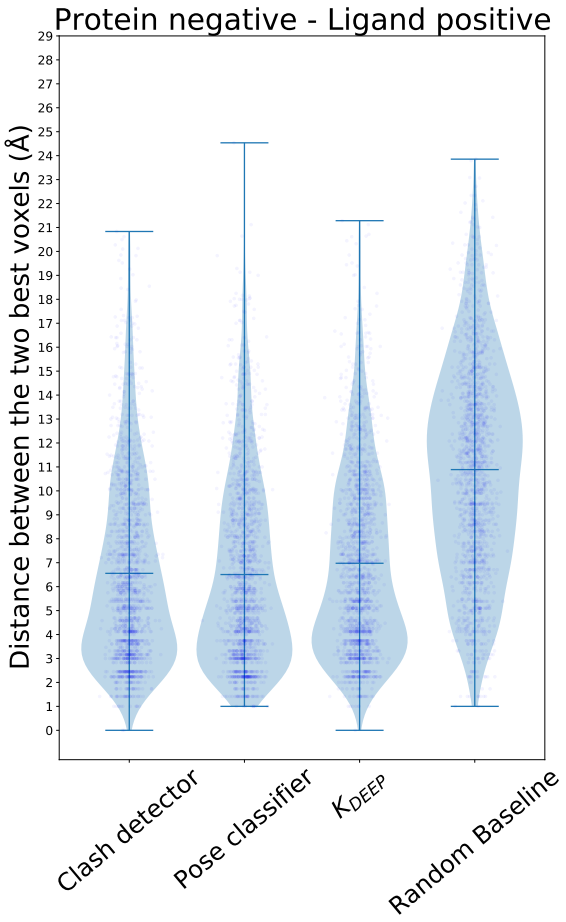


Figure S6: Distance distribution between the two voxels with highest, absolute attribution value in protein negative and ligand positive channels.

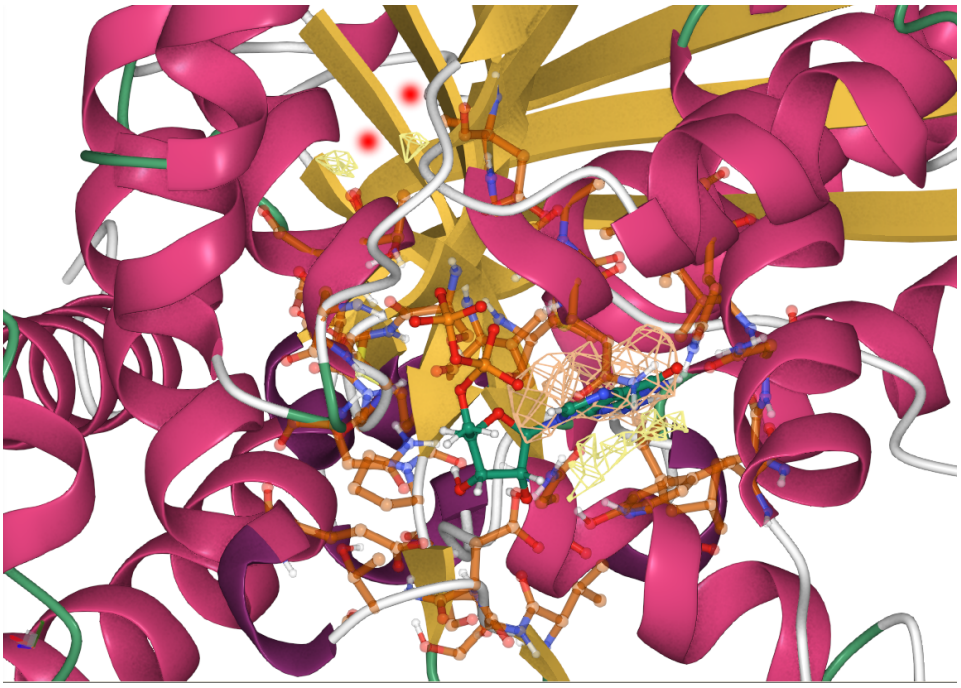


Figure S7: Attributions computed for the pose classifier model for the protein and ligand aromatic channels (yellow and brown, respectively). In addition to highlighting the aromatic ring in the Tyr residue engaging in a pi-stacking interaction with the ligand (lower right corner), two other regions in the aromatic protein channel appear highlighted (red dots at the top), despite being far apart from the ligand. A His residue can be found at that location (not shown for clarity) but no interaction with the ligand is possible as a beta sheet sits between the two parties. PDB code: 5JVD

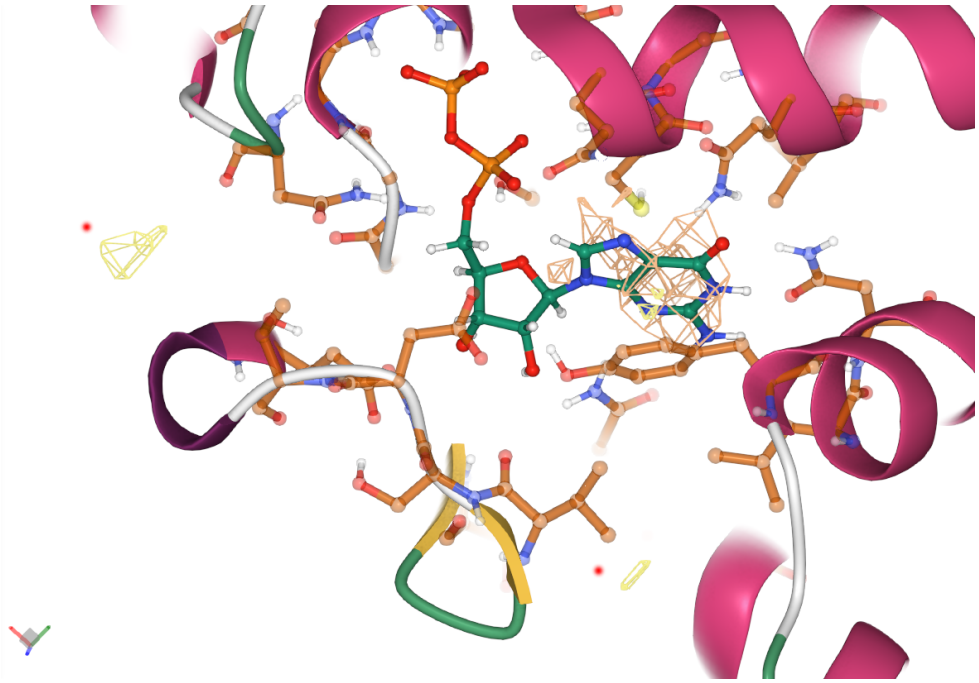


Figure S8: Attributions computed for K_{DEEP} model for the protein and ligand aromatic channels (yellow and brown, respectively). In addition to highlighting the aromatic ring in the Tyr residue engaging in a pi-stacking interaction with the ligand, two extra regions appear highlighted in the protein aromatic channel (identified by two red dots), despite being far apart from the ligand. The aromatic residues present at those locations (not shown for clarity) are a Trp (left red dot) and a Tyr (bottom right dot). Both are far away and shielded from the ligand by other protein residues. PDB code: 5JVD

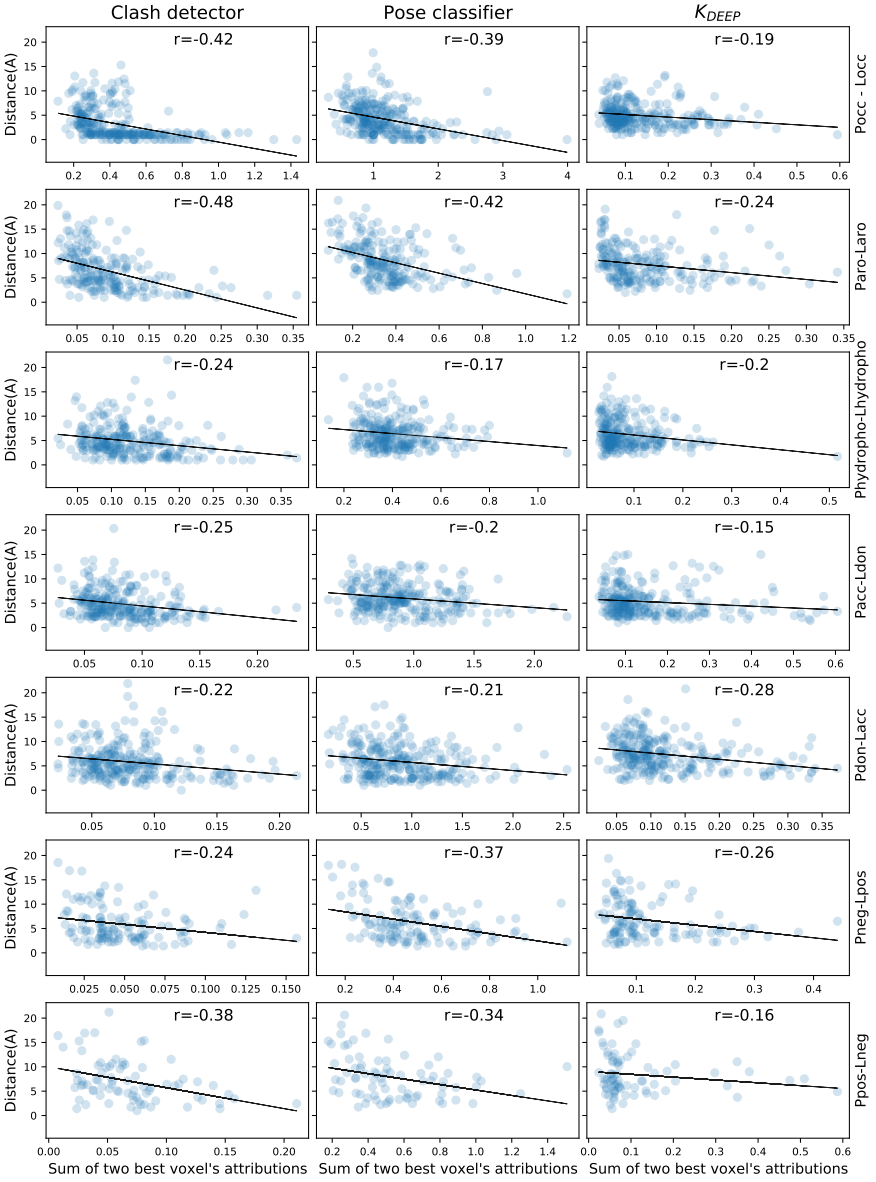


Figure S9: Correlation between the sum of the attributions of the two best voxels and distance between them. Each column represents a model and each row represents a protein-ligand channel combination (P is for protein, L for ligand, occ is occupancy, aro is aromatic, hydropho is hydrophobic, acc is acceptor, don is donor, neg is negative and pos is positive.)

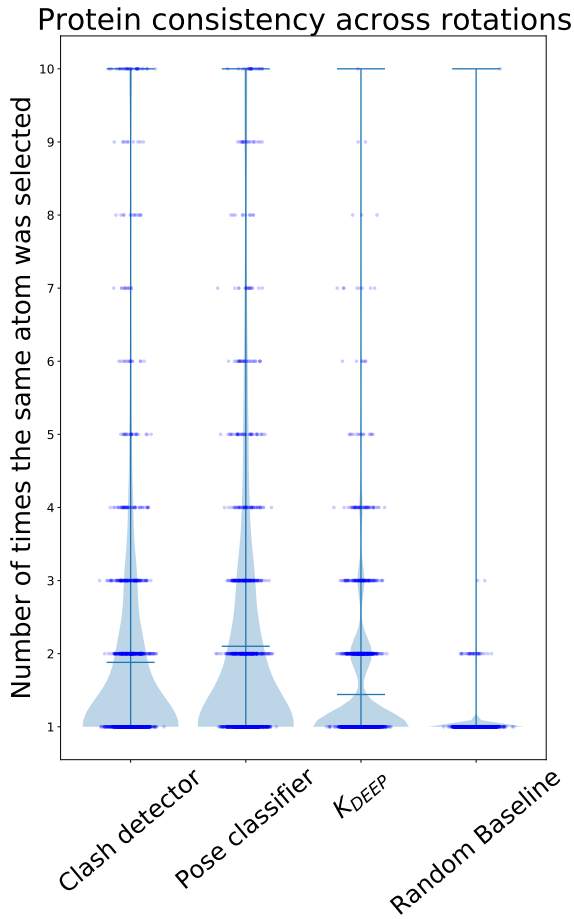


Figure S10: Consistency in protein atom attributions across 10 different orientations. The protein atom closest to the best voxel in the protein occupancy channel is identified in each of the 10 orientations. For each complex, we plot how many times the same atom was selected. As can be seen, in the clash detector and pose classifier models, for some complexes, the exact same protein atom is picked in all 10 rotations. In all three models, the distribution is clearly shifted upwards (more consistent) compared to the random baseline.

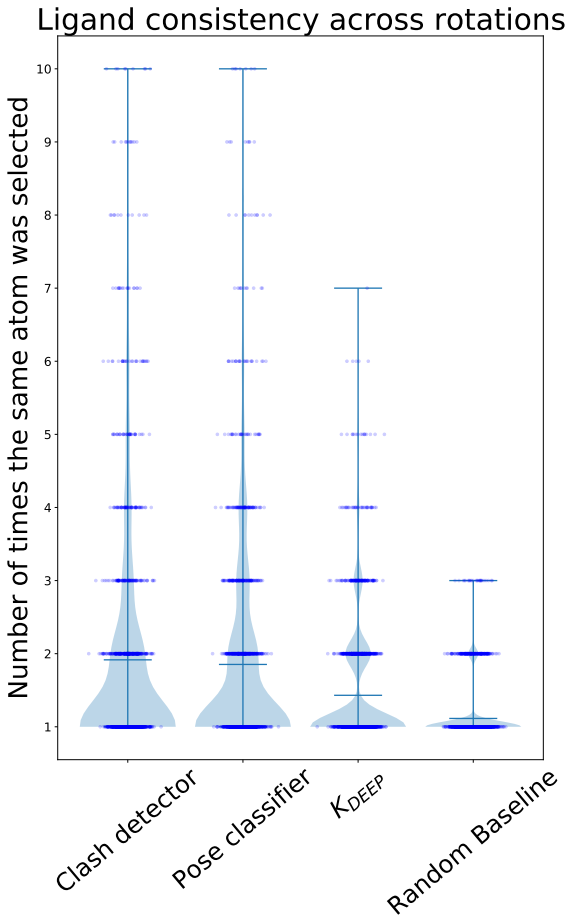


Figure S11: Consistency in ligand atom attributions across 10 different orientations. The ligand atom closest to the best voxel in the ligand occupancy channel is identified in each of the 10 orientations. For each complex, we plot how many times the same ligand atom was selected. All three models show a distribution more shifted towards greater values than the random baseline.

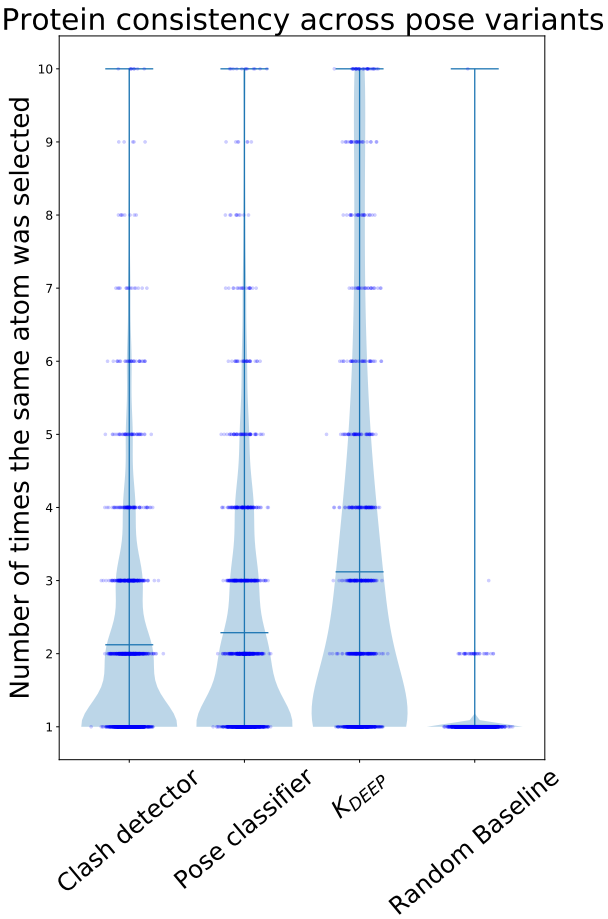


Figure S12: Consistency in protein atom attributions across 10 pose variations. The protein atom closest to the best voxel in the protein occupancy channel is identified in each of the 10 variants. For each complex, we plot how many times the same atom was selected. In K_{DEEP}, the exact same protein atom is select in all 10 pose variants for a large number of complexes.

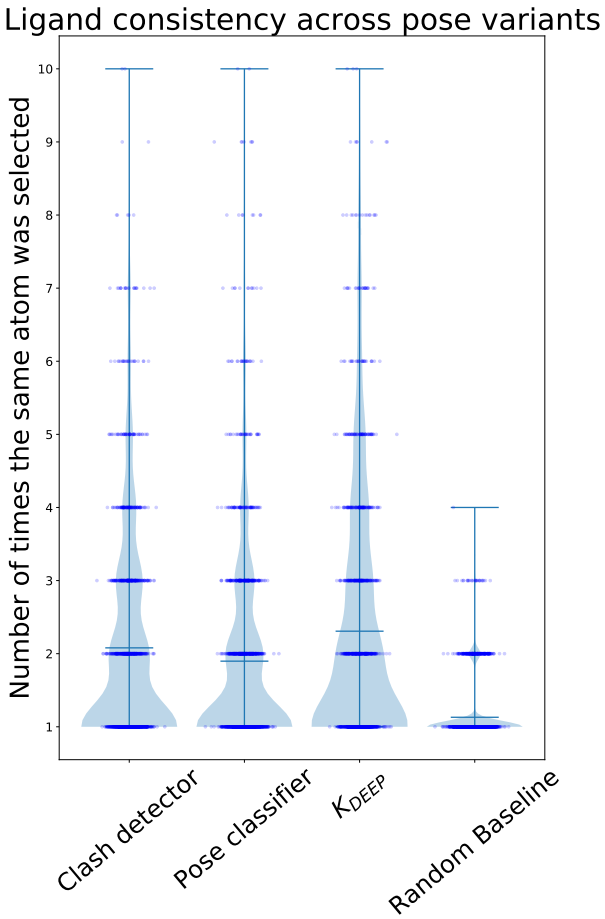


Figure S13: Consistency in ligand atom attributions across 10 pose variations. The ligand atom closest to the best voxel in the ligand occupancy channel is identified in each of the 10 poses. For each complex, we plot how many times the same ligand atom was selected. Although all three models show a distribution better than the baseline, unfortunately, there are few complexes on which the exact same ligand atom is picked in all 10 poses.

References

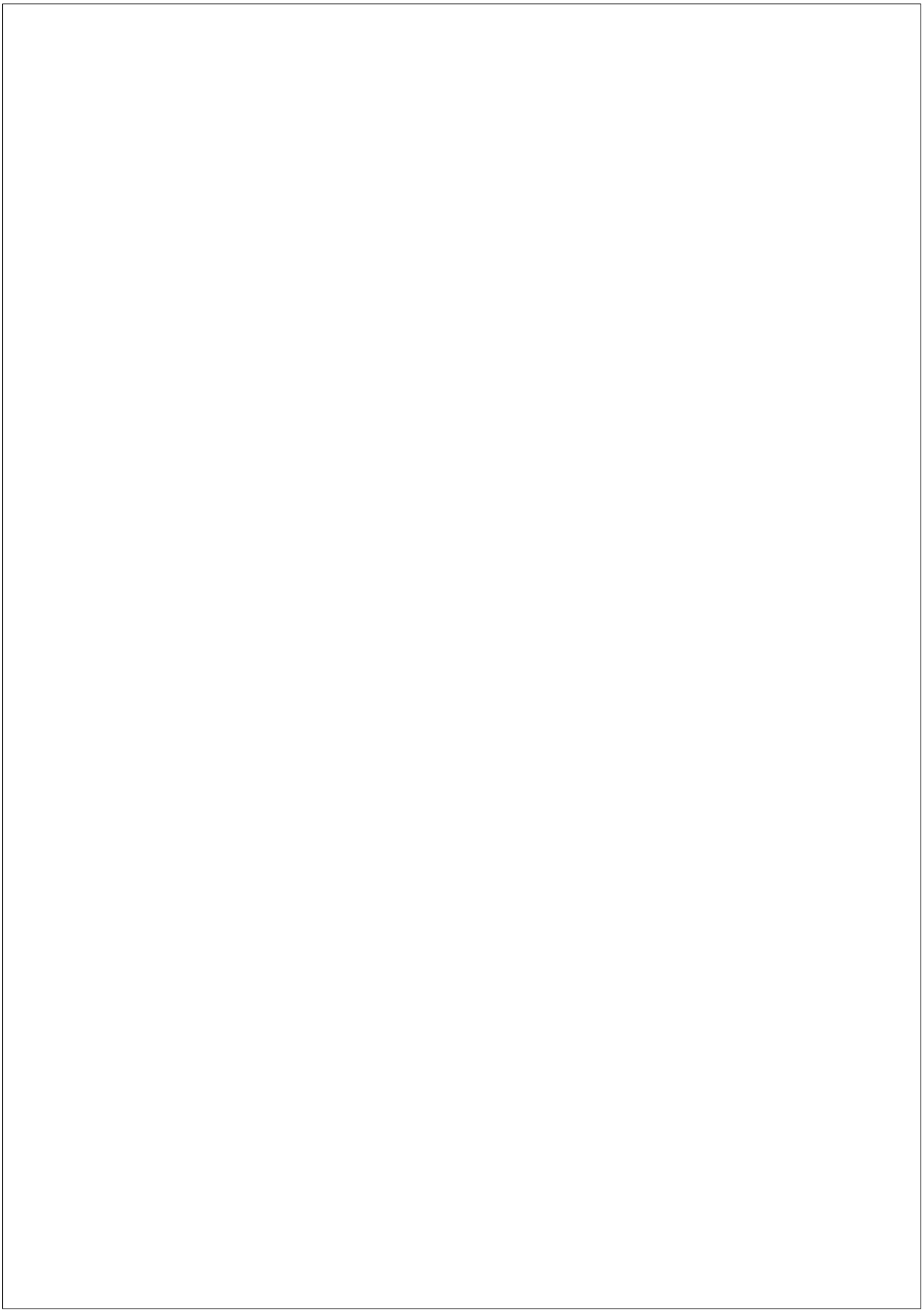
(1) Liu, Z.; Li, Y.; Han, L.; Li, J.; Liu, J.; Zhao, Z.; Nie, W.; Liu, Y.; Wang, R. PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics* **2015**, *31*, 405–412.

(2) Ragoza, M.; Hochuli, J.; Idrobo, E.; Sunseri, J.; Koes, D. R. Protein-Ligand Scoring with Convolutional Neural Networks. *J. Chem. Inf. Model.* **2017**, *57*, 942–957.

(3) BCEWithLogitsLoss — PyTorch 1.8.1 documentation. `https://pytorch.org/docs/stable/generated/torch.nn.BCEWithLogitsLoss.html`.

(4) Hu, L.; Benson, M. L.; Smith, R. D.; Lerner, M. G.; Carlson, H. A. Binding MOAD (Mother of All Databases). *Proteins: Struct., Funct., Genet.* **2005**, *60*, 333–340.

(5) Ruiz-Carmona, S.; Alvarez-Garcia, D.; Foloppe, N.; Garmendia-Doval, A. B.; Juhos, S.; Schmidtke, P.; Barril, X.; Hubbard, R. E.; Morley, S. D. rDock: A Fast, Versatile and Open Source Program for Docking Ligands to Proteins and Nucleic Acids. *PLoS Comput. Biol.* **2014**, *10*, e1003571.



Chapter 4

DISCUSSION

In this thesis, we have presented new methods to predict the binding mode of a ligand to a protein, to validate the capacity of a predicted pharmacofield to recover the right binding mode among several variants and to identify which atoms contributed the most to the prediction made by a convolutional neural network. All these tasks are important in computational chemistry research and the focus of strong research efforts.

4.1 LigVoxel

Previously, we discussed the benefits of using a pharmacophore when scoring compounds, as well as their limitations, namely, the need for, at the very least, one protein-ligand structure. LigVoxel eliminates that limitation by constructing a pharmacofield from the protein structure alone. My contribution to this work was limited to evaluating the capacity of the predicted pharmacofield to distinguish the right pose (RMSD below 2 Å with respect to the crystal pose) among several decoy poses. In all tests where poses of the crystal conformer were included among the set of decoy poses, the correct pose was recovered among the decoys in a great fraction of cases. Importantly, the recovery was still high when the voxelization center was modified by up to 3 Å, which indicates that the autoencoder is tolerant to such modifications, and when poses of non-

crystal conformers were included as decoys. However, when the crystal conformer was not included among the conformers, performance fell to almost zero, although this seems to be due to sampling limitations in the conformer generation, and not the autoencoder.

These results are evidence supporting the capacity of the autoencoder to generate reasonable pharmacofields from the protein structure alone, but also indicate that they could be used to re-score poses and compounds in a virtual screening campaign. In fact, if the voxelization of the poses was fast enough, one could build a docking software where each sampled pose would be scored based on the overlap with the predicted pharamacofield, which needs to be computed only once. Actually, the last experiment of that article, where we only generated poses for non-crystal conformers, can be thought of as a toy example of a docking algorithm.

Furthermore, improvements in machine learning –like generative adversarial networks– and increases in the number of structures to train on might lead to greater discriminatory capacity. Additionally, one benefit of pharmacofields is the ease to interpret the score, as one can simply visually inspect the overlap between the predicted clouds and the atoms in the pose. However, it must be stated that these experiments were all done in protein conformations where a ligand was bound and was simply extracted before performing the pharmacofield prediction. The question remains of how different protein conformations will impact LigVoxel’s prediction.

4.2 SkeleDock

Although several protocols already existed to perform scaffold docking, they were not readily available or had some limitations; for instance, while they aligned the query compound to the template, its conformation was not modeled after it, losing accuracy. Furthermore, the dihedral

autocompletion step proposed in SkeleDock is novel and helpful when modeling congeneric series where minor modifications are introduced in the scaffold, and particularly valuable when modeling macrocycles. An extra benefit of this protocol is that the resulting conformation closely resembles the crystal geometry, which could be key if downstream affinity predictors are expecting such geometries, like K_{DEEP}.

When using structure-based affinity predictors, and particularly when using CNNs, one must keep in mind the geometry distribution they expect. For instance, K_{DEEP} was trained only on crystal poses, hence, applying K_{DEEP} on docked poses where the predicted geometry might belong to a distribution slightly different from that found in crystals, could be detrimental for performance. More generally, scoring functions are made of various distance based functions, and during docking, the search engine produces poses that fall in local and global minima of these functions. Although scoring functions might be overall similar, slight changes in the shape of these functions might lead to different scores, similarly to how an hydrogen bond might be classified as such or not depending on the criteria used by the researcher.

An additional consideration to keep in mind is the difference between pose prediction and affinity prediction. In a congeneric series, SkeleDock will place the common scaffold in practically the exact same conformation for all the compounds in the series, as there is no extra minimization for those atoms; while this might not yield the best poses, it is probably beneficial when scoring, as the geometry of the common scaffold is no longer a source of noise, and only the atoms that are different contribute to the differences in score.

Additionally, SkeleDock was validated in a blind setup, which we believe to be key to improve reproducibility and increase trust in computational approaches. The results of the challenge also point out the necessity to use "the right tool for the job", as all best performing solutions use some variation of template docking. This hints the importance of understanding

the scenario that the practitioner faces: is there any binding affinity data that can be used to validate the predicted binding mode? Are there any crystal poses for a similar compound or fragment? Is there any literature describing the flexibility of the pocket? These and many other questions must be asked by the researcher and taken into account to select the software solution that better fits its situation.

Finally, it must be recognized that SkeleDock is limited to use cases where the structure of a similar compound is known. However, as discussed in the article, the rising popularity of fragment based drug discovery could result in an abundance of seed or template structures, which could be then used by SkeleDock to grow fragments into lead compounds. Overall, we believe that SkeleDock is a valuable tool with a narrow, but well defined, scope.

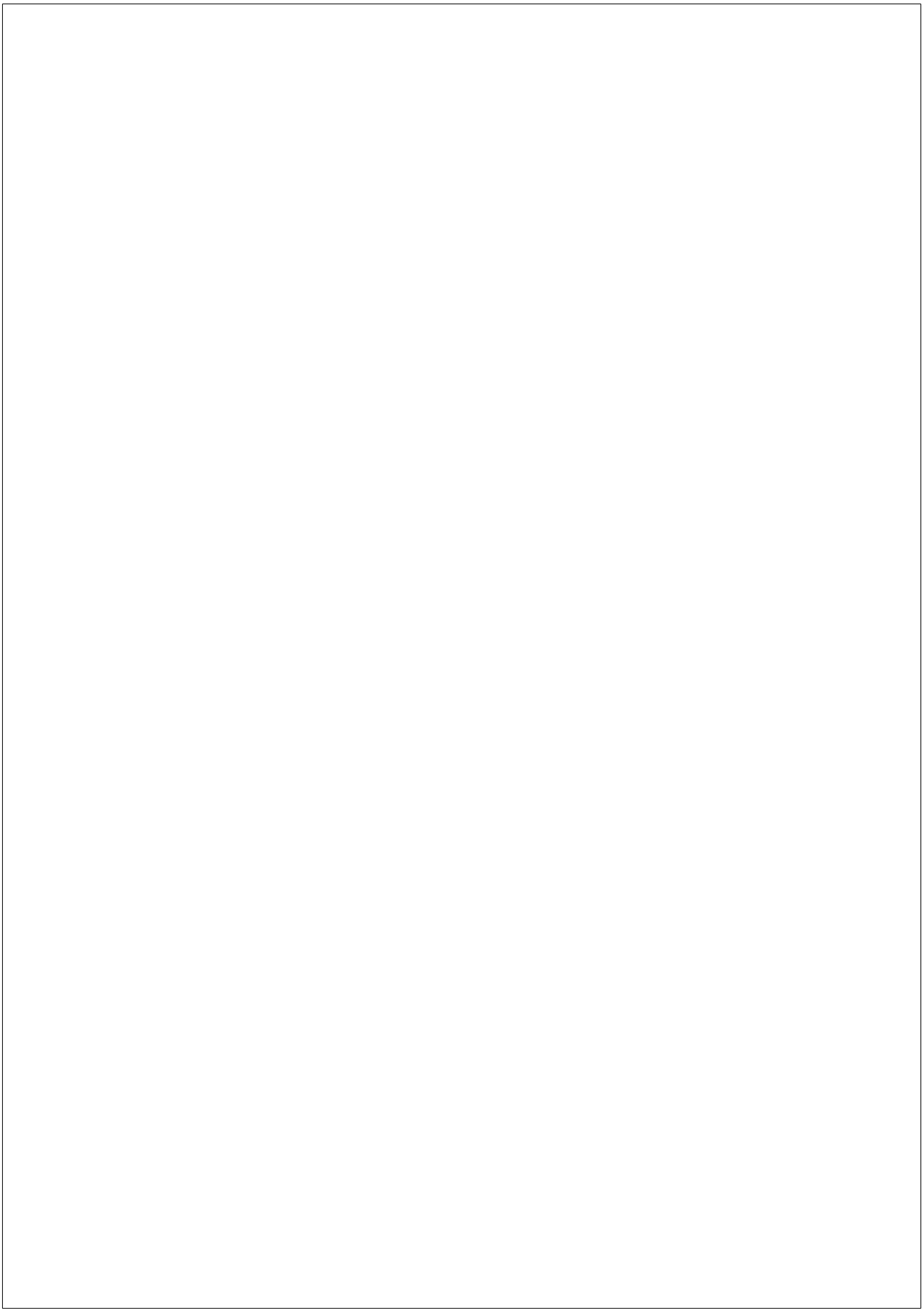
4.3 Glimpse

The adoption of computer vision algorithms, like CNNs, by computational chemistry has inherited both its advantages and its disadvantages: state-of-the-art has improved, but the black-box nature of these algorithms clouds this success, hampering its value, as the opacity of these predictions is in direct conflict with the nature of scientific thinking. Glimpse is an attempt to alleviate this problem by identifying the features in the input (atoms) which influence the prediction the most, improving the interpretability of the predictions. Although this is far from being a perfect explanation of the internal logic on the neural network, it provides a useful explanation of it.

Besides offering this application online, which includes the three models we presented in the article, so that users can extract attributions for their own predictions, we also studied if CNNs were engaging in shortcut learning or not when applied to chemistry problems. We were able to conclude that the interface of the protein-ligand complex accumulated most

of the attributions, which is where the interactions that determine affinity and binding mode occur. Furthermore, we were able to show reciprocity between protein and ligand atoms, further supporting the hypothesis that chemistry principles are learned by CNNs. However, some hints of shortcut learning were also identified, like far-away residues being highlighted. This constitutes a great argument for the use of Glimpse -and other interpretability methods- as diagnostics tools to improve model training.

Finally, it is also worth mentioning the benefits of providing this application, and others presented here, to the public, in a platform like PlayMolecule.org. First and foremost, it increases the value of the software by virtue of making it available to any user with an internet connection and a web browser. Second, it facilitates the evaluation of the reproducibility of the results reported in the article, as any user can potentially use the application and check if he or she obtains the same or comparable results. Third, users provide feedback, report bugs and ask for new features, inducing continuous improvement cycles which, ultimately, improve the software.



Chapter 5

CONCLUSIONS

The algorithms and applications presented here are novel and, when tested against other state-of-the-art methods, they compared favourably. Hence, we believe that they provide value for the scientific community. Importantly, they are publicly available to use at PlayMolecule.org. From the results presented here, we can draw some conclusions:

1. SkeleDock is a valuable application to model the binding mode of a ligand when the structure of a similar compound is available. Its dihedral autocompletion step allows it to overcome minor differences between the scaffold of the template and that of the query compound. It was validated in a blind setup, the D3R Grand Challenge, where it achieved first place in two tasks.

2. The pharmacofields predicted by LigVoxel are capable of discriminating the right pose among several decoy poses, proving that the predicted pharmacofield is accurate and that it has value as a scoring function.

3. Glimpse can identify which regions or atoms of the protein-ligand complex used as input influence the prediction the most, helping

users to understand the rationale behind the prediction they obtained from a neural network like K_{DEEP}. It can also be used as a diagnostics tool to evaluate if a neural network is learning what is expected or not.

4. Convolutional neural networks trained to predict affinity, detect steric clashes or discriminate accurate and inaccurate binding modes learn to focus on the interface between protein and ligand. Computed attributions show reciprocity, as the protein and ligand atoms in contact are both highlighted, supporting the idea that these networks can internally derive structural biology principles.

5. Making these applications available for free online increases their value and facilitates reproducibility evaluations.

Bibliography

[1] Blundell TL. Structure-based drug design. Nature. 1996 11;384:23–26. Available from: https://europepmc.org/article/med/8895597.

[2] Anderson AC. The Process of Structure-Based Drug Design. Chemistry Biology. 2003 9;10:787–797.

[3] Holyoak T. Molecular Recognition: Lock-and-Key, Induced Fit, and Conformational Selection. In: Encyclopedia of Biophysics. Springer Berlin Heidelberg; 2013. p. 1584–1588.

[4] Gurevich EV, Gurevich VV. Therapeutic Potential of Small Molecules and Engineered Proteins. Handbook of Experimental Pharmacology. 2014;219:1–12. Available from: https://link.springer.com/chapter/10.1007/978-3-642-41199-1_1.

[5] Klebe G, Böhm HJ. Energetic and Entropic Factors Determining Binding Affinity in Protein-Ligand Complexes. Journal of Receptor and Signal Transduction Research. 1997 jan;17(42007):459–473. Available from: https://www.tandfonline.com/doi/abs/10.3109/10799899709036621.

[6] Roberts BC, Mancera RL. Ligand - protein docking with water molecules. Journal of Chemical Information and Modeling. 2008 feb;48(2):397–408. Available from: https://pubs.acs.org/doi/abs/10.1021/ci700285e.

[7] Michel J, Tirado-Rives J, Jorgensen WL. Prediction of the water content in protein binding sites. Journal of Physical Chemistry B. 2009 oct;113(40):13337–13346. Available from: https://pubs.acs.org/doi/abs/10.1021/jp9047456.

[8] Chodera JD, Mobley DL. Entropy-Enthalpy Compensation: Role and Ramifications in Biomolecular Ligand Recognition and Design. Annual Review of Biophysics. 2013 5;42:121–142. Available from: https://www.annualreviews.org/doi/abs/10.1146/annurev-biophys-083012-130318.

[9] Du X, Li Y, Xia YL, Ai SM, Liang J, Sang P, et al. Insights into Protein-Ligand Interactions: Mechanisms, Models, and Methods. International Journal of Molecular Sciences 2016, Vol 17, Page 144. 2016 1;17:144. Available from: https://www.mdpi.com/1422-0067/17/2/144.

[10] Stank A, Kokh DB, Fuller JC, Wade RC. Protein Binding Pocket Dynamics. Accounts of Chemical Research. 2016 may;49(5):809–815. Available from: https://pubs.acs.org/doi/abs/10.1021/acs.accounts.5b00516.

[11] Zhou T, Commodore L, Huang WS, Wang Y, Sawyer TK, Shakespeare WC, et al. Structural Analysis of DFG-in and DFG-out Dual Src-Abl Inhibitors Sharing a Common Vinyl Purine Template. Chemical Biology & Drug Design. 2010 jan;75(1):18–28. Available from: http://doi.wiley.com/10.1111/j.1747-0285.2009.00905.x.

[12] Pantsar T, Poso A. Binding Affinity via Docking: Fact and Fiction. Molecules. 2018 jul;23(8):1899. Available from: http://www.mdpi.com/1420-3049/23/8/1899.

[13] Bauer MR, Mackey MD. Electrostatic Complementarity as a Fast and Effective Tool to Optimize Binding and Selectivity of Protein-Ligand Complexes. Journal of Medicinal Chemistry. 2019

3;62:3036–3050. Available from: https://pubs.acs.org/doi/abs/10.1021/acs.jmedchem.8b01925.

[14] Li D, Sun M, Li M, Li Z, Jiang Z. The impact of hardware improvement for molecular modeling in a grid environment. Expert Opinion on Drug Discovery. 2009 8;4:873–877. Available from: https://www.tandfonline.com/doi/abs/10.1517/17460440903061246.

[15] Ewing TJA, Makino S, Skillman AG, Kuntz ID. DOCK 4.0: Search strategies for automated molecular docking of flexible molecule databases. Journal of Computer-Aided Molecular Design. 2001;15(5):411–428. Available from: https://link.springer.com/article/10.1023/A:1011115820450.

[16] Verdonk ML, Cole JC, Hartshorn MJ, Murray CW, Taylor RD. Improved protein-ligand docking using GOLD. Proteins: Structure, Function, and Bioinformatics. 2003 aug;52(4):609–623. Available from: http://doi.wiley.com/10.1002/prot.10465.

[17] Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, et al. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. Journal of Medicinal Chemistry. 2004 mar;47(7):1739–1749. Available from: https://pubs.acs.org/doi/abs/10.1021/jm0306430.

[18] Ruiz-Carmona S, Alvarez-Garcia D, Foloppe N, Garmendia-Doval AB, Juhos S, Schmidtke P, et al. rDock: A Fast, Versatile and Open Source Program for Docking Ligands to Proteins and Nucleic Acids. PLoS Computational Biology. 2014 apr;10(4):e1003571. Available from: http://dx.plos.org/10.1371/journal.pcbi.1003571.

[19] Harvey MJ, Giupponi G, De Fabritiis G. ACEMD: Accelerating biomolecular dynamics in the microsecond time scale. Journal of Chemical Theory and Computation. 2009 jun;5(6):1632–1639. Available from: https://pubs.acs.org/doi/abs/10.1021/ct9000685.

[20] Mestres J. Virtual screening: a real screening complement to high-throughput screening. Biochemical Society Transactions. 2002 8;30:797–799. Available from: https://europepmc.org/article/med/12196200.

[21] Ackloo S, Al-awar R, Amaro RE, Arrowsmith CH, Azevedo H, Batey RA, et al. CACHE (Critical Assessment of Computational Hit-finding Experiments): A public-private partnership benchmarking initiative to enable the development of computational methods for hit-finding. 2021 10;Available from: https://chemrxiv.org/engage/chemrxiv/article-details/6168ba62f718dfc39bdee0db.

[22] Zhu T, Cao S, Su PC, Patel R, Shah D, Chokshi HB, et al.. Hit identification and optimization in virtual screening: Practical recommendations based on a critical literature analysis. American Chemical Society; 2013. Available from: https://pubs.acs.org/doi/abs/10.1021/jm301916b.

[23] Al Olaby RR, Cocquerel L, Zemla A, Saas L, Dubuisson J, Vielmetter J, et al. Identification of a novel drug lead that inhibits HCV infection and cell-to-cell transmission by targeting the HCV E2 glycoprotein. PLoS ONE. 2014 oct;9(10). Available from: https://pubmed.ncbi.nlm.nih.gov/25357246/.

[24] Perryman AL, Yu W, Wang X, Ekins S, Forli S, Li SG, et al. A virtual screen discovers novel, fragment-sized inhibitors of mycobacterium tuberculosis InhA. Journal of Chemical Information and Modeling. 2015 mar;55(3):645–659. Available from: https://pubs.acs.org/doi/10.1021/ci500672v.

[25] Luciani R, Saxena P, Surade S, Santucci M, Venturelli A, Borsari C, et al. Virtual Screening and X-ray Crystallography Identify Non-Substrate Analog Inhibitors of Flavin-Dependent Thymidylate Synthase. Journal of Medicinal Chemistry. 2016 oct;59(19):9269–9275. Available from: https://pubmed.ncbi.nlm.nih.gov/27589670/.

[26] Babaoglu K, Simconov A, Irwin JJ, Nelson ME, Feng B, Thomas CJ, et al. Comprehensive mechanistic analysis of hits from high-throughput and docking screens against β-lactamase. Journal of Medicinal Chemistry. 2008 apr;51(8):2502–2511. Available from: https://pubs.acs.org/doi/abs/10.1021/jm701500e.

[27] Pereira HM, Berdini V, Cleasby A, Garratt RC. Crystal structure of calf spleen purine nucleoside phosphorylase complexed to a novel purine analogue. FEBS Letters. 2007 oct;581(26):5082–5086. Available from: http://doi.wiley.com/10.1016/j.febslet.2007.09.051.

[28] Lyu J, Wang S, Balius TE, Singh I, Levit A, Moroz YS, et al. Ultra-large library docking for discovering new chemotypes. Nature. 2019 feb;566(7743):224–229. Available from: https://www.nature.com/articles/s41586-019-0917-9.

[29] Kairys V, Baranauskiene L, Kazlauskiene M, Matulis D, Kazlauskas E. Binding affinity in drug design: experimental and computational techniques. Expert Opinion on Drug Discovery. 2019 8;14:755–768. Available from: https://www.tandfonline.com/doi/abs/10.1080/17460441.2019.1623202.

[30] Di L, Kerns E, Carter G. Drug-Like Property Concepts in Pharmaceutical Design. Current Pharmaceutical Design. 2009 7;15:2184–2194.

[31] Shen C, Weng G, Zhang X, Leung ELH, Yao X, Pang J, et al. Accuracy or novelty: what can we gain from target-specific machine-learning-based scoring functions in virtual screening? Briefings in Bioinformatics. 2021 9;22. Available from: https://academic.oup.com/bib/article/22/5/bbaa410/6070382.

[32] Santana K, do Nascimento LD, e Lima AL, Damasceno V, Nahum C, Braga RC, et al. Applications of Virtual Screening in Bioprospecting: Facts, Shifts, and Perspectives to Explore the Chemo-Structural Diversity of Natural Products. Frontiers in Chemistry. 2021 4;9:155.

[33] Borrelli KW, Vitalis A, Alcantara R, Guallar V. PELE: Protein energy landscape exploration. A novel Monte Carlo based technique. Journal of Chemical Theory and Computation. 2005;1(6):1304–1311. Available from: https://pubs.acs.org/doi/pdf/10.1021/ct0501811.

[34] Muegge I, Mukherjee P. An overview of molecular fingerprint similarity search in virtual screening. Taylor and Francis Ltd; 2016. Available from: https://www.tandfonline.com/doi/abs/10.1517/17460441.2016.1117070.

[35] Kumar A, Zhang KYJ. Advances in the development of shape similarity methods and their application in drug discovery. Frontiers Media S.A.; 2018. Available from: https://www.frontiersin.org/articles/10.3389/fchem.2018.00315/full.

[36] Wang DD, Zhu M, Yan H. Computationally predicting binding affinity in protein-ligand complexes: free energy-based simulations and machine learning-based scoring functions. Briefings in Bioinformatics. 2021 5;22:1–24. Available from: https://academic.oup.com/bib/article/22/3/bbaa107/5860693.

[37] Maia EHB, Assis LC, de Oliveira TA, da Silva AM, Taranto AG. Structure-Based Virtual Screening: From Classical to Artificial Intelligence. Frontiers in Chemistry. 2020 4;8:343.

[38] Trott O, Olson AJ. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. Journal of Computational Chemistry. 2009 jan;31(2):NA–NA. Available from: http://doi.wiley.com/10.1002/jcc.21334.

[39] Dror RO, Pan AC, Arlow DH, Borhani DW, Maragakis P, Shan Y, et al. Pathway and mechanism of drug binding to G-protein-coupled receptors. Proceedings of the National Academy of Sciences of the United States of America. 2011 aug;108(32):13118–13123. Available from: https://www.pnas.org/content/108/32/13118.

[40] Spitaleri A, Decherchi S, Cavalli A, Rocchia W. Fast Dynamic Docking Guided by Adaptive Electrostatic Bias: The MD-Binding Approach. Journal of Chemical Theory and Computation. 2018 mar;14(3):1727–1736. Available from: https://pubs.acs.org/doi/abs/10.1021/acs.jctc.7b01088.

[41] Aguirre C, Brink TT, Guichou JF, Cala O, Krimm I. Comparing Binding Modes of Analogous Fragments Using NMR in Fragment-Based Drug Design: Application to PRDX5. PLOS ONE. 2014 7;9:e102300. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0102300.

[42] Miller EB, Murphy RB, Sindhikara D, Borrelli KW, Grisewood MJ, Ranalli F, et al. Reliable and Accurate Solution to the Induced Fit Docking Problem for Protein-Ligand Binding. Journal of Chemical Theory and Computation. 2021 4;17:2630–2639. Available from: https://pubs.acs.org/doi/abs/10.1021/acs.jctc.1c00136.

[43] Varela-Rial A, Majewski M, Fabritiis GD. Structure based virtual screening: Fast and slow. Wiley Interdisciplinary Reviews: Computational Molecular Science. 2021;p. e1544. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/wcms.1544.

[44] Huey R, Morris GM, Olson AJ, Goodsell DS. A semiempirical free energy force field with charge-based desolvation. Journal of Computational Chemistry. 2007 apr;28(6):1145–1152. Available from: http://doi.wiley.com/10.1002/jcc.20634.

[45] Verdonk ML, Mortenson PN, Hall RJ, Hartshorn MJ, Murray CW. Protein-ligand docking against non-native protein conformers. Journal of Chemical Information and Modeling. 2008 nov;48(11):2214–2225. Available from: https://pubs.acs.org/doi/abs/10.1021/ci8002254.

[46] Hu X, Balaz S, Shelver WH. A practical approach to docking of zinc metalloproteinase inhibitors. Journal of Molecular Graphics and Modelling. 2004 mar;22(4):293–307.

[47] Lamoree B, Hubbard RE. Current perspectives in fragment-based lead discovery (FBLD). Portland Press Ltd; 2017. Available from: https://doi.org/10.1042/EBC20170028.

[48] Mizutani R, Shimizu Y, Saiga R, Ueno G, Nakamura Y, Takeuchi A, et al. Spatiotemporal development of soaked protein crystal. Scientific Reports 2014 4:1. 2014 7;4:1–7. Available from: https://www.nature.com/articles/srep05731.

[49] Doerr S, De Fabritiis G. On-the-fly learning and sampling of ligand binding by high-throughput molecular simulations. Journal of Chemical Theory and Computation. 2014 may;10(5):2064–2069. Available from: https://pubs.acs.org/doi/abs/10.1021/ct400919u.

[50] Souza PCT, Thallmair S, Conflitti P, Ramírez-Palacios C, Alessandri R, Raniolo S, et al. Proteinâligand binding with the coarse-grained Martini model. Nature Communications. 2020 dec;11(1):1–11. Available from: https://doi.org/10.1038/s41467-020-17437-5.

[51] Shan Y, Kim ET, Eastwood MP, Dror RO, Seeliger MA, Shaw DE. How does a drug molecule find its target binding site? Journal of the American Chemical Society. 2011 6;133:9181–9183. Available from: https://pubs.acs.org/doi/abs/10.1021/ja202726y.

[52] Buch I, Giorgino T, De Fabritiis G. Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. Proceedings of the National Academy of Sciences of the United States of America. 2011 jun;108(25):10184–10189. Available from: https://www.pnas.org/content/108/25/10184.

[53] Zhu J, Lv Y, Han X, Xu D, Han W. Understanding the differences of the ligand binding/unbinding pathways between phosphorylated and non-phosphorylated ARH1 using molecular dynamics simulations. Scientific Reports. 2017 dec;7(1):1–14. Available from: https://www.nature.com/articles/s41598-017-12031-0.

[54] Clark AJ, Tiwary P, Borrelli K, Feng S, Miller EB, Abel R, et al. Prediction of protein–ligand binding poses via a combination of induced fit docking and metadynamics simulations. Journal of chemical theory and computation. 2016;12(6):2990–2998.

[55] Ruiz-Carmona S, Schmidtke P, Luque FJ, Baker L, Matassova N, Davis B, et al. Dynamic undocking and the quasi-bound state as tools for drug discovery. Nature Chemistry. 2017 mar;9(3):201–206. Available from: https://www.nature.com/articles/nchem.2660.

[56] Majewski M, Barril X. Structural Stability Predicts the Binding Mode of Protein-Ligand Complexes. Journal of Chemical Information and Modeling. 2020 mar;60(3):1644–1651. Available from: https://pubs.acs.org/doi/abs/10.1021/acs.jcim.9b01062.

[57] Martinez-Rosell G, Harvey MJ, De Fabritiis G. Molecular-Simulation-Driven Fragment Screening for the Discovery of New CXCL12 Inhibitors. Journal of Chemical Information and Modeling. 2018 mar;58(3):683–691. Available from: https://pubs.acs.org/doi/abs/10.1021/acs.jcim.7b00625.

[58] Ferruz N, Tresadern G, Pineda-Lucena A, De Fabritiis G. Multibody cofactor and substrate molecular recognition in the myo-inositol monophosphatase enzyme. Scientific Reports. 2016 jul;6(1):1–10. Available from: https://www.nature.com/articles/srep30275.

[59] Ferruz N, Doerr S, Vanase-Frawley MA, Zou Y, Chen X, Marr ES, et al. Dopamine D3 receptor antagonist reveals a cryptic pocket in aminergic GPCRs. Scientific Reports. 2018 dec;8(1):1–10. Available from: https://www.nature.com/articles/s41598-018-19345-7.

[60] Gioia D, Bertazzo M, Recanatini M, Masetti M, Cavalli A. Dynamic Docking: A Paradigm Shift in Computational Drug Discovery. Molecules. 2017 nov;22(11):2029. Available from: http://www.mdpi.com/1420-3049/22/11/2029.

[61] Tripathi A, Bankaitis VA. Molecular Docking: From Lock and Key to Combination Lock. Journal of molecular medicine and clinical applications. 2017;2. Available from: https://www.sciforschenonline.org/journals/molecular-biology-medicine/IJMBM-2-106.php.

[62] Ferrari AM, Wei BQ, Costantino L, Shoichet BK. Soft docking and multiple receptor conformations in virtual screening.

Journal of Medicinal Chemistry. 2004 oct;47(21):5076–5084. Available from: https://pubs.acs.org/doi/abs/10.1021/jm049756p.

[63] Sherman W, Day T, Jacobson MP, Friesner RA, Farid R. Novel procedure for modeling ligand/receptor induced fit effects. Journal of Medicinal Chemistry. 2006 jan;49(2):534–553. Available from: https://pubs.acs.org/doi/abs/10.1021/jm050540c.

[64] Amaro RE, Baudry J, Chodera J, Demir Ö, McCammon JA, Miao Y, et al.. Ensemble Docking in Drug Discovery. Biophysical Society; 2018.

[65] Zavodszky MI. Side-chain flexibility in protein-ligand binding: The minimal rotation hypothesis. Protein Science. 2005 mar;14(4):1104–1114. Available from: http://doi.wiley.com/10.1110/ps.041153605.

[66] Cozzini P, Kellogg GE, Spyrakis F, Abraham DJ, Costantino G, Emerson A, et al.. Target flexibility: An emerging consideration in drug discovery and design. American Chemical Society; 2008. Available from: https://pubs.acs.org/doi/abs/10.1021/jm800562d.

[67] Jacobson MP, Friesner RA, Xiang Z, Honig B. On the role of the crystal environment in determining protein side-chain conformations. Journal of Molecular Biology. 2002 jul;320(3):597–608.

[68] Jacobson MP, Pincus DL, Rapp CS, Day TJF, Honig B, Shaw DE, et al. A hierarchical approach to all-atom protein loop prediction. Proteins: Structure, Function, and Bioinformatics. 2004 mar;55(2):351–367. Available from: http://doi.wiley.com/10.1002/prot.10613.

[69] Schrödinger Release 2021-1: Prime, Schrödinger, LLC, New York, NY,. 2021;.

[70] Lecina D, Gilabert JF, Guallar V. Adaptive simulations, towards interactive protein-ligand modeling. Scientific Reports. 2017 dec;7(1):1–11. Available from: https://www.nature.com/articles/s41598-017-08445-5.

[71] Klebe G. On the validity of popular assumptions in computational drug design. Journal of Cheminformatics. 2011 4;3:1–1. Available from: https://jcheminf.biomedcentral.com/articles/10.1186/1758-2946-3-S1-O18.

[72] Huang SY. Comprehensive assessment of flexible-ligand docking algorithms: current effectiveness and challenges. Briefings in Bioinformatics. 2018 sep;19(5):982–994. Available from: https://academic.oup.com/bib/article/19/5/982/3071431.

[73] Plattner N, Noé F. Protein conformational plasticity and complex ligand-binding kinetics explored by atomistic simulations and Markov models. Nature Communications 2015 6:1. 2015 7;6:1–10. Available from: https://www.nature.com/articles/ncomms8653.

[74] Murray CW, Auton TR, Eldridge MD. Empirical scoring functions. II. The testing of an empirical scoring function for the prediction of ligand-receptor binding affinities and the use of Bayesian regression to improve the quality of the model. Journal of Computer-Aided Molecular Design 1998 12:5. 1998;12:503–519. Available from: https://link.springer.com/article/10.1023/A:1008040323669.

[75] Jiménez J, Škalič M, Martínez-Rosell G, De Fabritiis G. KDEEP: Protein-Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks. Journal of Chemical Information and Modeling. 2018 feb;58(2):287–296. Available from: https://pubs.acs.org/doi/abs/10.1021/acs.jcim.7b00650.

[76] Muegge I, Martin YC. A general and fast scoring function for protein-ligand interactions: A simplified potential approach. Journal of Medicinal Chemistry. 1999 3;42:791–804. Available from: https://pubs.acs.org/doi/abs/10.1021/jm980536j.

[77] Li J, Fu A, Zhang L. An Overview of Scoring Functions Used for Protein-Ligand Interactions in Molecular Docking. Interdisciplinary Sciences: Computational Life Sciences 2019 11:2. 2019 3;11:320–328. Available from: https://link.springer.com/article/10.1007/s12539-019-00327-w.

[78] Warren GL, Andrews CW, Capelli AM, Clarke B, LaLonde J, Lambert MH, et al. A critical assessment of docking programs and scoring functions. Journal of Medicinal Chemistry. 2006 oct;49(20):5912–5931. Available from: https://pubmed.ncbi.nlm.nih.gov/17004707/.

[79] Guedes IA, Barreto AMS, Marinho D, Krempser E, Kuenemann MA, Sperandio O, et al. New machine learning and physics-based scoring functions for drug discovery. Scientific Reports 2021 11:1. 2021 2;11:1–19. Available from: https://www.nature.com/articles/s41598-021-82410-1.

[80] Hu B, Lill MA. Exploring the potential of protein-based pharmacophore models in ligand pose prediction and ranking. Journal of Chemical Information and Modeling. 2013 may;53(5):1179–1190. Available from: https://pubs.acs.org/doi/abs/10.1021/ci400143r.

[81] Peach ML, Nicklaus MC. Combining docking with pharmacophore filtering for improved virtual screening. Journal of Cheminformatics. 2009 dec;1(1):6. Available from: https://jcheminf.biomedcentral.com/articles/10.1186/1758-2946-1-6.

[82] Da C, Kireev D. Structural protein-ligand interaction fingerprints (SPLIF) for structure-based virtual screening: Method and benchmark study. Journal of Chemical Information and Modeling. 2014 sep;54(9):2555–2561. Available from: https://pubs.acs.org/doi/10.1021/ci500319f.

[83] Deng Z, Chuaqui C, Singh J. Structural Interaction Fingerprint (SIFt): A Novel Method for Analyzing Three-Dimensional Protein-Ligand Binding Interactions. Journal of Medicinal Chemistry. 2004 jan;47(2):337–344. Available from: https://pubs.acs.org/doi/abs/10.1021/jm030331x.

[84] Jasper JB, Humbeck L, Brinkjost T, Koch O. A novel interaction fingerprint derived from per atom score contributions: exhaustive evaluation of interaction fingerprint performance in docking based virtual screening. Journal of Cheminformatics. 2018 dec;10(1):15. Available from: https://jcheminf.biomedcentral.com/articles/10.1186/s13321-018-0264-0.

[85] Mysinger MM, Carchia M, Irwin JJ, Shoichet BK. Directory of useful decoys, enhanced (DUD-E): Better ligands and decoys for better benchmarking. Journal of Medicinal Chemistry. 2012 jul;55(14):6582–6594. Available from: https://pubs.acs.org/doi/10.1021/jm300687e.

[86] Caporuscio F, Rastelli G, Imbriano C, Del Rio A. Structure-based design of potent aromatase inhibitors by high-throughput docking. Journal of Medicinal Chemistry. 2011 jun;54(12):4006–4017. Available from: https://pubs.acs.org/doi/abs/10.1021/jm2000689.

[87] Varela-Rial A, Majewski M, Cuzzolin A, Martínez-Rosell G, De Fabritiis G. SkeleDock: A Web Application for Scaffold Docking in PlayMolecule. Journal of Chemical Information and Modeling. 2020 jun;60(6):2673–2677. Available from: https://pubs.acs.org/doi/abs/10.1021/acs.jcim.0c00143.

[88] Hou T, Wang J, Li Y, Wang W. Assessing the performance of the molecular mechanics/Poisson Boltzmann surface area and molecular mechanics/generalized Born surface area methods. II. The accuracy of ranking poses generated from docking. Journal of computational chemistry. 2011;32(5):866–877.

[89] Genheden S, Ryde U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. Expert opinion on drug discovery. 2015;10(5):449–461.

[90] Pu C, Yan G, Shi J, Li R. Assessing the performance of docking scoring function, FEP, MM-GBSA, and QM/MM-GBSA approaches on a series of PLK1 inhibitors. MedChemComm. 2017;8(7):1452–1458.

[91] Brandsdal BO, Österberg F, Almlöf M, Feierberg I, Luzhkov VB, Åqvist J. Free energy calculations and ligand binding. Advances in protein chemistry. 2003;66:123–158.

[92] Parenti MD, Rastelli G. Advances and applications of binding affinity prediction methods in drug discovery. Biotechnology advances. 2012;30(1):244–250.

[93] Limongelli V, Bonomi M, Parrinello M. Funnel metadynamics as accurate binding free-energy method. Proceedings of the National Academy of Sciences. 2013;110(16):6358–6363.

[94] Jorgensen WL, Thomas LL. Perspective on free-energy perturbation calculations for chemical equilibria. Journal of chemical theory and computation. 2008;4(6):869–876.

[95] De Vivo M, Masetti M, Bottegoni G, Cavalli A. Role of molecular dynamics and related methods in drug discovery. Journal of medicinal chemistry. 2016;59(9):4035–4061.

[96] Cournia Z, Allen B, Sherman W. Relative Binding Free Energy Calculations in Drug Discovery: Recent Advances and

Practical Considerations. American Chemical Society; 2017. Available from: https://pubs.acs.org/doi/10.1021/acs.jcim.7b00564.

[97] Schindler CEM, Baumann H, Blum A, Böse D, Buchstaller HP, Burgdorf L, et al. Large-scale assessment of binding free energy calculations in active drug discovery projects. Journal of Chemical Information and Modeling. 2020 nov;60(11):5457–5474. Available from: https://pubs.acs.org/doi/abs/10.1021/acs.jcim.0c00900.

[98] Sherborne B, Shanmugasundaram V, Cheng AC, Christ CD, DesJarlais RL, Duca JS, et al. Collaborating to improve the use of free-energy and other quantitative methods in drug discovery. Journal of computer-aided molecular design. 2016;30(12):1139–1141.

[99] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res. 2000 jan;28(1):235–242. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/28.1.235.

[100] Ballester PJ, Mitchell JBO. A machine learning approach to predicting proteinâligand binding affinity with applications to molecular docking. Bioinformatics. 2010 may;26(9):1169–1175. Available from: https://academic.oup.com/bioinformatics/article/26/9/1169/199938.

[101] Ragoza M, Hochuli J, Idrobo E, Sunseri J, Koes DR. Protein-Ligand Scoring with Convolutional Neural Networks. Journal of Chemical Information and Modeling. 2017 apr;57(4):942–957. Available from: https://pubmed.ncbi.nlm.nih.gov/28368587/.

[102] Kolb P, Irwin J. Docking Screens: Right for the Right Reasons? Current Topics in Medicinal Chemistry. 2009 jun;9(9):755–

770. Available from: https://www.eurekaselect.com/85032/article.

[103] Lolli G, Caflisch A. High-Throughput Fragment Docking into the BAZ2B Bromodomain: Efficient in Silico Screening for X-Ray Crystallography. ACS Chemical Biology. 2016 mar;11(3):800–807. Available from: https://pubs.acs.org/doi/10.1021/acschembio.5b00914.

[104] Hermann JC, Marti-Arbona R, Fedorov AA, Fedorov E, Almo SC, Shoichet BK, et al. Structure-based activity prediction for an enzyme of unknown function. Nature. 2007 aug;448(7155):775–779. Available from: https://www.nature.com/articles/nature05981.

[105] Wagner JR, Churas CP, Liu S, Swift RV, Chiu M, Shao C, et al. Continuous Evaluation of Ligand Protein Predictions: A Weekly Community Challenge for Drug Docking. Structure. 2019 aug;27(8):1326–1335.e4.

[106] Martinez-Rosell G, Lovera S, Sands ZA, De Fabritiis G. PlayMolecule CrypticScout: Predicting Protein Cryptic Sites Using Mixed-Solvent Molecular Simulations. Journal of chemical information and modeling. 2020 apr;60(4):2314–2324. Available from: https://pubs.acs.org/doi/abs/10.1021/acs.jcim.9b01209.

[107] Shan Y, Arkhipov A, Kim ET, Pan AC, Shawa DE. Transitions to catalytically inactive conformations in EGFR kinase. Proceedings of the National Academy of Sciences of the United States of America. 2013 apr;110(18):7270–7275. Available from: www.pnas.org/cgi/doi/10.1073/pnas.1220843110.

[108] Gervasio FL, Laio A, Parrinello M. Flexible docking in solution using metadynamics. Journal of the American Chemical Society. 2005 mar;127(8):2600–2607. Available from: https://pubs.acs.org/doi/abs/10.1021/ja0445950.

[109] Macalino SJY, Gosu V, Hong S, Choi S. Role of computer-aided drug design in modern drug discovery. Pharmaceutical Society of Korea; 2015. Available from: https://link.springer.com/article/10.1007/s12272-015-0640-5.

[110] Usha T, Shanmugarajan D, Goyal AK, Kumar CS, Middha SK. Recent Updates on Computer-aided Drug Discovery: Time for a Paradigm Shift. Current Topics in Medicinal Chemistry. 2018 feb;17(30):3296–3307. Available from: http://www.eurekaselect.com/158758/article.

[111] Kumar A, Ito A, Hirohama M, Yoshida M, Zhang KYJ. Identification of sumoylation activating enzyme 1 inhibitors by structure-based virtual screening. Journal of Chemical Information and Modeling. 2013 apr;53(4):809–820. Available from: https://pubs.acs.org/doi/abs/10.1021/ci300618e.

[112] Freedman DH. Hunting for New Drugs with AI. Nature Research; 2019.

[113] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021 596:7873. 2021 7;596:583–589. Available from: https://www.nature.com/articles/s41586-021-03819-2.