Send Orders for Reprints to reprints@benthamscience.net

Current Medicinal Chemistry, XXXX, *XX*, 1-00

REVIEW ARTICLE

Developing Generalizable Scoring Functions for Molecular Docking: Challenges and Perspectives

Rodrigo Quiroga[1,#,*] and Marcos Villarreal[1,#,*]

[1]*Departamento de Química Teórica y Computacional, Facultad de Ciencias Químicas, Universidad Nacional de Córdoba. Instituto de Investigaciones en Físico-Química de Córdoba, INFIQC-CONICET, Argentina*

ARTICLE HISTORY

Received: June 27, 2024
Revised: September 18, 2024
Accepted: September 25, 2024

DOI:
10.2174/0109298673334469241017053508

Abstract: Structure-based drug discovery methods, such as molecular docking and virtual screening, have become invaluable tools in developing novel drugs. At the core of these methods are Scoring Functions (SFs), which predict the binding affinity between ligands and protein targets. This study aims to review and contextualize the challenges and best practices in training novel scoring functions to improve their accuracy and generalizability in predicting protein-ligand binding affinities. Effective training of scoring functions requires careful attention to the quality of training data and methodologies. We emphasize the need for robust training strategies to produce consistent and generalizable SFs. Key considerations include addressing hidden biases and overfitting in machine-learning models, as well as ensuring the use of high-quality, unbiased datasets for both training and evaluation of SFs. Innovative hybrid methods, combining the advantages of empirical and machine-learning approaches, hold promise for outperforming current scoring functions while displaying greater generalizability and versatility.

Keywords: Molecular docking, scoring function, computational drug discovery, virtual screening, machine learning, deep learning.

1. INTRODUCTION

1.1. Computer-aided Drug Discovery

Over the past few decades, small-molecule drug discovery has been achieved mainly by expensive and time-consuming high-throughput screening, with relatively low hit rates [1]. In recent years, the increasing computational resources available to researchers worldwide, coupled with the improving quality and quantity of virtual compound libraries, suggest that computer-aided drug discovery could have the potential to redefine the drug discovery and development process [2]. Computer-Aided Drug Discovery (CADD) encompasses various methodologies that attempt to use computational methods to discover or design new drugs. We can briefly attempt to classify these methodologies into two types of methods: Ligand-based drug discovery (LDBB) and structure-based drug discovery (SBDD). In this review, we will discuss structure-based methods such as molecular docking and virtual screening, focusing on SFs and the methods used to develop, parameterize, train, and test them.

focusing on scoring functions (SFs)

1.2. Overview of Structure-based Drug Discovery, Molecular Docking

Structure-based drug discovery often depends on the availability of high-quality protein structures. Although this has often represented an important limitation, the increasing availability and recent short turnaround for the structures of membrane-associated proteins like G protein-coupled receptors [3] have mitigated this limitation [2]. It is also worth noting that although Alphafold2 [4] and similar programs have represented a major breakthrough in predicting protein structures, these predicted structures still appear to be a worse resource for molecular docking than experimental structures [5]. Molecular docking is a computational method that attempts to predict the most likely position, orientation, and conformation with which a ligand (often a small organic molecule) can bind to a protein [6]. Docking programs mostly consist of two interacting parts. A scoring function that describes the molecu-

*Address correspondence to this autor at the Departamento de Química Teórica y Computacional, Facultad de Ciencias Químicas, Universidad Nacional de Córdoba. Instituto de Investigaciones en Físico-Química de Córdoba, INFIQC-CONICET, Argentina; E-mails: rquiroga@unc.edu.ar; mvillarreal@unc.edu.ar
# *These authors contributed equally to this work.*

0929-8673/XX $65.00+.00

lar interactions between protein and ligand, and a search algorithm that attempts to find the global minimum of the scoring function with its corresponding ligand position, orientation, and conformation. In the following sections, we will provide a brief overview of scoring functions and search algorithms.

1.3. Overview of Scoring Functions

The binding free energy of a ligand to a protein can be estimated in a number of different ways, and docking programs can, therefore, be classified into one of the following five categories [6]:

1- Physics-based (also known as force-field based)
2- Knowledge-based potentials
3- Empirical scoring functions
4- Descriptor based (also known as machine learning based)
5- Hybrid strategies

Various docking programs using all five strategies have been used successfully in different drug discovery projects [3, 7, 8]. Physics-based scoring functions were amongst the first to be used in docking software, as early versions of both Autodock and DOCK employed scoring functions based on the popular AMBER [9] force field. In some cases, these functions incorporated an additional H-bond term and/or solvation energy terms, generally computed using Poisson-Boltzmann or Generalized Born models [6]. Recently, physics-based scoring function terms have been used to train machine learning algorithms to derive both general and target-specific scoring functions [10].

1.4. Search Algorithms

Many different search algorithms have been developed and used to find scoring function minimums [11]. Simulated annealing, genetic algorithms, Lamarckian genetic algorithms, Monte Carlo algorithms, Tabu search, particle swarm optimization, as well as combinations of these algorithms, or combinations with local minimization algorithms such as BFGS, LBFGS, steepest descent, and the Solis and Wets algorithm have all been implemented in different docking software [11-13].

2. SCORING FUNCTION TYPES

2.1. Physics-based Scoring Functions

Physics-based scoring functions use energy terms from well-established force fields used in molecular dynamics or Monte Carlo simulations to estimate the non-covalent interaction energy between protein and ligand atoms. The first physics-based scoring functions in docking programs were designed to compute potential energy in the gas phase, which is only one component of the free energy change in a protein−ligand binding process [6]. Later developments incorporated continuum solvation models. Perhaps the most cited and used scoring functions of this group are the initial Autodock software [14], the GoldScore function used within the GOLD docking software [15], and scoring functions based on AMBER within the DOCK docking software [16]. MM-PBSA/GBSA methods can also be considered part of this group, although these methods employ standard force fields to compute potential energies, and they usually rely on molecular dynamics simulations in explicit solvents for configurational sampling [6].

The main strengths of physics-based scoring functions are an accurate representation of the physical interaction between the protein and the ligand, the general applicability of these functions to diverse sets of both proteins and ligand molecules, as well as the ability to possibly analyze complex interactions such as metal coordination, although this may also be construed as a weakness since these complex interactions are sometimes highly sensitive to parametrization and it can be challenging to obtain correct bond lengths and geometries [17]. Another important weakness of these methods is the need for intensive computational resource use, which limits their application in mass virtual screening experiments [18].

2.2. Knowledge-based Scoring Functions

Knowledge-based scoring functions arise from the application of Boltzmann's principle to construct potentials of mean force from databases of protein-ligand structures [19]. Perhaps the most notable knowledge-based scoring function is Drugscore [20, 21] and its successor DSX [22]. The main advantages of these scoring functions are that they can be trained on crystallographic data without requiring experimental binding affinity data and also the low computational resources expended in their calculation [23]. Their main weakness is the difficulty in adequately representing rare atomic interactions, as well as the need to couple docking runs performed with these functions with other types of scoring functions to perform virtual screening [24].

2.3. Empirical Scoring Functions

Empirical scoring functions aim to estimate the protein-ligand binding energy by adding the individual

contributions from different terms. These terms aim to represent different physicochemical phenomena observed in protein-ligand interactions. The almost ubiquitous terms used in different empirical scoring functions are a Van der Waals term, a solvation/electrostatic term, a hydrophobic interaction term, an H-bond term, and a term to account for entropy [6]. Each of these terms is normally considered to contribute to the overall score in a linear manner, although some scoring functions have explored non-linear relationships [25].

Once these terms have been defined, most empirical scoring functions are derived by training a set of parameters, such as atomic radii, constants, and atom--type-dependent parameters, by performing Multivariate Linear Regression [MLR) or partial least-squares [PLS) on a dataset of protein-ligand complexes with known three-dimensional structures and experimentally measured binding affinities. This is true for the first empirical scoring function developed, LUDI [26], as well as for some of the most used and cited functions such as Chemscore [27], X-score [28], GlideScore [29], the Autodock Vina scoring function [30], and the Vina-based Vinardo [31] scoring function.

One of the main advantages of empirical scoring functions is their low computational cost, especially when individual terms only involve measuring diatomic distances between ligand and protein atoms. This allows for the use of empirical scoring functions for ultra-,large-scale virtual screening, where millions of virtual compounds are screened to predict potential binders to a protein of interest [32]. Another advantage worthy of mention is the simplicity of empirical scoring functions and the interpretability of docking results. This interpretability is key in developing protein-specific scoring functions [33].

Amongst the limitations of empirical scoring functions is the fact that over-simplification of complex physical interactions between atoms can lead to problems when performing docking on proteins such as metallo-enzymes or highly hydrated and/or water-bridge dependent protein-ligand interactions [34-36]. For this reason, some programs have been developed to perform hydrated or "wet" docking [37-39]. Also, the general applicability of a scoring function to novel protein-ligand complexes will depend strongly on the training set used [for example, the variety of atom types observed in said training set structures). The need for training set complexes to include experimental binding affinity data limits the number of structures to be used for training.

Finally, it should be mentioned that if one considers the possible "scoring function space" for empirical SFs, it can be said that we have explored a relatively small portion of the possibilities [33, 40]. The scoring function space is vast and mainly populated with non--generalizable or badly performing functions. This renders parametrization and empirical SFs design (the selection of terms and the possible linear and non-linear ways to combine them) a challenging endeavor. As we explore in forthcoming sections, training SFs to simultaneously optimize scoring, docking, and ranking capabilities should allow further exploration of this space and possibly lead to SF innovation.

2.4. Descriptor-based Scoring Functions (Machine Learning)

These SFs use a variety of ligand-protein interaction descriptors such as ligand atom type and number, ligand-protein atom pair distances, structural interaction fingerprints, geometrical descriptors and/or empirical-function-like interaction terms that have been used to train traditional Machine Learning (ML) algorithms, such as Random Forest (RF), Bayesian classifiers, Support Vector Machine (SVM), Gradient Boosting, feed--forward and convolutional Neural Networks [41]. These functions are normally trained on protein-ligand complexes for which both structure and binding affinity data are available. In this sense, they are similar to empirical SFs. Unlike empirical SFs, the descriptors used are often not physics-inspired, and they can be extremely complex models and often not a linear combination of descriptors [6].

An advantage of descriptor-based SFs is that they are normally able to achieve higher performance metrics than empirical scoring functions for the task they are trained to perform. This is true for scoring capabilities [42, 43] as well as for virtual screening [44-46]. However, overfitting [47], memorization of ligand features without actual "learning" [47-49], learning the inherent bias of training datasets [44, 48, 50], and noticeable drops in performance when applying SFs to tasks other than those they were trained for [51, 52], or performing the same task on other datasets [44, 48], have all shown the limitations of the application of this family of SFs for SBDD.

Another limitation of machine-learning SFs is that the derivatives of the SF with respect to the geometrical coordinates are difficult to calculate, making the searches to find the global minimum challenging and costly. As a result, these SF are typically used for re-scoring (see section 3.C.1) rather than to perform actual molecular docking.

In recent years, exciting new progress has been made with several deep learning-based docking meth-

ods, the pioneering work of DeepDock [53], then EquiBind [54], TankBind [55], DiffDock [56], and Uni-Mol [57]. However, it was shown that this first generation of methods did not outperform "classical" docking software and SFs such as Vina and Gold at docking and were not apparently generalizable to datasets that differ from their training sets [58]. Also, diffusion-inspired pose minimization methods, such as PLANTAIN [59], improved upon these results. More recently, a second generation of programs that use diffusion methods promises large improvements at reproducing experimental structures of protein-ligand complexes by performing dockings, such as Uni-Molv2 [60], DiffDock-L [61], and Alphafold3 [62]. It remains to be seen if this docking performance is generalizable and also how plausible and effective these methods are at performing virtual screening with large and ultra-large virtual screening libraries, although there are recent advances based on promising transformer-based [63] and pharmacophore-based [64] approaches to performing virtual screening.

2.5. Hybrid Methodologies

Some scoring functions were developed using a mixture of the SF types mentioned above [65-67], or consensus methods were used for a group of established SFs [49]. Other more recent advances that combine empirical and machine-learning-based approaches have also proven promising [68]. However, hybrid methodologies have been the focus of little recent research when compared to machine-learning-based approaches.

3. METHODOLOGY BEHIND SCORING FUNCTION TRAINING

3.1. Data Collection and Preprocessing

Whatever the type of SF, and although low-quality data has been shown to sometimes be helpful, high-quality training data is necessary for training. Although others exist and have been reviewed elsewhere [41], the most utilized databases for structures of protein-ligand complexes and their binding data are PDBBIND [69] and CASF [70] for protein-ligand structures and binding affinity, and DUD-E [71] for virtual screening.

Of these databases, it should be noted that the last PDBBIND release that is openly available is v2020, while starting from v2021, the database has switched to a subscription-based system (https://www.pdbbind--plus.org.cn/).

The CASF database is a subset of PDBBIND, with the latest release version being v2016 [72]. CASF offers the enticing possibility to standardize SF evaluation, which is a promising concept. However, it should be noted that the evaluation of docking power and ranking power (see section 3. C) performed with CASF is conductedin a non-self-consistent manner, which entails multiple caveats, especially for complex scoring functions such as machine-learning-based SFs.

Compounds experimentally determined to not bind to certain proteins constitute valuable information for virtual screening dataset construction [73]. However, this experimental information is severely lacking. Initial datasets using only these experimentally verified inactive compounds or decoys were small [74], but recent efforts in this direction have constructed promising, larger datasets [75].

Due to the scarcity of "negative examples," "non-binders," or low-binding affinity data and structures, data augmentation has been proposed as a way to generate more informative training sets that should increase the generalizability of SFs as well as improve their docking and virtual screening performance [71, 76]. Although this is a step in the right direction, this augmented data is generated in a non-self-consistent way. This diminishes its potential to actually improve the generalizability and performance of SFs trained on this augmented data. Adapting such strategies to augment data in a self-consistent manner could be one of the possible routes to pave the way for developing better SFs, although this methodology may be more plausible and easily applied to empirical or hybrid SFs.

Regarding virtual screening datasets, the DUD-E database is one of the most used and cited, but has been shown to be inherently biased [48] and should be used with caution for training purposes, especially for machine-learning-based SFs. As highlighted above, these methods can easily "learn" these biases and can display outstanding performance metrics that are not expected to be extrapolated to other datasets. Several alternatives have been developed in the last few years, with some notable examples being the AD dataset [48], LIT-PCBA [75], and TocoDecoy [77].

Recently, several methods have been proposed for generating decoys with less bias or to generate a determined bias [78]. This is a sound idea, but as is always the case with the programmatic generation of decoy compounds, special care must be taken to ensure that SFs trained using this data do not "learn" the inverse function of the methods used to generate the decoys, therefore being able to achieve spuriously high virtual screening performance with little generalizability.

3.2. Feature Selection and Representation

Regarding both empirical and descriptor-based (machine-learning) SFs, feature selection is a vital part of training a novel scoring function. Great care should be taken when including non-physics based descriptors, and even non-interaction protein and ligand descriptors, as these by themselves have been shown to be enough to achieve high correlations with experimental binding data when used to train machine learning algorithms such as random forest and convolutional neural networks [47, 48, 52]. Reducing the number of non-physical and non-interaction descriptors should reduce the probability of overfitting and memorization. All SFs, which include ligand and protein descriptors as well as interaction-based descriptors, should be trained and tested both with and without interaction-based descriptors to analyze the extent to which the SF shows memorization issues.

3.3. Training SFs for Different Tasks and the Success Metrics used for Each

There are three main tasks which we ask of scoring functions:

3.3.1. Scoring Power

Scoring power can be defined as the ability to predict a protein-ligand interaction score for experimental structures, which correlates with experimental binding affinities [70, 79]. This is the "classical" task that empirical and physics-based SFs are trained to perform [6]. However, when the SF is complex, and the scoring task is the only task used to train the SF, there is no guarantee that the resulting scoring function has global or even local minima corresponding to protein-ligand complexes that closely resemble the experimental training structures. In other words, the parameters of the SF (which are normally trained by MLR in the case of empirical scoring functions or diverse machine learning algorithms in the case of descriptor-based methods) are highly degenerated, which means solutions do not converge to a single parameter combination, but instead there are multiple possibilities which represent diverging SFs, albeit with similar scoring power results. The amount of information provided to our training model is scarce, and this may partially explain why machine-learning-based scoring functions are able to outperform physics and empirical scoring functions without the need to even include features that describe protein-ligand interactions [48, 52]. Great care should be taken when applying an SF trained only for scoring power to other tasks, such as docking or virtual screening, or when applying it to a dataset on which it was not trained since there are little to no guarantees that its capacities extrapolate to cases or tasks not included in the training dataset.

3.3.2. Docking Power

Docking power can be defined as the ability of a Scoring Function (SF) to find a pose that represents the global minimum of the protein-ligand interaction score and which also closely resembles the experimentally determined structure of the complex [70, 79]. As we mentioned above, the scoring function space for SF that has good scoring capabilities is very large and diverse. However, docking is a more demanding task and should prove a bottleneck for some of the widely varying parameter combinations that display high-scoring power. Docking power is quite straightforward to evaluate and test for physics-based, knowledge-based, and empirical scoring functions, as nearly all docking programs combine SFs with a search algorithm. This combination allows the exploration of different poses to identify those representing the global and local minima of the scoring function, which can then be compared to the experimentally determined structure of each protein-ligand complex. To this end, Root Mean Squared Deviation [RMSD] is the most used structural distance measure, with the standard being a 2A cutoff to establish similarity between predicted poses and the experimental pose. It should be noted that this is not always ideal. Different ligand or protein-ligand characteristics (large ligands, total or partial ligand symmetry, solvent-exposed ligand poses, non-specific ligand binding, *etc.*) can lead to high RMSD values. Other methodologies have been proposed which are promising replacements for using a 2A RMSD threshold, such as SuCOS [80]. Docking power is more difficult to measure for "black-box" SFs, and generally, this is performed by generating poses with empirical scoring functions and then scoring one or more of the generated poses [44, 45, 81, 82]. It should be noted that this is not self-consistent. There is no guarantee that the poses generated also represent global or local minima for the machine-learning scoring function. Therefore, using CASF to evaluate machine-learning SFs may be an unfair comparison, which can lead to an overestimation of their docking power since we are not comprehensively searching for their global minima for each protein-ligand complex in the training set. Finally, scoring functions trained only to optimize docking power could still probably fail to adequately compare the binding affinities of different ligands to the same or different proteins since the training data only provides information about the relative binding affinity of different poses of the same ligand, while providing no infor-

mation on the relative binding affinity of different ligands.

3.3.3. Virtual Screening (Ranking) Power

Virtual Screening power, or ranking power, can be defined as the ability of an SF to correctly rank a variety of ligands that have been determined to be "active" or "inactive" for a particular protein (depending on whether they bind to the protein or not), or by their experimental binding affinities if available [70, 79]. This task requires that the scoring function be able to adequately compare the binding scores of ligands, which many times have very different sizes, atomic composition, and binding modes.

Virtual screening benchmarks are used for both training and evaluating the ranking power of SFs.

Akin to scoring power, virtual screening power is often used to train machine-learning SFs. The SF is trained to rank known active compounds above "decoy" compounds, which are expected to not bind to the protein. The pose for which each decoy will be evaluated for scoring the binding to the target is usually generated by using established, traditional SFs to perform docking. In summary, the SF is both trained and evaluated using poses generated by other SFs, so there is again no guarantee that these poses used to train and evaluate the machine-learning-based SF represent global or local minima.

These decoys are sometimes generated programmatically [71, 83] and display hidden biases [48], which can be learned by complex SF producing apparently excellent virtual screening power [44]. However, when evaluated with other datasets, this performance drops and caution should be exercised when using these SFs to perform virtual screening for real-world targets.

3.3.4. Training Scoring Functions Combining the Capabilities to Perform all Three Tasks

We argue that improved outcomes for SBDD in drug discovery would follow from the use of a self-consistent approach for training SFs, *i.e.*, a methodology that allows scoring, docking, and virtual screening with the same scoring function. The drawback and limitation of using different SFs for each of these activities is that, most probably, the scoring functions might not be in register, *i.e.*, their minima do not occur in similar coordinates of the conformational space. The use of two or more different scoring functions is a very common strategy in virtual screening for SBDD. The idea is first to use a function with low computational requirements for docking and then to rank the obtained poses using a second function that has a better overall performance but is more expensive to use for docking or that cannot be used for docking at all, as discussed in section 2.4 section 2.D. The consequences of the lack of registry between the two functions can be understood by looking at (Fig. **1**). In this scenario, the docking is made with the first function, providing the configurations for the second function to score them. The lack of registration between these two functions results in that the minima in one function may not be minima in the other, and then we do not get the true answer from the second and better function.

The question then arises as to how we can develop self-consistent scoring functions. We can think of two approaches to train an SF; one includes regression, and the other is by classification. In the first approach, which is how most functions are developed [14, 30, 31], some measure of the difference between the score and the experimental binding energy value is minimized. This regression is performedwith space missing data gathered from the databases mentioned in section 3. A, section 3.1, such as the aforementioned PDBBIND. It is important to note that this parameter fitting process defines not only the score value for the experimental pose but also the whole landscape, as shown in Fig. (**1**). In this figure, we schematize the landscape for two scoring functions, both of which give the same score for the experimental pose. For this reason, both will have the same scoring power [see section 3.3.1 section 3.C.1]. However, it is easy to see that they will give very different results in other applications. For example, minimizing the experimental pose with SF 1 will yield energy and pose very similar to the starting one, as it is close to a minimum. More interestingly, docking with this function will produce a successful case since the minimum close to the experimental pose is the global minimum. On the other hand, function 2 will not produce good results in either minimization or docking despite having equal scoring power. It is clear then that training using only regression is not guaranteed to generate functions with good docking power. This indicates that the SF space must shrink when the number of tasks to be performed increases. Another way to train an SF could be based on generated pose decoys instead of relying on experimental results alone. Here, using some pre-existing docking methodology, a database is created with poses that are "different" from the experimentally determined pose for each ligand. As mentioned in section 3.C.2 section 3.3.2, structural comparisons are generally conducted by using a value of 2A in RMSD as a cutoff. This is equivalent to assuming that poses with less than 2A have a score equal to or similar to that of the experimental pose, while the rest of the poses will have to be of a higher score since the experimental pose is assumed as the one with the

lowest free energy. In this scenario, the remaining task is to classify the experimental pose as active or with a lower assigned score, while the rest of the decoy poses will have to be classified as inactive or have a higher score. In this way, taking a representative number of decoys will enforce an overall minimum for the experimental pose on the SF under training. Generating the decoy structures consistent with the function being trained is key. As a theoretical example, let's assume that the decoys have been generated with a previous function that uses small atomic radii. This will cause the decoys to be very close to the protein and thus easily detectable with a function with larger radii since the decoys will, in this case, give high Van der Waals repulsion energies. In this scenario, the training can again achieve very good metrics even in the test sets but give very poor docking results. This is because when docking the comparison of the score of the experimental pose is no longer compared to the decoys generated by another SF but to poses generated by the function itself, and therefore represent local or global minimums.

comma missing after the word "docking"

This decoy-based training can also be extended to calibrate a function for virtual screening [76]. This task has the added complexity that the decoy poses have to be generated for both active and inactive molecules of each protein target. Again, the lack of self-consistency of the decoy poses can generate good virtual screening metric values in training but serious flaws in model generalization. This situation is depicted in Fig. (**2**). Each panel of the Figure represents the landscapes obtained with different SF for an active and two inactive molecules in the same protein target. Both SFs perform equally well in the virtual screening task as they are able to correctly distinguish the active from the inactive when fed with the poses indicated by the dots. Nevertheless, in an actual application of virtual screening, where docking is performed to find the global minimum before comparing scores, the SF in panel B will rank both decoys as better candidates than the active molecule, while the SF in panel A ranks the active molecule as the best candidate. This theoretical example shows that the scoring function space shrinks when we optimize the performance of SFs to perform multiple tasks. This observation is important and suggests that SF innovation can not only be achieved by increasing the amount and quality of training data (big data thinking) but also by including several different tasks to really learn all the different nuances that are at play when applying SFs in real life SBDD projects.

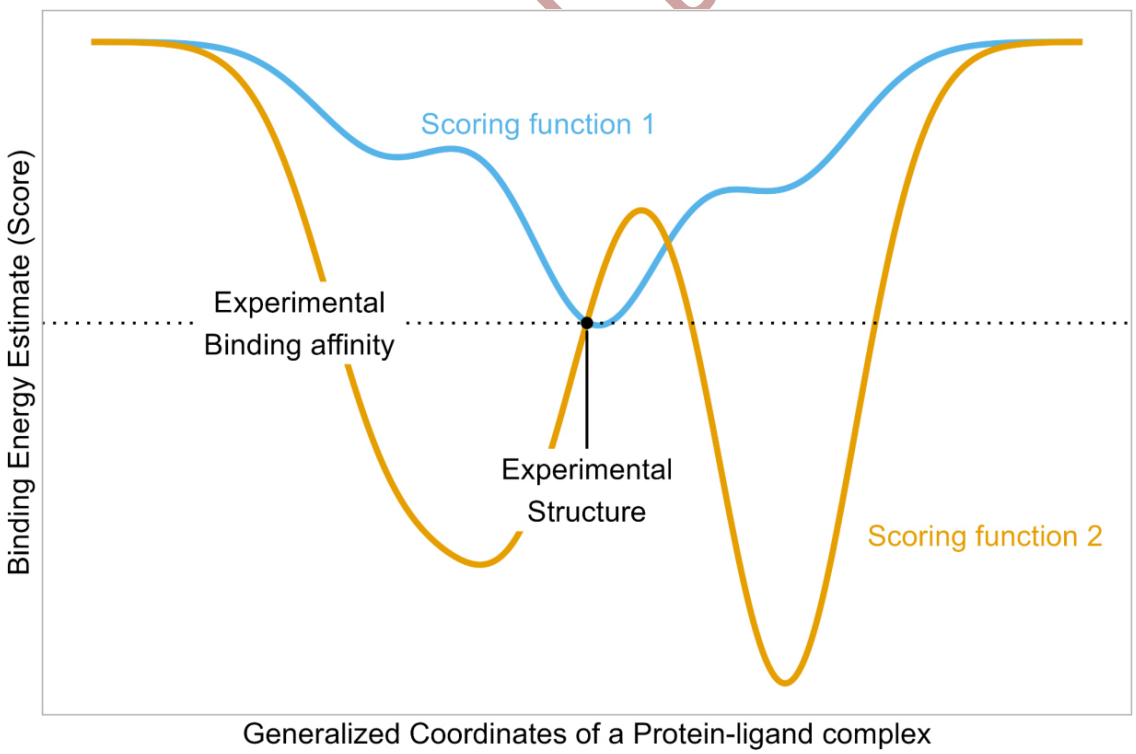

Fig. (1). Schematic binding affinity landscapes for a given protein-ligand complex for two scoring functions. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

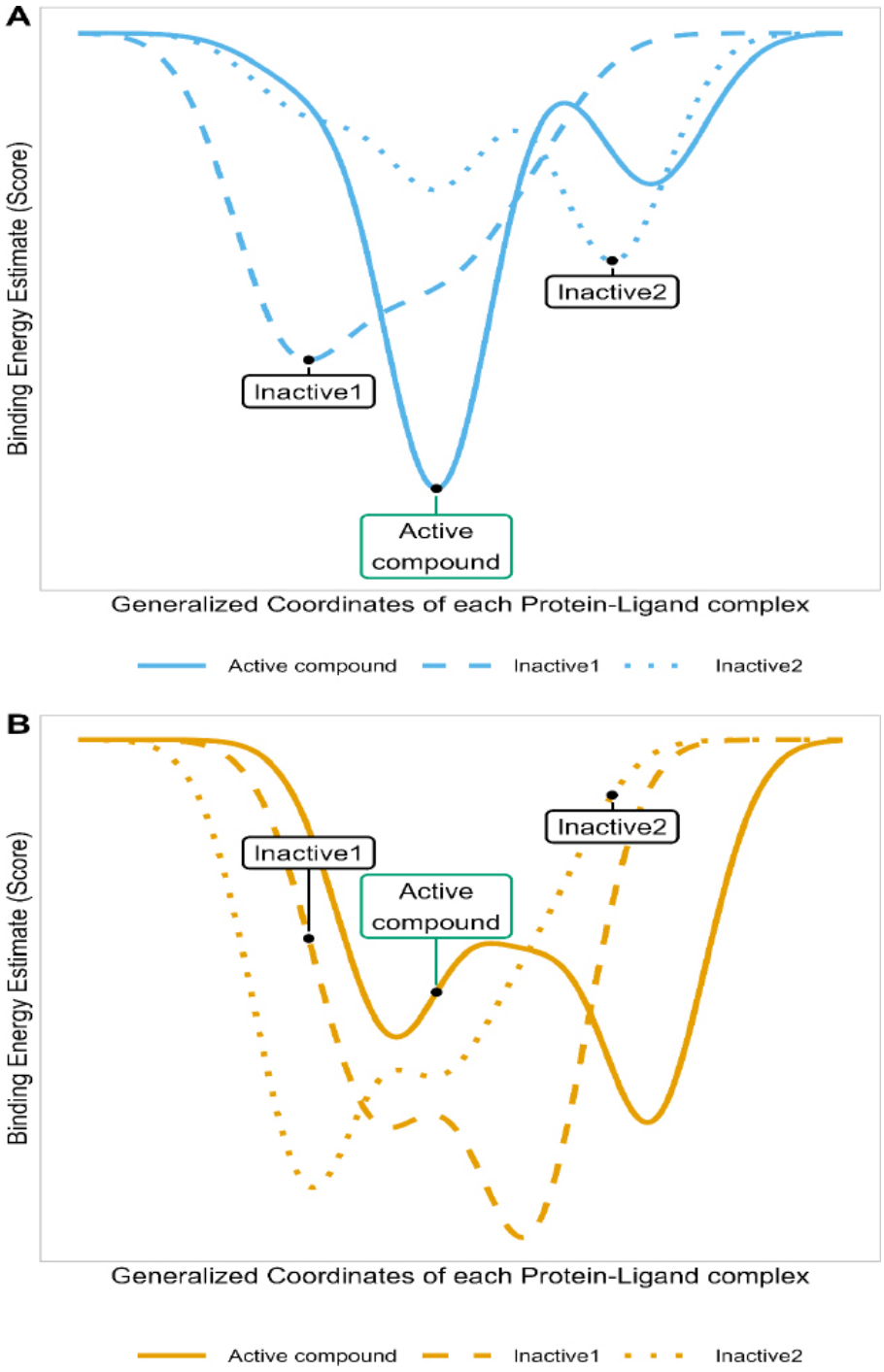

Fig. (2). Schematic binding affinity landscapes for active and inactive compounds on the same target for two scoring functions. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

Based on these considerations, a possible way to train a self-consistent SF is using an iterative procedure. The first iteration would minimize two terms. The first is to measure the difference between the calculated score and the experimental binding energy. The second term is the forces (the derivatives of the SF with respect to the geometrical coordinates) felt by the experimental pose, which serves to enforce the existence of a minimum at this position. Now, a redocking of ligands is performed, and wrong poses are incorpo-

rated in the database of decoy poses for that ligand. In the next iterations, a third term is incorporated that measures the difference in score between the experimental and the decoy poses in order to enforce the experimental pose as the global minimum. As the iterations proceed, a fraction of older decoys must be removed. In this way, the pose decoy database will better reflect the current state of the SF under training. After a finite number of cycles and with some convergence threshold, this procedure should result in a self-consistent SF capable of simultaneously optimizing Scoring, Docking, and Virtual screening capabilities.

Finally, another interesting application of scoring functions is the development of functions that are tailor-made for SBDD with a particular protein or protein family. Training SFs to perform well for a given protein of interest is an enticing idea [33, 40]. However, this idea entails several limitations. The main limitation is that for a given protein, the amount of training data available is greatly diminished in comparison to SFs trained on many different protein families. In turn, training on these small datasets can spuriously improve the performance of complex scoring functions due to overfitting. Akin to the concept of foundational models in AI, a possible route would be to first train a general function using the largest dataset available and then fine-tune it for the desired target protein.

can lead to spurious improvements in

The plot shows the binding energy estimates (scores) corresponding to two different scoring functions (SF1 in light blue and SF2 in orange) across all possible generalized coordinates values of a protein-ligand complex. The point in black is located at the experimental binding energy (dotted black line) and at the experimentally determined coordinate (full black line). The plot shows a putative case where the error of both SF in the scoring task is very low, but the local and global minima of each scoring function correspond to different areas. With SF1, performing either minimization of the experimental structure or docking should result in a similar pose, while the application of SF2 will result in a completely different binding mode.

Each line represents the binding landscape of a molecule to the same protein target, with the active compound in a solid line and inactive compounds in dashed and dotted lines. Two different theoretical SFs are represented in each panel. Dots indicate protein-ligand complex structures used for training/evaluation of ranking power (see section 3.C.3), positioned at the same generalized coordinates in each panel. For the active molecule, the coordinates correspond to the experimental structure; for inactive compounds, coordinates are either experimentally determined or generated through molecular docking. Panel A: A self-consistent SF (light blue) where each training pose represents the global minimum, correctly ranking the active compound as the best binding candidate. Panel B: A non-self-consistent SF (orange). Though this SF ranks the active compound as the best candidate using the given poses, a complete energy landscape exploration would reveal both decoys ranking higher than the active compound.

section 3.3.4

Please change all instances of "an SF" in the text (there are 6), to "a SF"

CONCLUSION

As we reviewed above, training an SF to perform a single task is risky, especially for machine-learning SFs, as in this case, scoring and virtual screening power can only be evaluated in a non-self-consistent manner. Therefore, simultaneously optimizing the capacity of an SF to perform all three tasks could provide a way to outperform existing SFs while possibly granting capabilities to adequately predict novel binding modes and ligand types that were scarce or not present in the training set. It is crucial to address memorization and overfitting in machine-learning SFs, both for training new SFs as well as when using these functions for scoring and/or virtual screening rescoring. Although there have been recent advances, there is an unsatisfied need for high-quality, unbiased datasets for both training and evaluation of SFs.

Novel groundbreaking results for diffusion methods such as AlphaFold 3 do not imply the obsolescence of empirical scoring functions but rather an opportunity to enhance performance through the development of novel methods that combine empirical and machine-learning approaches. These offer a promising solution to possibly surpass current SF limitations while optimizing both generalizability and versatility, as well as the capacity to perform virtual screening using ultra-large libraries. The success of future SF development lies in their ability to generalize across diverse protein families and ligand types, giving the best possible performance in real-world applications.

LIST OF ABBREVIATIONS

RMSD	=	Root Mean Squared Deviation
SF	=	Scoring Function
CADD	=	Computer-aided Drug Discovery
LBDD	=	ligand-based Drug Discovery
SBDD	=	structure-based Drug Discovery
MLR	=	Linear Regression
PLS	=	Partial Least-squares

remove all text in red, these paragraphs correspond to Fig. 1 and Fig. 2 captions, please see Proof_correction_list.docx file for a corrected version of Fig. 1 and Fig. 2 captions, as well as a detailed list of required Proof modifications.

AUTHORS' CONTRIBUTIONS

It is hereby acknowledged that all authors have accepted responsibility for the manuscript's content and consented to its submission. They have meticulously reviewed all results and unanimously approved the final version of the manuscript.

CONSENT FOR PUBLICATION

Not applicable.

FUNDING

This work was supported by the Agencia Nacional de Promoción de la Investigación, el Desarrollo Tecnológico y la Innovación (Agencia I+D+i) grant PICT2019-2019-01979, and Secretaría de Ciencia y Técnica de la Universidad Nacional de Córdoba (SECYT-UNC). R.Q and M.A.V were supported by Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET, Argentina).

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

Declared none.

REFERENCES

[1] Bender, A.; Bojanic, D.; Davies, J.W.; Crisman, T.J.; Mikhailov, D.; Scheiber, J.; Jenkins, J.L.; Deng, Z.; Hill, W.A.; Popov, M.; Jacoby, E.; Glick, M. Which aspects of HTS are empirically correlated with downstream success? *Curr. Opin. Drug Discov. Devel.,* **2008**, *11*(3), 327-337. PMID: 18428086

[2] Sadybekov, A.V.; Katritch, V. Computational approaches streamlining drug discovery. *Nature,* **2023**, *616*(7958), 673-685. http://dx.doi.org/10.1038/s41586-023-05905-z PMID: 37100941

[3] Congreve, M.; de Graaf, C.; Swain, N.A.; Tate, C.G. Impact of GPCR Structures on Drug Discovery. *Cell,* **2020**, *181*(1), 81-91. http://dx.doi.org/10.1016/j.cell.2020.03.003 PMID: 32243800

[4] Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S.A.A.; Ballard, A.J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A.W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly accurate protein structure prediction with AlphaFold. *Nature,* **2021**, *596*(7873), 583-589. http://dx.doi.org/10.1038/s41586-021-03819-2 PMID: 34265844

[5] Holcomb, M.; Chang, Y.T.; Goodsell, D.S.; Forli, S. Evaluation of ALPHAFOLD2 structures as docking targets. *Protein Sci.,* **2023**, *32*(1), e4530. http://dx.doi.org/10.1002/pro.4530 PMID: 36479776

[6] Liu, J.; Wang, R. Classification of current scoring functions. *J. Chem. Inf. Model.,* **2015**, *55*(3), 475-482. http://dx.doi.org/10.1021/ci500731a PMID: 25647463

[7] Sliwoski, G.; Kothiwale, S.; Meiler, J.; Lowe, E.W. Computational methods in drug discovery. *Pharmacol Rev.,* **2014**, *66*(1), 334-395. http://dx.doi.org/10.1124/pr.112.007336

[8] Bender, B.J.; Gahbauer, S.; Luttens, A.; Lyu, J.; Webb, C.M.; Stein, R.M.; Fink, E.A.; Balius, T.E.; Carlsson, J.; Irwin, J.J.; Shoichet, B.K. A practical guide to large-scale docking. *Nat. Protoc.,* **2021**, *16*(10), 4799-4832. http://dx.doi.org/10.1038/s41596-021-00597-z PMID: 34561691

[9] Weiner, S.J.; Kollman, P.A.; Case, D.A.; Singh, U.C.; Ghio, C.; Alagona, G.; Profeta, S.; Weiner, P. A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.,* **1984**, *106*(3), 765-784. http://dx.doi.org/10.1021/ja00315a051

[10] Guedes, I.A.; Barreto, A.M.S.; Marinho, D.; Krempser, E.; Kuenemann, M.A.; Sperandio, O.; Dardenne, L.E.; Miteva, M.A. New machine learning and physics-based scoring functions for drug discovery. *Sci. Rep.,* **2021**, *11*(1), 3198. http://dx.doi.org/10.1038/s41598-021-82410-1 PMID: 33542326

[11] Dias, R.; de Azevedo, W., Jr Molecular docking algorithms. *Curr. Drug Targets,* **2008**, *9*(12), 1040-1047. http://dx.doi.org/10.2174/138945008786949432 PMID: 19128213

[12] Yadava, U. Search algorithms and scoring methods in protein-ligand docking. *Endocrinology&Metabolism International Journal,* **2018**, *6*(6), 359-367. http://dx.doi.org/10.15406/emij.2018.06.00212

[13] Halperin, I.; Ma, B.; Wolfson, H.; Nussinov, R. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins,* **2002**, *47*(4), 409-443. http://dx.doi.org/10.1002/prot.10115 PMID: 12001221

[14] Goodsell, D.S.; Olson, A.J. Automated docking of substrates to proteins by simulated annealing. *Proteins,* **1990**, *8*(3), 195-202. http://dx.doi.org/10.1002/prot.340080302 PMID: 2281083

[15] 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*(3), 727-748. http://dx.doi.org/10.1006/jmbi.1996.0897 PMID: 9126849

[16] DesJarlais, R.L.; Sheridan, R.P.; Seibel, G.L.; Dixon, J.S.; Kuntz, I.D.; Venkataraghavan, R. Using shape complementarity as an initial screen in designing ligands for a receptor binding site of known three-dimensional structure. *J. Med. Chem.,* **1988**, *31*(4), 722-729. http://dx.doi.org/10.1021/jm00399a006 PMID: 3127588

[17] Scrima, S.; Tiberti, M.; Ryde, U.; Lambrughi, M.; Papaleo, E. Comparison of force fields to study the zinc-finger containing protein NPL4, a target for disulfiram in cancer therapy. *Biochim Biophys Acta BBA,* **2023**, *1871*(4), 140921.

[18] Varela-Rial, A.; Majewski, M.; De Fabritiis, G. Structure based virtual screening: Fast and slow. *Wiley Interdiscip. Rev. Comput. Mol. Sci.,* **2022**, *12*(2), e1544. http://dx.doi.org/10.1002/wcms.1544

[19] Sippl, M.J. Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. *J. Comput. Aided*

Mol. Des., **1993**, *7*(4), 473-501. http://dx.doi.org/10.1007/BF02337562 PMID: 8229096

[20] Velec, H.F.G.; Gohlke, H.; Klebe, G. DrugScore(CSD)--knowledge-based scoring function derived from small molecule crystal data with superior recognition rate of near-native ligand poses and better affinity prediction. *J. Med. Chem.,* **2005**, *48*(20), 6296-6303. http://dx.doi.org/10.1021/jm050436v PMID: 16190756

[21] Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based scoring function to predict protein-ligand interactions. *J. Mol. Biol.,* **2000**, *295*(2), 337-356. http://dx.doi.org/10.1006/jmbi.1999.3371 PMID: 10623530

[22] Neudert, G.; Klebe, G. *DSX*: a knowledge-based scoring function for the assessment of protein-ligand complexes. *J. Chem. Inf. Model.,* **2011**, *51*(10), 2731-2745. http://dx.doi.org/10.1021/ci200274q PMID: 21863864

[23] Muegge, I.; Martin, Y.C. A general and fast scoring function for protein-ligand interactions: a simplified potential approach. *J. Med. Chem.,* **1999**, *42*(5), 791-804. http://dx.doi.org/10.1021/jm980536j PMID: 10072678

[24] Hsieh, J.H.; Yin, S.; Wang, X.S.; Liu, S.; Dokholyan, N.V.; Tropsha, A. Cheminformatics meets molecular mechanics: a combined application of knowledge-based pose scoring and physical force field-based hit scoring functions improves the accuracy of structure-based virtual screening. *J. Chem. Inf. Model.,* **2012**, *52*(1), 16-28. http://dx.doi.org/10.1021/ci2002507 PMID: 22017385

[25] Dias, R.; Macedo Timmers, L.F.; Caceres, R.; de Azevedo, W., Jr Evaluation of molecular docking using polynomial empirical scoring functions. *Curr. Drug Targets,* **2008**, *9*(12), 1062-1070. http://dx.doi.org/10.2174/138945008786949450 PMID: 19128216

[26] Böhm, H.J. The computer program LUDI: a new method for the de novo design of enzyme inhibitors. *J. Comput. Aided Mol. Des.,* **1992**, *6*(1), 61-78. http://dx.doi.org/10.1007/BF00124387 PMID: 1583540

[27] Eldridge, M.D.; Murray, C.W.; Auton, T.R.; Paolini, G.V.; Mee, R.P. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput. Aided Mol. Des.,* **1997**, *11*(5), 425-445. http://dx.doi.org/10.1023/A:1007996124545 PMID: 9385547

[28] Wang, R.; Lai, L.; Wang, S. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J. Comput. Aided Mol. Des.,* **2002**, *16*(1), 11-26. http://dx.doi.org/10.1023/A:1016357811882 PMID: 12197663

[29] 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*(7), 1739-1749. http://dx.doi.org/10.1021/jm0306430 PMID: 15027865

[30] 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*(2), 455-461. http://dx.doi.org/10.1002/jcc.21334 PMID: 19499576

[31] Quiroga, R.; Villarreal, M.A. Vinardo: A scoring function based on autodock vina improves scoring, docking, and virtual screening. *PLOS ONE,* **2016**, *11*(5), e0155183.

[32] Li, H.; Leung, K.S.; Wong, M.H. idock: A multithreaded virtual screening tool for flexible ligand docking. *2012 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB),* 09-12 May, 2012, San Diego, CA, USA, 2012, pp. 77-84. http://dx.doi.org/10.1109/CIBCB.2012.6217214

[33] Bitencourt-Ferreira, G.; Villarreal, M.A.; Quiroga, R.; Biziukova, N.; Poroikov, V.; Tarasova, O.; de Azevedo Junior, W.F. Exploring Scoring Function Space: Developing Computational Models for Drug Discovery. *Curr. Med. Chem.,* **2024**, *31*(17), 2361-2377. http://dx.doi.org/10.2174/0929867330666230321103731 PMID: 36944627

[34] Irwin, J.J.; Raushel, F.M.; Shoichet, B.K. Virtual screening against metalloenzymes for inhibitors and substrates. *Biochemistry,* **2005**, *44*(37), 12316-12328. http://dx.doi.org/10.1021/bi050801k PMID: 16156645

[35] Pottel, J.; Therrien, E.; Gleason, J.L.; Moitessier, N. Docking ligands into flexible and solvated macromolecules. 6. Development and application to the docking of HDACs and other zinc metalloenzymes inhibitors. *J. Chem. Inf. Model.,* **2014**, *54*(1), 254-265. http://dx.doi.org/10.1021/ci400550m PMID: 24364808

[36] García-Sosa, A.T. Hydration properties of ligands and drugs in protein binding sites: tightly-bound, bridging water molecules and their effects and consequences on molecular design strategies. *J. Chem. Inf. Model.,* **2013**, *53*(6), 1388-1405. http://dx.doi.org/10.1021/ci3005786 PMID: 23662606

[37] van Dijk, A.D.J.; Bonvin, A.M.J.J. Solvated docking: introducing water into the modelling of biomolecular complexes. *Bioinformatics,* **2006**, *22*(19), 2340-2347. http://dx.doi.org/10.1093/bioinformatics/btl395 PMID: 16899489

[38] Forli, S.; Olson, A.J. A force field with discrete displaceable waters and desolvation entropy for hydrated ligand docking. *J. Med. Chem.,* **2012**, *55*(2), 623-638. http://dx.doi.org/10.1021/jm2005145 PMID: 22148468

[39] Eberhardt, J.; Santos-Martins, D.; Tillack, A.F.; Forli, S. AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. *J. Chem. Inf. Model.,* **2021**, *61*(8), 3891-3898. http://dx.doi.org/10.1021/acs.jcim.1c00203 PMID: 34278794

[40] Bitencourt-Ferreira, G.; De Azevedo, W.F. Exploring the Scoring Function Space. In: *Docking Screens for Drug Discovery*; Springer, **2019**. http://dx.doi.org/10.1007/978-1-4939-9752-7_17

[41] Meli, R.; Morris, G.M.; Biggin, P.C. Scoring Functions for Protein-Ligand Binding Affinity Prediction Using Structure-based Deep Learning: A Review. *Front. Bioinform.,* **2022**, *2*, 885983. http://dx.doi.org/10.3389/fbinf.2022.885983 PMID: 36187180

[42] Ballester, P.J.; Mitchell, J.B.O. A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. *Bioinformatics,* **2010**, *26*(9), 1169-1175. http://dx.doi.org/10.1093/bioinformatics/btq112 PMID: 20236947

[43] Zilian, D.; Sotriffer, C.A. SFCscore(^{RF}): a random forest-based scoring function for improved affinity prediction of protein-ligand complexes. *J. Chem. Inf. Model.,* **2013**, *53*(8), 1923-1933.

http://dx.doi.org/10.1021/ci400120b PMID: 23705795

[44] Zhang, X.; Shen, C.; Jiang, D.; Zhang, J.; Ye, Q.; Xu, L.; Hou, T.; Pan, P.; Kang, Y. TB-IECS: an accurate machine learning-based scoring function for virtual screening. *J. Cheminform.,* **2023**, *15*(1), 63. http://dx.doi.org/10.1186/s13321-023-00731-x PMID: 37403155

[45] Stafford, K.A.; Anderson, B.M.; Sorenson, J.; van den Bedem, H. AtomNet PoseRanker: Enriching Ligand Pose Quality for Dynamic Proteins in Virtual High-Throughput Screens. *J. Chem. Inf. Model.,* **2022**, *62*(5), 1178-1189. http://dx.doi.org/10.1021/acs.jcim.1c01250 PMID: 35235748

[46] Ragoza, M.; Hochuli, J.; Idrobo, E.; Sunseri, J.; Koes, D.R. Protein–Ligand Scoring with Convolutional Neural Networks. *J. Chem. Inf. Model.,* **2017**, *57*(4), 942-957. http://dx.doi.org/10.1021/acs.jcim.6b00740 PMID: 28368587

[47] Volkov, M.; Turk, J.A.; Drizard, N.; Martin, N.; Hoffmann, B.; Gaston-Mathé, Y.; Rognan, D. On the Frustration to Predict Binding Affinities from Protein–Ligand Structures with Deep Neural Networks. *J. Med. Chem.,* **2022**, *65*(11), 7946-7958. http://dx.doi.org/10.1021/acs.jmedchem.2c00487 PMID: 35608179

[48] Chen, L.; Cruz, A.; Ramsey, S.; Dickson, C.J.; Duca, J.S.; Hornak, V. Hidden bias in the DUD-E dataset leads to misleading performance of deep learning in structure-based virtual screening. *Plos one,* **2019**, *14*(8), e0220113. http://dx.doi.org/10.1371/journal.pone.0220113

[49] Morris, C.J.; Stern, J.A.; Stark, B.; Christopherson, M.; Della Corte, D. MILCDock: Machine Learning Enhanced Consensus Docking for Virtual Screening in Drug Discovery. *J. Chem. Inf. Model.,* **2022**, *62*(22), 5342-5350. http://dx.doi.org/10.1021/acs.jcim.2c00705 PMID: 36342217

[50] 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*(3), 947-961. http://dx.doi.org/10.1021/acs.jcim.8b00712 PMID: 30835112

[51] Ashtawy, H.M.; Mahapatra, N.R. Task-Specific Scoring Functions for Predicting Ligand Binding Poses and Affinity and for Screening Enrichment. *J. Chem. Inf. Model.,* **2018**, *58*(1), 119-133. http://dx.doi.org/10.1021/acs.jcim.7b00309 PMID: 29190087

[52] Gabel, J.; Desaphy, J.; Rognan, D. Beware of machine learning-based scoring functions-on the danger of developing black boxes. *J. Chem. Inf. Model.,* **2014**, *54*(10), 2807-2815. http://dx.doi.org/10.1021/ci500406k PMID: 25207678

[53] Méndez-Lucio, O.; Ahmad, M.; del Rio-Chanona, E.A.; Wegner, J.K. A geometric deep learning approach to predict binding conformations of bioactive molecules. *Nat. Mach. Intell.,* **2021**, *3*(12), 1033-1039. http://dx.doi.org/10.1038/s42256-021-00409-9

[54] Stark, H.; Ganea, O.E.; Pattanaik, L.; Barzilay, R.; Jaakkola, T. EQUIBIND: Geometric deep learning for drug binding structure prediction. *arXiv: 2202.05146,* **2023**.

[55] Lu, W.; Wu, Q.; Zhang, J.; Rao, J.; Li, C.; Zheng, S. TANKBind: trigonometry-aware neural networks for drug-protein binding structure prediction. *Proceedings of the 36th International Conference on Neural Information Processing Systems,* 28 Nov- 9 Dec, 2022, New Orleans, LA, USA, pp. 1-14.

[56] Corso, G.; Stark, H.; Jing, B.; Barzilay, R.; Jaakkola, T. DiffDock: Diffusion steps, twists, and turns for molecular docking. **2023**. Available from:https://openreview.net/forum?id=kKF8_K-mBbS(accessed on 2-10-2024)

[57] Zhou, G.; Gao, Z.; Ding, Q.; Zheng, H.; Xu, H.; Wei, Z. Uni-mol: A universal 3d molecular representation learning framework. **2023**. Available from:https://openreview.net/-forum?id=6K2RM6wVqKu(accessed on 2-10-2024)

[58] Buttenschoen, M.; Morris, G.M.; Deane, C.M. PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chem. Sci. (Camb.),* **2024**, *15*(9), 3130-3139. http://dx.doi.org/10.1039/D3SC04185A PMID: 38425520

[59] Brocidiacono, M.; Popov, K.I.; Koes, D.R.; Tropsha, A. PLANTAIN: Diffusion-inspired pose score minimization for fast and accurate molecular docking *arxiv. 2307.12090,* **2023**.http://arxiv.org/abs/2307.12090

[60] Alcaide, E.; Gao, Z.; Ke, G.; Li, Y.; Zhang, L.; Zheng, H. Uni-Mol docking V2: Towards realistic and accurate binding pose prediction. *arxiv. 2405.11769,* **2024**.

[61] Corso, G.; Deng, A.; Fry, B.; Polizzi, N.; Barzilay, R.; Jaakkola, T. Deep Confident Steps to New Pockets: Strategies for Docking Generalization. *arxiv. 2402.18396,* **2024**.

[62] Abramson, J.; Adler, J.; Dunger, J.; Evans, R.; Green, T.; Pritzel, A.; Ronneberger, O.; Willmore, L.; Ballard, A.J.; Bambrick, J.; Bodenstein, S.W.; Evans, D.A.; Hung, C.C.; O'Neill, M.; Reiman, D.; Tunyasuvunakool, K.; Wu, Z.; Žemgulytė, A.; Arvaniti, E.; Beattie, C.; Bertolli, O.; Bridgland, A.; Cherepanov, A.; Congreve, M.; Cowen-Rivers, A.I.; Cowie, A.; Figurnov, M.; Fuchs, F.B.; Gladman, H.; Jain, R.; Khan, Y.A.; Low, C.M.R.; Perlin, K.; Potapenko, A.; Savy, P.; Singh, S.; Stecula, A.; Thillaisundaram, A.; Tong, C.; Yakneen, S.; Zhong, E.D.; Zielinski, M.; Žídek, A.; Bapst, V.; Kohli, P.; Jaderberg, M.; Hassabis, D.; Jumper, J.M. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature,* **2024**, *630*(8016), 493-500. http://dx.doi.org/10.1038/s41586-024-07487-w PMID: 38718835

[63] Xue, M.; Liu, B.; Cao, S.; Huang, X. FeatureDock: Protein-ligand docking guided by physicochemical feature-based local environment learning using transformer *chemrxiv,* **2024**. http://dx.doi.org/10.26434/chemrxiv-2024-dh2rw

[64] Zhang, Z.; He, X.; Long, D.; Luo, G.; Chen, S. Enhancing generalizability and performance in drug–target interaction identification by integrating pharmacophore and pre--trained models. *Bioinformatics,* **2024**, *40*(Suppl. 1), i539-i547. http://dx.doi.org/10.1093/bioinformatics/btae240 PMID: 38940179

[65] de Magalhães, C.S.; Almeida, D.M.; Barbosa, H.J.C.; Dardenne, L.E. A dynamic niching genetic algorithm strategy for docking highly flexible ligands. *Inf. Sci.,* **2014**, *289*, 206-224. http://dx.doi.org/10.1016/j.ins.2014.08.002

[66] Debroise, T.; Shakhnovich, E.I.; Chéron, N. A Hybrid Knowledge-Based and Empirical Scoring Function for Protein–Ligand Interaction: SMoG2016. *J. Chem. Inf. Model.,* **2017**, *57*(3), 584-593. http://dx.doi.org/10.1021/acs.jcim.6b00610 PMID: 28191941

[67] Baek, M.; Shin, W.H.; Chung, H.W.; Seok, C. GalaxyDock BP2 score: a hybrid scoring function for accurate protein–ligand docking. *J. Comput. Aided Mol. Des.,* **2017**, *31*(7), 653-666. http://dx.doi.org/10.1007/s10822-017-0030-9 PMID: 28623486

[68] Li, Y.; Lin, H.; Yang, H.; Yuan, Y.; Zou, R.; Zhou, G. Synergistic application of molecular docking and machine learning for improved binding pose. *Natl Sci Open,* **2024**, *3*(2), 0230058. http://dx.doi.org/10.1360/nso/20230058

[69] Wang, R.; Fang, X.; Lu, Y.; Yang, C.Y.; Wang, S. The PDBbind database: methodologies and updates. *J. Med. Chem.,* **2005**, *48*(12), 4111-4119. http://dx.doi.org/10.1021/jm048957q PMID: 15943484

[70] Cheng, T.; Li, X.; Li, Y.; Liu, Z.; Wang, R. Comparative assessment of scoring functions on a diverse test set. *J. Chem. Inf. Model.,* **2009**, *49*(4), 1079-1093. http://dx.doi.org/10.1021/ci9000053 PMID: 19358517

[71] Mysinger, M.M.; Carchia, M.; Irwin, J.J.; Shoichet, B.K. Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J. Med. Chem.,* **2012**, *55*(14), 6582-6594. http://dx.doi.org/10.1021/jm300687e PMID: 22716043

[72] Su, M.; Yang, Q.; Du, Y.; Feng, G.; Liu, Z.; Li, Y.; Wang, R. Comparative Assessment of Scoring Functions: The CASF-2016 Update. *J. Chem. Inf. Model.,* **2019**, *59*(2), 895-913. http://dx.doi.org/10.1021/acs.jcim.8b00545 PMID: 30481020

[73] Lagarde, N.; Zagury, J.F.; Montes, M. Benchmarking Data Sets for the Evaluation of Virtual Ligand Screening Methods: Review and Perspectives. *J. Chem. Inf. Model.,* **2015**, *55*(7), 1297-1307. http://dx.doi.org/10.1021/acs.jcim.5b00090 PMID: 26038804

[74] Rohrer, S.G.; Baumann, K. Maximum unbiased validation (MUV) data sets for virtual screening based on PubChem bioactivity data. *J. Chem. Inf. Model.,* **2009**, *49*(2), 169-184. http://dx.doi.org/10.1021/ci8002649 PMID: 19434821

[75] Tran-Nguyen, V.K.; Jacquemard, C.; Rognan, D. LIT-PCBA: An Unbiased Data Set for Machine Learning and Virtual Screening. *J. Chem. Inf. Model.,* **2020**, *60*(9), 4263-4273. http://dx.doi.org/10.1021/acs.jcim.0c00155 PMID: 32282202

[76] Scantlebury, J.; Brown, N.; Von Delft, F.; Deane, C.M. Data Set Augmentation Allows Deep Learning-Based Virtual Screening to Better Generalize to Unseen Target Classes and Highlight Important Binding Interactions. *J. Chem. Inf. Model.,* **2020**, *60*(8), 3722-3730. http://dx.doi.org/10.1021/acs.jcim.0c00263 PMID: 32701288

[77] Zhang, X.; Shen, C.; Liao, B.; Jiang, D.; Wang, J.; Wu, Z.; Du, H.; Wang, T.; Huo, W.; Xu, L.; Cao, D.; Hsieh, C.Y.; Hou, T. TocoDecoy: A New Approach to Design Unbiased Datasets for Training and Benchmarking Machine-Learning Scoring Functions. *J. Med. Chem.,* **2022**, *65*(11), 7918-7932. http://dx.doi.org/10.1021/acs.jmedchem.2c00460 PMID: 35642777

[78] Imrie, F.; Bradley, A.R.; Deane, C.M. Generating property-matched decoy molecules using deep learning. *Bioinformatics,* **2021**, *37*(15), 2134-2141. http://dx.doi.org/10.1093/bioinformatics/btab080

[79] 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*(6), 1717-1736. http://dx.doi.org/10.1021/ci500081m PMID: 24708446

[80] Leung, S.; Bodkin, M.; Von Delft, F.; Brennan, P.; Morris, G. SuCOS is better than RMSD for evaluating fragment elaboration and docking poses. *chemrxiv 8100203,* **2019**. http://dx.doi.org/10.26434/chemrxiv.8100203.v1

[81] Wójcikowski, M.; Ballester, P.J.; Siedlecki, P. Performance of machine-learning scoring functions in structure-based virtual screening. *Sci. Rep.,* **2017**, *7*(1), 46710. http://dx.doi.org/10.1038/srep46710 PMID: 28440302

[82] McGibbon, M.; Money-Kyrle, S.; Blay, V.; Houston, D.R. SCORCH: Improving structure-based virtual screening with machine learning classifiers, data augmentation, and uncertainty estimation. *J. Adv. Res.,* **2023**, *46*, 135-147. http://dx.doi.org/10.1016/j.jare.2022.07.001 PMID: 35901959

[83] Adeshina, Y.O.; Deeds, E.J.; Karanicolas, J. Machine learning classification can reduce false positives in structure-based virtual screening. *Proc. Natl. Acad. Sci. USA,* **2020**, *117*(31), 18477-18488. http://dx.doi.org/10.1073/pnas.2000585117 PMID: 32669436