frontiers | Frontiers in Bioinformatics

REVIEW
published: 17 June 2022
doi: 10.3389/fbinf.2022.885983

Scoring Functions for Protein-Ligand Binding Affinity Prediction Using Structure-based Deep Learning: A Review

Rocco Meli[1], *Garrett M. Morris*[2] *and Philip C. Biggin*[1]*

[1]Department of Biochemistry, University of Oxford, Oxford, United Kingdom, [2]Department of Statistics, University of Oxford, Oxford, United Kingdom

The rapid and accurate *in silico* prediction of protein-ligand binding free energies or binding affinities has the potential to transform drug discovery. In recent years, there has been a rapid growth of interest in deep learning methods for the prediction of protein-ligand binding affinities based on the structural information of protein-ligand complexes. These structure-based scoring functions often obtain better results than classical scoring functions when applied within their applicability domain. Here we review structure-based scoring functions for binding affinity prediction based on deep learning, focussing on different types of architectures, featurization strategies, data sets, methods for training and evaluation, and the role of explainable artificial intelligence in building useful models for real drug-discovery applications.

Keywords: drug discovery, scoring function, artificial intelligence, explainable AI, deep learning, neural network, affinity

OPEN ACCESS

Edited by:
Tunca Dogan,
Hacettepe University, Turkey

Reviewed by:
Ahmet Sureyya Rifaioglu,
Iskenderun Technical University, Turkey
Xiaolei Zhu,
Anhui Agricultural University, China
Tingjun Hou,
Zhejiang University, China

***Correspondence:**
Philip C. Biggin
philip.biggin@bioch.ox.ac.uk

Specialty section:
This article was submitted to Drug Discovery in Bioinformatics, a section of the journal Frontiers in Bioinformatics

Received: *28 February 2022*
Accepted: *11 May 2022*
Published: *17 June 2022*

Citation:
Meli R, Morris GM and Biggin PC (2022) Scoring Functions for Protein-Ligand Binding Affinity Prediction Using Structure-based Deep Learning: A Review. Front. Bioinform. 2:885983. doi: 10.3389/fbinf.2022.885983

1 INTRODUCTION

The discovery and development of new small-molecule drugs is a very challenging and expensive process (Drews 2000; Dickson and Gagnon 2004; Schneider and Schneider 2016). Only a handful of new drugs are approved each year (Brown and Wobst 2021), which is minuscule compared to the vastness of chemical space (Reymond et al., 2010) and the billions of dollars poured into drug discovery campaigns (DiMasi et al., 2016). The discovery pipeline for small-molecule drugs usually starts with the identification of a protein target against which a hit compound is identified by high throughput screening (HTS) (Mayr and Bojanic 2009; Macarron et al., 2011). The hit compound is subsequently optimized to obtain a lead compound with good potency and favorable pharmacodynamics and pharmacokinetics properties.

Thanks to significant methodological and hardware advances, computer-aided drug discovery (CADD) has played an important role in the development of new small-molecule drugs over the last decades (Sliwoski et al., 2013). CADD speeds up the early stages of the drug discovery process—hit identification and hit-to-lead optimization—and lowers the costs of these phases by reducing time and experimental resources needed. CADD methods fall into two broad classes: (explicit) structure-based, and ligand-based (or implicit structure-based) methods. For the latter, similarities to known active molecules play an important role since either the protein target is unknown, or information about the protein target is either unavailable or not included. For structure-based methods, the target structure is known and this additional information is exploited in the modelling and optimization of drug-target interactions (DTIs).

One of the main goals in the computational elucidation of DTIs is the calculation of relative or absolute binding free energies to distinguish potent binders from weak binders (or non-binders) against a target of interest. A fast and accurate prediction of protein-ligand binding affinities would circumvent the need for many time-consuming and complex experiments. Rigorous computational methods based on all-atom molecular dynamics simulations in explicit solvent—such as free energy perturbation and thermodynamic integration (Adcock and McCammon 2006)—can compute accurate relative and absolute binding free energies (Bash et al., 1987; Boresch et al., 2003; Mobley et al., 2007; Aldeghi et al., 2016, Aldeghi et al., 2018a; Cournia et al., 2017), predict ligand selectivity (Aldeghi et al., 2017) and mutation effects (Aldeghi et al., 2018b; Hauser et al., 2018), and guide fragment elaborations (Alibay et al., 2022). Unfortunately, such rigorous methods are computationally expensive and often require a lot of expert knowledge and domain expertise (Mey et al., 2020; Hahn et al., 2021). This remains true even for simpler methods such as ligand-interaction energy (LIE) (Åqvist et al., 1994; Jones-Hertzog and Jorgensen 1997). Methods treating the solvent implicitly, such as the Poisson-Boltzmann and generalized Born models (Genheden and Ryde 2015), can offer significant speed increase but sometimes at the expense of accuracy.

The great successes of deep learning (DL) in the fields of computer vision (Voulodimos et al., 2018), natural language processing (NLP) (Young et al., 2018), and other fields of computer science in recent years kick-started the research and application of deep learning in many scientific disciplines including physics, chemistry, biology, and medicine (Baldi 2021). In the field of drug discovery, machine learning (ML) has been in use for a long time, and the potential usefulness of the use of deep learning in virtual screening was identified early on (Unterthiner et al., 2014). The application of modern deep learning architectures to all stages of the drug discovery pipeline is a very active area of research today (Jing et al., 2018; Brown, 2020; Muratov et al., 2020; Jiménez-Luna et al., 2021a; Gaudelet et al., 2021). The main applications in small-molecule drug design consists in the prediction of DTIs, identification of binding sites (Jiménez et al., 2017; Pu et al., 2019; Aggarwal et al., 2021), the generation of novel molecular entities (Schneider and Clark 2019; Meyers et al., 2021), and the prediction of absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties (Huang D. Z. et al., 2021).

Bioactivity prediction can be performed as a classification task—where binders/actives are distinguished from non-binders/inactives—or as a regression task. Machine learning and deep learning scoring functions (SFs) for the prediction of binding affinities (regression) are useful in lead optimization, in contrast with SFs that try to identify binders amongst a large pool of non-binders (classification) and are used in virtual screening to identify a hit. Another task where SFs are commonly used is pose prediction, where near-native poses are distinguished from incorrect poses (classification). Pose prediction and binding affinity prediction are complementary tasks in molecular docking, where a pose is generated and subsequently scored according to the predicted binding affinity.

In this review, we will focus on SFs for binding affinity prediction (inhibition constant K_i or dissociation constant K_d) or binding free energy prediction, but we will inevitably mention related SFs used in pose prediction and virtual screening—which often share the same algorithms and ideas. Recent reviews of structure-based SFs and deep learning for virtual screening are given by Li et al. (2021b), Kimber et al. (2021), and Rifaioglu et al. (2019). Additionally, to narrow the scope of the review, we focus on structure-based deep-learning methods and we refer the reader interested in ligand-based methods to Tropsha (2010), Muratov et al. (2020), Baskin (2020), and Palazzesi and Pozzan (2022). More general and broad reviews about the application of machine learning and deep learning in drug discovery are provided by Chen H. et al. (2018), Vamathevan et al. (2019), and Schneider et al. (2019).

2 CLASSICAL SCORING FUNCTIONS

Historically, SFs for binding affinity prediction and virtual screening have been classified into three categories: force-field-based, empirical, and knowledge-based (Muegge and Rarey 2001; Böhm and Stahl 2002). However, recently Liu and Wang (2015) argued that this historical classification overlooks more recent developments in the field and thus proposed an updated classification scheme with four classes of scoring functions: force-field-based or physics-based, empirical or regression-based, knowledge-based or potential of mean force-based, and descriptor-based or machine learning-based.

This classification is useful to distinguish different methodologies and ideas appearing in the development of SFs. However, some SFs can't be precisely assigned to only one category and the boundary between the four different classes remains rather fuzzy.

In this section we will briefly discuss the first three classes of SFs, often termed "classical" SFs. A good overview of the different SFs can be found in the paper of Liu and Wang (2015)—which proposed the current classification of SFs—and a more recent overview of different SFs used in protein-ligand docking is provided by Li et al. (2019a). While classical scoring functions are still actively developed and refined today, the research focus has certainly shifted to ML/DL based scoring functions.

2.1 Physics-Based (Force-Field Based) Scoring Functions

Physics-based (or force-field-based) SFs use energy terms of a molecular mechanics force-field—whose parameters are determined to reproduce experimental observables or *ab initio* quantum mechanical calculations (Monticelli and Tieleman 2012)—to evaluate protein-ligand interactions. The non-covalent interaction energy between protein and ligand atoms is usually expressed as the sum of van der Waals and electrostatic interaction terms. In their simplest form, such pairwise interactions are represented by a Lennard-Jones potential and Coulomb interaction between point charges.

Different physics-based scoring functions use different potentials to describe van der Waals and electrostatic interactions, depending on the design of the underlying force field. For example, the dielectric constant can be distance-dependent to take into account electrostatic screening due to the solvent and the lower dielectric constant in protein-ligand binding sites (Hingerty et al., 1985; Gilson and Honig 1986; Huang et al., 2010).

Often, additional shorter-range (and sometimes directional) terms are added to account for hydrogen bonding as well as solvation energy and therefore physics-based scoring function can take the following form:

$$\Delta G_{\text{binding}} = \Delta E_{\text{VdW}} + \Delta E_{\text{el}} + \Delta E_{\text{H-bond}} + \Delta G_{\text{sol}}. \quad (1)$$

The solvation energy term can take into account both polar and non-polar contributions. The former accounts for the loss of polar interactions between charged groups and water, while the latter accounts for the desolvation of hydrophobic groups upon binding.

Finally, empirical terms accounting for the loss of torsional degrees of freedom upon complexation can also be included. Oftentimes, simple approximations based on the number of rotatable bonds are used (Böhm 1994; Chang et al., 2007; Huey et al., 2007; Huang and Zou 2010), although more advanced treatments have been suggested (Guedes et al., 2021b). The same corrections are applied to empirical and knowledge-based scoring functions, discussed below.

Force-field-based scoring functions are attractive because of their physical origin and because they can leverage advances in force-field developments, including the latest advances in ML force-fields (Unke et al., 2021). However, describing solvent effects in ligand binding remains an outstanding challenge (Limongelli et al., 2012; Ross et al., 2012; Darby et al., 2019).

Notable examples of physics-based (force field-based) scoring functions are DOCK (DesJarlais et al., 1988; Meng et al., 1992; Shoichet et al., 1992; Ewing et al., 2001; Moustakas et al., 2006; Allen et al., 2015), AutoDock (Goodsell and Olson 1990) and AutoDock 2 (Morris et al., 1996) (AutoDock 3 and AutoDock 4 use hybrid scoring functions (Morris et al., 1998; Huey et al., 2007; Morris et al., 2009)), GoldScore (Jones et al., 1995, Jones et al., 1997), and GalaxyDock (Shin and Seok 2012; Shin et al., 2013).

2.2 Empirical (Regression-Based) Scoring Functions

Empirical or regression-based scoring functions are based on regression analysis to determine the coefficient of different pre-defined terms based on experimental data. This is also what machine learning (or descriptor-based) scoring functions do, however in empirical or regression-based scoring functions the functional form of the scoring function is predetermined and it is often quite simple (such as a linear combination of different contributions) (Ain et al., 2015). As we mentioned previously, the line between the four different classes of scoring functions suggested by Li et al. (2019a) is sometimes blurry.

Empirical scoring functions assuming a linear functional form take the following form (Guedes et al., 2018):

$$\Delta G_{\text{binding}} = w_0 + w_1 \Delta G_{\text{VdW}} + w_2 \Delta G_{\text{H-bond}} + w_3 \Delta G_{\text{entropy}}. \quad (2)$$

The functional form of empirical scoring functions is similar to physics-based scoring functions. However, in empirical scoring functions the parameters **w** are determined by regression analysis—usually multivariate linear regression or partial least squares (Li et al., 2019a)—to reproduce experimentally determined values.

Often, the different terms in empirical scoring functions are simple reward or penalty scores. For example, the ChemScore (Eldridge et al., 1997; Verdonk et al., 2003) scoring function has the following functional form:

$$\text{ChemScore} = w_0 + w_1 S_{\text{hbond}} + w_2 S_{\text{metal}} + w_3 S_{\text{lipo}} + w_4 H_{\text{rot}} + E_{\text{int}} + E_{\text{clash}} + E_{\text{cov}} \quad (3)$$

where S_{hbond} is the score assigned to hydrogen bonds, S_{metal} scores acceptor-metal interactions, S_{lipo} scores lipophilic interactions, H_{rot} describes the loss in conformational entropy upon complexation, E_{int} is the ligand's internal energy, E_{cov} is the covalent energy term, and E_{clash} represents the energetic penalty of clashes between protein and ligand atoms.

One of the first empirical scoring functions was introduced by Böhm (1994) and notable examples include ChemScore (Eldridge et al., 1997; Verdonk et al., 2003), X-Score (Wang et al., 2002), Glide (Friesner et al., 2004, 2006) DockThor (de Magalhães et al., 2014), SFCscore (Sotriffer et al., 2008). More recent scoring functions are Vinardo (Quiroga and Villarreal 2016), Lin_F9 (Yang and Zhang 2021), DockTScore (Guedes et al., 2021a) (combined with ML), and AA-Score (Pan et al., 2022).

A fairly recent review of empirical scoring functions for structure-based virtual screening is provided by Guedes et al. (2018).

2.3 Knowledge-Based (Potential-Based) Scoring Functions

Knowledge-based or potential-based scoring functions are based on pairwise statistical potentials of the form:

$$S = \sum_{i \in \text{lig}} \sum_{j \in \text{prot}} \omega_{ij}(r), \quad (4)$$

where the distance-dependent pairwise potential $\omega_{ij}(r)$ is given by:

$$\omega_{ij}(r) = -k_b T \ln\left(\frac{\rho_{ij}(r)}{\rho_{ij}^0} s\right). \quad (5)$$

$\rho_{ij}(r)$ is the number density of pairs of type *i*-*j* at distance *r* while ρ_{ij}^0 is the same quantity for a reference state where there is no interaction between types *i* and *j* (Muegge and Martin 1999). Therefore, if $\rho_{ij}(r)$ is larger than the reference state ρ_{ij}^0 it contributes favorably to the scoring function while if $\rho_{ij}(r)$ is smaller than the reference state ρ_{ij}^0 then it contributes unfavorably to the scoring function. The pairwise potentials $\omega_{ij}(r)$ are obtained

TABLE 1 | Main data sets providing protein-ligand complexes (crystal structures) and corresponding binding affinities. *N* is the number of protein-ligand complexes (co-crystal structures) with associated binding affinities.

Data Set	*N*	Superset	Website
PDBbind 2020	19 443	—	pdbbind.org.cn
CASF-2016	285	PDBbind 2016	pdbbind.org.cn
CASF-2013	195	PDBbind 2013	pdbbind.org.cn
CASF-2007	195	PDBbind 2007	pdbbind.org.cn
Binding MOAD 2020	15 223	—	bindingmoad.org
CSAR-NCS HiQ	343	Binding MOAD + PDBbind	csardock.org
CSAR-NCS HiQ Update	123	Binding MOAD + PDBbind	csardock.org
Astex Diverse Set	74	—	doi.org/10.1021/jm061277y
BindingDB	11 442	—	bindingdb.org
D3R GC 4	20	—	drugdesigndata.org
D3R GC 3	24	—	drugdesigndata.org
D3R GC 2	36	—	drugdesigndata.org
D3R GC 2015	24	—	drugdesigndata.org

from the analysis of interactions in a large data set of protein-ligand complexes and usually, only pairs of protein and ligand atoms within a certain cutoff are considered ($r < r_{\text{cutoff}}$).

One of the advantages of knowledge-based scoring functions is that entropic and solvation contributions are taken into account implicitly (Muegge and Martin 1999). However, some knowledge-based scoring functions include solvation and entropy effects explicitly (Huang and Zou 2010).

Notable examples of knowledge-based (potential-based) scoring function are the SMoG (DeWitte and Shakhnovich 1996; DeWitte et al., 1997) (later extended to a hybrid knowledge-based and empirical scoring function (Debroise et al., 2017a)), the PMF scoring function developed by Muegge and co-workers (Muegge and Martin 1999; Muegge 2000, Muegge 2001), DrugScore (Gohlke et al., 2000; Velec et al., 2005; Neudert and Klebe 2011; Dittrich et al., 2018), ITScore (Huang and Zou 2006a; Huang and Zou 2006b, Huang and Zou 2010), KECSA (Zheng and Merz 2013), and M-score (Yang et al., 2005). More recent knowledge-based scoring functions are SMoG 2016 (Debroise et al., 2017b), Convex-PL (Kadukova and Grudinin 2017), DLIGAND2 (Chen P. et al., 2019), and KORP-PL (Kadukova et al., 2021).

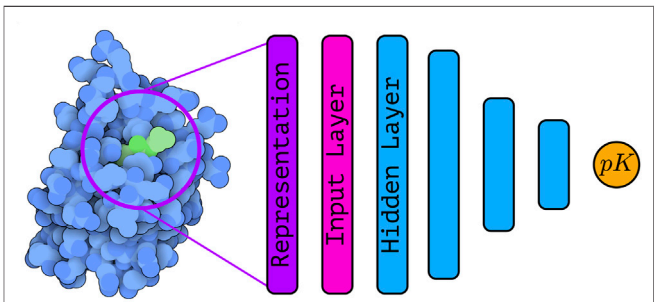

FIGURE 1 | | Schematic representation of a structure-based deep learning architecture—with several hidden layers—for binding affinity prediction. The protein-ligand complex (PDB ID: 5S9H, rendered with Illustrate (Goodsell et al., 2019a)) is encoded into a suitable representation that is used as input to the deep learning architecture. Using a series of stacked hidden layers, a prediction of the binding affinity is finally obtained. The exact nature of the input layer as well as the hidden layers depends on the type of architecture under consideration.

3 DATA SETS

To train ML and DL SFs, high-quality and reasonably large data sets are essential. The success of supervised machine learning and deep learning algorithms strongly depends on the quality and the size of the data set used for training. Thanks to the advances in high-throughput X-ray crystallography and cryo-electron microscopy (cryoEM), the number of available high-resolution structures in the Protein Data Bank (PDB) is constantly increasing (Goodsell et al., 2019b).

In this section, we briefly discuss some of the most common data sets encountered in the training and evaluation of machine learning and deep learning structure-based SFs for binding affinity prediction. The main data sets providing both co-crystal structures and experimental binding affinities are listed in **Table 1**.

3.1 PDBbind

The PDBbind dataset (Wang et al., 2004) is a curated subset of the PDB and it is arguably one of the most common data sets used to train ML and DL SFs for protein-ligand binding affinity prediction. The dataset also contains protein-protein and ligand-nucleic acid complexes.

The origin of the database can be traced back to 2004, when Wang et al. (2004) collected protein-ligand complexes from the PDB (release 103, January 2003) and screened the primary references of the identified complexes to extract binding affinity data (K_d, K_i, IC_{50}).

To train ML and DL SFs, high-quality data is essential—although it has been demonstrated that including lower quality data can improve performance (Li et al., 2015; Francoeur et al., 2020). The PDBbind database is therefore split into a "refined" set and a "general" set (Wang et al., 2004, Wang et al., 2005). The "refined" set is a selection of protein-ligand crystal structures with a resolution of 2.5 Å or lower, where there is a single ligand that is non-covalently bound without significant

steric clashes (Wang et al., 2005). Only systems with associated equilibrium constants K_i and K_d are included in the refined set—IC_{50} values depend on the design of the binding assay—and complexes are filtered to only contain common organic elements.

The same approach was used to build the PDBbind refined set version 2007 (Cheng et al., 2009), but it was improved to produce the PDBbind refined set 2013 and subsequent versions (Li et al., 2014b; Liu et al., 2014, 2017). In addition to the previous criteria used to compile the PDBbind refined set 2007, the complexes added to the PDBbind refined set 2013 satisfy the following additional criteria (Li et al., 2014b): no missing backbone or side chain fragments within 8 Å from the ligand, no extreme values of binding affinity (1 pm < K < 10 mm, where $K = \{K_i, K_d\}$), no multiple binding sites with significantly different binding affinities (> 10 folds difference), no non-standard amino acids within 5 Å from the ligand, and no shallow binders (< 15% of buried ligand surface). The rules for selecting protein-ligand complexes into the PDBbind refined set 2013, together with their rationale and the difference with the rules used for the PDBbind refined set 2007, are very clearly summarized by Li et al. (2014b).

The PDBbind dataset can be downloaded from `pdbbind.org.cn`. The current release (PDBbind 2020) collects binding affinities and structural data for 23 496 biomolecular complexes, 19 443 of which are protein-ligand complexes.

3.1.1 CASF

The CASF benchmarks are a series of comparative assessments of scoring functions originally introduced by Cheng et al. (2009). They evaluate different scoring functions for their performance on scoring, ranking, docking, and screening on a diverse and high-quality set of protein-ligand complexes. Originally employed to compare mostly classical SFs, it has become the *de facto* standard for an initial evaluation of ML and DL SFs (especially for protein-ligand binding affinity prediction).

To test different scoring functions on a diverse and high-quality data set of protein-ligand complexes, a data set is extracted from the PDBbind refined set (where high-quality complexes have already been identified). The PDBbind refined set is clustered according to sequence similarity using BLAST (Altschul et al., 1990), with a similarity threshold of 90% (Cheng et al., 2009). This means that proteins with a sequence similarity higher than 90% are collected in the same cluster since they are likely to represent the same protein or the same protein family.

Once proteins from the PDBbind refined set are clustered by sequence similarity, clusters containing at least four complexes are retained (Cheng et al., 2009). This results in a total of 65 clusters, from which three complexes are sampled: the complex with lower binding affinity, the complex with higher binding affinity, and the complex with binding affinity closer to the mean between the highest and lowest binding affinities (Cheng et al., 2009). This clustered sub-sampling of the PDBbind refined set (called PDBbind core set) results in a total of 65 × 3 = 195 protein-ligand complexes used for the first comparative assessment of scoring functions (CASF-2007).

For the CASF-2013 comparative assessment of scoring functions (Li et al., 2014a), the construction of the PDBbind core set was improved by using the same sequence similarity program used by the PDB, and only clusters with five (and not four) proteins were retained (Li et al., 2014b). Additionally, the best binding affinity has to differ at least 10-fold from the median binding affinity, and the median binding affinity has to differ at least 10-fold from the poorest binding affinity (Li et al., 2014b). The electron density maps of the remaining complexes were visually assessed; if a complex failed at this step, the next best candidate was selected amongst the same cluster (Li et al., 2014b). The final PDBbind core set 2013 still consists of 195 protein-ligand complexes from 65 protein clusters (Li et al., 2014b).

The core set for CASF-2016 (Su et al., 2018) brought additional refinements and more data. As usual, the systems within the high-quality benchmark set are selected from the 4057 protein-ligand complexes in the PDBbind refined set (version 2016). The clustering of complexes based on protein sequence similarity remains the same. However, for CASF-2016, five representatives of each cluster were selected instead of the three selected for CASF-2007 and CASF-2013 (Su et al., 2018). The representative complexes were selected according to their binding affinity: the complex with the lowest binding affinity, the complex with the highest binding affinity, and three complexes distributed as evenly as possible between the lowest and highest binding affinity (Su et al., 2018). The lowest and highest binding affinities differ at least 100-fold and the difference between consecutive binding affinities is at least 1-fold. All ligands were inspected to ensure that there are no identical ligands or stereoisomers (Su et al., 2018). The final PDBbind core set (CASF-2016 benchmark set) consists of 57 × 5 = 285 protein-ligand complexes and it is arguably one of the test sets encountered more frequently in the development of ML and DL SFs.

Unlike the PDBbind data set, the CASF benchmark is not updated annually and therefore the latest release to date remains CASF-2016. The CASF benchmark packages can be downloaded from `pdbbind.org.cn/casf.php`.

It is very common for ML and DL SFs to be trained on the PDBbind refined or general set and subsequently tested on the CASF benchmark set. Recently, non-redundant subsets of the PDBbind refined set were introduced by Boyles et al. (2019) and Su et al. (2020) to evaluate the ability of ML and DL SFs to generalize when removing increasingly dissimilar examples from the training set that have some similarities with the CASF benchmark set.

3.2 Binding MOAD

The Binding MOAD (Mother Of All Databases) (Hu et al., 2005; Benson et al., 2007; Ahmed et al., 2014; Smith et al., 2019) is a subset of the PDB that collects high-quality and biologically relevant crystal structures of protein-ligand complexes together with experimentally determined binding affinities. Ligands available in the Binding MOAD include small peptides (ten amino acids or less), small oligonucleotides (four nucleotides or fewer), small and drug-like organic molecules, and enzymatic cofactors. Crystal structures have a resolution better than 2.5 Å.

As for the PDBbind data set, experimental binding affinities are collected from the primary reference of the deposited PDB structure and consists of only K_i, K_d or IC_{50} values.

The Binding MOAD was first introduced in 2005, containing 5331 protein-ligand complexes from 1780 unique protein families and 2630 unique ligands (Hu et al., 2005). 1375 protein-ligand complexes were associated with binding affinity data spanning 13 orders of magnitude (Hu et al., 2005). The 1780 unique protein families were used to create a non-redundant subset for which 475 complexes have binding affinity data (Hu et al., 2005).

The Binding MOAD is extracted from the PDB as follows (Hu et al., 2005). The full PDB database is screened for high-resolution structures (better than 2.5 Å) excluding theoretical models and NMR structures. Structures containing nucleic acids larger than four nucleotides and peptides longer than ten amino acids were also discarded. Subsequently, complexes with covalently bound ligands as well as invalid ligand structures were filtered out. This reduced database of protein-ligand complexes is hand-curated: the primary citation associated with each structure is screened for binding affinity data while some "suspect ligands" were flagged for visual inspection, resulting in the final database of 5331 protein-ligand complexes.

The Binding MOAD has been expanded annually over the years by adding new protein-ligand complexes deposited on the PDB (together with binding affinity data), resulting in 23 269 total entries and 8156 entries with associated binding affinities in 2015 (Ahmed et al., 2014). In 2019, the Binding MOAD contained 32 747 structures comprising 9117 unique protein families and 16 044 unique ligands.

The Binding MOAD and the PDBbind databases are curated in a similar fashion, to the point that the two data sets could be compared to find and fix disagreements in overlapping systems (Hu et al., 2005). However, the Binding MOAD includes complexes with only binding cofactors, complexes with both a ligand and a cofactor present, and also includes high-quality complexes without binding affinity data (Hu et al., 2005).

Given that the development of the PDBbind was mostly driven by the development of scoring functions (Cheng et al., 2009; Li et al., 2014b,a; Su et al., 2018) while the development of the Binding MOAD was primarily driven by research on protein binding site prediction (Clark et al., 2020) and protein flexibility (Clark et al., 2019), it is more common to encounter the former in the development of ML and DL SFs. However, Binding MOAD can be certainly used for assessing the performance of scoring functions in binding affinity prediction (Xavier et al., 2016) and has been used to build the CSAR dataset discussed below.

3.2.1 CSAR

The CSAR dataset is a data set associated with the Community Structure-Activity Resource (CSAR) which has the goal of collecting high-quality data from both academia and industry to improve docking scoring functions and to organize community-wide assessments of current methods (Dunbar et al., 2011).

The first CSAR data set consisted of protein-ligand complexes from the PDB for which experimental binding affinities (K_i or K_d values) were available in the Binding MOAD database, augmented with data from the PDBbind; Dunbar et al. (2011) describe the CSAR data set as "the best of the PDB [...] augmented with binding data from Binding MOAD and PDBbind". The data set consists of 343 protein-ligand complexes which span binding affinities over several orders of magnitude.

The CSAR data set is subdivided into two subsets: Set 1, and Set 2. Initially, 2916 protein-ligand complexes were identified in the Binding MOAD database (version 2006) and filtered down to 1241 entries according to the quality of the crystal structures. Further processing consisted of the removal of ligands for which hybridization states and bond orders could not be automatically inferred, and for which the experimental binding affinity was expressed in terms of IC_{50} values. This resulted in a total of 309 complexes with associated K_a, K_d and K_i values. Later on, an additional 1228 complexes from Binding MOAD (versions 2007 and 2008) were processed to obtain an additional 230 complexes with associated binding affinity data. After moving some complexes between the two groups to balance physicochemical properties, the final data set representing the initial release consisted of Set 1 (242 entries) and Set 2 (297 entries). Following community feedback, a more stringent quality assessment of the crystal structures was applied, thus reducing the size of the two sets, and errors concerning binding affinities were corrected. Following the CSAR benchmark exercise (Smith et al., 2011), the two sets were further processed resulting in the CSAR-NCS HiQ data set (September 2010), subdivided into Set 1 (176 entries) and Set 2 (167 entries). The CSAR-NCS HiQ data set consists of 52 protein targets with 2 or more structures and 191 targets with a single structure.

The CSAR-NCS HiQ dataset was subsequently updated with an additional 123 structures (set 3) applying the same criteria of the CSAR-NCS HiQ data set to structures in Binding MOAD added between 1/1/2009 and 12/31/2011.

The CSAR-NCS HiQ data set Dunbar et al. (2011), its update, and other data sets associated with the CSAR benchmark exercises (Damm-Ganamet et al., 2013; Dunbar et al., 2013; Smith et al., 2015; Carlson et al., 2016) can be downloaded from csardock.org or bindingmoad.org.

3.3 Astex Diverse Set

The Astex Diverse Set is another common data set encountered in the validation of protein-ligand scoring functions (Hartshorn et al., 2007), alongside the CASF and CSAR benchmarks. This data set contains 85 protein-ligand complexes, most of which are associated with experimentally measured binding potency.

The diverse set was obtained as follows. First, proteins from the PDB database were clustered based on sequence similarity leading to 9188 clusters of distinct proteins. Then, ligands bound to the clustered proteins were then screened to select high-quality structures of pharmaceutical or agrochemical interest and were filtered according to drug-likeliness criteria. The selected protein-ligand complexes were further assessed in terms of ligand clashes with the binding site residues, possible problems related to spurious interactions, and quality of the ligand electron density. This automated filtering procedure resulted in 427 clusters with high-quality protein-ligand complexes.

The final Astex Diverse Set was manually curated from the 427 clusters resulting in 85 complexes. Potency data for 74 of the 85 complexes was finally extracted from the literature.

3.4 Other Data Sets

The data sets described above are curated collections of binding affinities and structures and are therefore useful for the development and assessment of structure-based SFs for protein-ligand binding affinity predictions, both using classical and ML/DL scoring functions. However, there are several other data sets that might be useful to build and assess scoring functions, and some are briefly described below.

ChEMBL (Gaulton et al., 2011; Bento et al., 2013; Mendez et al., 2018) is a manually curated database of bioactive molecules, where data about drug-like molecules are collected together with results from bioactivity assays and genomic information. ChEMBL version 29 (`10.6019/CHEMBL.database.29`) contains data about 21 05 464 compounds and 14 554 targets. While ChEMBL is an extremely valuable resource and provides a large amount of binding affinity data, it does not contain structural data and it is, therefore, more commonly encountered in the development and assessment of ligand-based models (such as in Riniker and Landrum (2013a)).

The bioactivity data in ChEMBL is also exchanged with PubChem Bioassay (Wang et al., 2009, 2011) and BindingDB (Chen et al., 2001; Liu et al., 2007). The PubChem Bioassay database is a public repository containing bioactivity data for small molecules collecting more than 130 million assay results together with their protocols, while the BindingDB is a public database of experimental binding affinities between proteins (8,644 as of 8 November 2021) and drug-like molecules (1,023,385 as of 8 November 2021) which is accessible via a web interface. The BindingDB also contains 5988 protein-ligand crystal structures with associated binding affinity measurements.

Data sets released as part of the Drug Design Data Resources (D3R) Grand Challenges also constitute important datasets on which several ML and DL scoring functions have been designed or tested. D3R Grand Challenges promote the development and benchmarking of computational methods for binding pose and binding affinity prediction, by organizing blinded community challenges using high-quality data sets of pharmaceutical interest. The first D3R Grand Challenge was based on two targets (Gathiaka et al., 2016) using data from industrial drug discovery programs. Subsequent challenges (Gaieb et al., 2017, 2019; Parks et al., 2020) introduced novel targets and associated data for the blind prediction of binding poses, affinity rankings, and relative binding free energies. All the data sets are now easily accessible on the D3R website (`drugdesigndata.org`) as additional test sets for the development and evaluation of ML and DL scoring functions. Interestingly, in the D3R Grand Challenge 3 an increased number of ML methods was observed but overall they did not seem to perform any better than standard methods (Gaieb et al., 2019).

The databases that do not contain target structures are often employed to build ligand-based models or are used to put together new data sets with three-dimensional structures by generating different conformers for the ligand and collecting target structures from the PDB (Bernstein et al., 1977; Berman et al., 2000) to subsequently build structure-based models. For example, Boyles et al. (2021) recently released a new dataset—called the Updated DUD-E Diverse Subset—which combines data from the Directory of Useful Decoys Enhanced (DUD-E) data set (Mysinger et al., 2012) and ChEMBL.

Some data sets for binding affinity prediction discussed above are collected into benchmark data sets such as MoleculeNet (Wu et al., 2018) and Therapeutic Data Commons (Huang K. et al., 2021), which provide much-needed collections for the evaluation of different machine learning and deep learning methods for molecular properties prediction as well as drug discovery and development.

4 MACHINE LEARNING AND DEEP LEARNING SCORING FUNCTIONS

Machine learning (or descriptor-based) scoring functions have been developed and used for decades (Brown 2020). The simplest "scoring functions"—more commonly known as QSAR (Quantitative Structure-activity Relationship) models—were based on a small set of handcrafted descriptors and simple models (such as multiple linear regression Morris et al. (1998); Böhm (1992)), and typically ligand-based. Later, other machine learning (ML) algorithms—such as support vector machines (SVM) (Boser et al., 1992; Cortes and Vapnik 1995), random forests (RFs) (Ho 1995, Ho 1998; Breiman 2001), and gradient boosting (Mason et al., 1999; Friedman 2002)—have been applied in attempt to learn non-linear relationships between descriptors and the binding affinity. **Figure 1** shows a schematic representation of a structure based deep learning architecture for binding affinity prediction.

For an in-depth and rigorous introduction to the deep learning (DL) architectures described below the reader should consult Goodfellow et al. (2016), while classical ML methods are described thoroughly in Bishop (2006), to which we refer the interested reader. For a more hands-on introduction to both ML and DL the reader should consult Géron (2019).

4.1 Descriptors

Descriptors for ML and DL SFs can encode information about the ligand, about the protein, or about intermolecular interactions in the protein-ligand complex. Ligand descriptors are commonly used in cheminformatics applications, quantitative structure-activity relationship (QSAR) modelling, and ligand-based virtual screening (Lo et al., 2018). Ligand-based descriptors can be combined with descriptors for the protein commonly employed in ML-based protein engineering (Xu et al., 2020) to obtain a SF that combines separate information about the ligand and the protein. However, here we focus on structure-based descriptors that encode the protein-ligand complex as a whole and form the basis for structure-based SFs. Methods that combine separate ligand and protein descriptors (van Westen et al., 2011), known as "pair methods" or "proteochemometric models", have

been reviewed by Qiu et al. (2016) and more recently by Kimber et al. (2021).

One common distinction between ML and DL models is that the latter are usually based on a simpler representation and learn descriptors directly from the data; this distinction is however somewhat arbitrary and most DL models still require some pre-processing to convert atom types and coordinates in the correct format for the architecture being used. Here we briefly review structure-based descriptors commonly employed with ML algorithms as well as the input representation used in DL architectures.

One common type of descriptor employed with ML models is an interaction fingerprint (IFP). Structural interaction FPs (SIFts) encode the 3D structure of a protein-ligand complex into a one-dimensional binary vector (Deng et al., 2003). Seven different interaction types involving the ligand and binding site residues are identified and a FP for the whole protein-ligand complex is obtained by concatenating the binding bit string of each binding site residue. Simple ligand-receptor interaction descriptors (SILIRID) are instead obtained from binary IFPs by summing the bits corresponding to the same amino-acids (Chupakhin et al., 2014), thus resulting in a FP with 168 elements (20 amino acids and one cofactor, and 8 interaction types per amino acid). Da and Kireev (2014) developed structural protein-ligand interaction fingerprints (SPLIF), where protein-ligand atom pairs within 4.5 Å are identified and expanded into circular fragments described by extended connectivity fingerprints (ECFPs) (Rogers and Hahn 2010). In this way, protein-ligand interactions are encoded implicitly instead of needing explicit *ad-hoc* interaction classes and therefore can encode all local interactions (Da and Kireev 2014). Similarly, Wójcikowski et al. (2018) developed the protein-ligand extended connectivity (PLEC) FP which combines the ECFP environments of the protein and the ligand atoms in contact to describe protein-ligand interactions. The atomic features employed to construct PLEC FPs—atomic number, isotope, number of neighboring heavy atoms, number of hydrogens, formal charge, ring membership, and aromaticity—are the same used to construct ECFP, but only pairs of interacting atoms within 4.5 Å are considered. The FPs computed for ligand and protein atoms are hashed together to a final bit position (Wójcikowski et al., 2018). The PLEC FP implementation is available as part of the Open Drug Discovery Toolkit (Wójcikowski et al., 2015) and has been successfully used in combination with different ML models for binding affinity prediction (Wójcikowski et al., 2018). There are several other IFPs such as APIF (Pérez-Nueno et al., 2009), PADIF (Jasper et al., 2018), and PyLIF (Radifar et al., 2013).

Ballester et al. (2014) evaluated the impact of the choice of chemical descriptors on ML scoring functions. They showed that more complex descriptors do not necessarily lead to more accurate scoring functions and they identify and discuss the factors that might be contributing to this observation: modelling assumptions, co-dependence of representation and regression method, and data set features.

In structure-based methods, the goal is to exploit the 3D information of protein-ligand complexes. One natural representation of the 3D structure of protein-ligand complexes is the electron density, which is used in X-ray crystallography to model the structure of both the protein and the bound ligand. To encode information about the nuclei available from resolved structures, a representation that clearly encodes the spatial relationship between the protein and the ligand are three-dimensional (3D) grids which discretize volumetric data. The voxel occupancy is often defined as the sum (or maximum) of decaying density functions centered at the different atoms, while atoms of different types are represented in different grids—which can be thought of as a generalization of the RGB channels used in 2D images to represent the different colors. Different representations have been proposed, but they are mostly based on atom-centered density functions $g(r; t_i)$ centered at atom i of type t whose contributions are aggregated together:

$$G(\mathrm{r}; t, \mathrm{R}) = \bigoplus_{i}^{N} g(\|\mathrm{r} - \mathrm{R}_i\|; t_i)\delta_{t_i,t}. \tag{6}$$

$\mathbf{r}$ represents the coordinates of the voxel, $\mathbf{R}_i$ represents the coordinates of atom i, while $\delta_{i,j}$ is Kronecker delta so that only atoms of type t contribute to $G(\mathbf{r}; t, \mathbf{R})$. $\oplus$ is an aggregation function such as sum or maximum.

In Jiménez et al. (2018),

$$g(r; t) = 1 - \exp\left[-\left(\frac{R_t}{r}\right)^{12}\right] \tag{7}$$

and the different channel represent hydrophobic, hydrogen-bond donor/acceptor, aromatic, ionizable, metallic, and excluded volume properties. R_t represents an atom type-dependent characteristic length, often set to the van der Waals radius. The properties are duplicated to represent protein and ligand atoms in different channels, and the density for different atoms in the same channel is aggregated by taking the maximum. In Ragoza et al. (2017a) and subsequent publications (Sunseri et al., 2018; Francoeur et al., 2020) the following functional form is used:

$$g(r; t) = \begin{cases} e^{-2r^2/R_t^2} & 0 \le r < R_t, \\ \dfrac{4}{e^2 R_t^2} r^2 - \dfrac{12}{e^2 R_t} r + \dfrac{9}{e^2} & R_t \le r < 1.5R_t, \\ 0 & r \ge 1.5R_t, \end{cases} \tag{8}$$

and the different channels represent the different atom types from the SMINA docking software (Koes et al., 2013), resulting in 16 channels for the receptor and 18 channels for the ligand (Ragoza et al., 2017a). Contributions from different atoms on the same channel are summed together.

The advantage of using 3D grid representations is that they encode clear spatial relationships between the different atom types and the computation can be performed very efficiently (Sunseri and Koes 2020) thus allowing on-the-fly data augmentation during training. However, grid representations have also several limitations. While computation of $G(\mathbf{r}; t)$ can be performed very efficiently, their dependence on the coordinate frame requires extensive data augmentation (Ragoza et al., 2017a) at increased computational costs, and the sparsity of some channels (such as the ones representing halogens or metals)

implies wasteful computations. Additionally, the memory footprint of grid-based representations increases with the number of atom types. Despite the limitations, the close connection to the field of computer vision has led to the successful development of several SFs based on this representation, as discussed below.

For graph-based models such as graph neural networks (GNNs), descriptors are associated to atoms—the nodes of the graph—and bonds—the edges of the graph. A node descriptor is a vector containing information about the atom. An edge descriptor is a vector describing the chemical bond—such as the bond order. There are several descriptors employed in the literature, and they depend on the task at hand. For protein-ligand binding affinities, simple quantities related to an atom or a bond are commonly employed since higher-level features are learned by intermediate GNN layers (Feinberg et al., 2018). Such simple features for the nodes can include one-hot-encoded elements or atom types, formal charges, hybridization states, aromaticity, and other atomic properties (Jiang et al., 2021). Edge features can include both 2D and 3D information such as bond order, conjugation, bond length, and other bond properties (Jiang et al., 2021).

Descriptors commonly used in ML/DL for quantum chemistry have been successfully applied to the classification of active and decoys against different protein families (Bartók et al., 2017). Recently, the smooth overlap of atomic position (SOAP) descriptor (Bartók et al., 2013)—which allows comparing molecules across the structural and chemical space (De et al., 2016)—have been used together with Gaussian processes models to predict $p\mathrm{IC}_{50}$ values (McCorkindale et al., 2020). At the same time, atomic environment vectors developed for the ANI family of neural network potentials (Smith et al., 2017; Gao et al., 2020) and based on Behler-Parrinello symmetry functions (Behler and Parrinello 2007) have been used as descriptors of protein-ligand complexes for binding affinity predictions (Meli et al., 2021). Behler-Parrinello symmetry functions have also been employed as node features in GNNs for binding affinity prediction (Karlov et al., 2020) and inspired the atomic convolution architecture from Gomes et al. (2017). Both descriptors are strongly related (Musil et al., 2021) and provide a local descriptor of the structural and chemical environment of atoms in a way that is translationally and rotationally invariant.

Learned molecular representations also play an important role as descriptors (Chuang et al., 2020; Menke and Koch 2021). The characteristic of deep learning architecture is that useful and efficient internal representations are learned directly from the input data. Therefore, the fixed and *ad-hoc* descriptors or fingerprints described above can be substituted with learned representations. Yang et al. (2019) performed an extensive analysis of learned molecular representation for property predictions, showing that they achieve similar or better performance than fixed descriptors. While many learned representations for computational chemistry include only 2D information, learned representation for three-dimensional structures have been developed (Kuzminykh et al., 2018) but their application in structure-based drug discovery is still under-explored. The interest in DL architectures is that they can leverage the simple inputs described above (such as 3D atomic densities or coordinates and atom types) to automatically learn internally complex representations that can be used for molecular property prediction.

Some authors extracted descriptors from molecular dynamics (MD) trajectories, instead of using the crystal structure or docked poses, although the use of trajectory data remains rare (Wang and Riniker 2020). Yakovenko and Jones (2017) use atomic densities but trained their model on both docked poses and MD trajectory frames to obtain learned representations later used to predict LIE. Berishvili et al. (2019) developed 1D descriptors based on GROMACS (Berendsen et al., 1995; Abraham et al., 2015), AutoDock Vina (Trott and Olson 2009), and SMINA (Koes et al., 2013) terms to describe frames from MD trajectories. The descriptor for each frame where stacked together into a matrix of size $n_{\mathrm{descriptor}} \times n_{\mathrm{frames}}$, representing the whole MD trajectory.

A more in-depth overview of featurization strategies for protein-ligand interactions that are commonly employed in the development of ML and DL SFs is given by Xiong et al. (2021), while an overview of common molecular representations used in AI-driven drug discovery is provided by David et al. (2020).

4.2 Overview of Classical Machine Learning Scoring Functions

Classical ML algorithms such as SVMs and RFs have been used in quantitative structure-activity relationship (QSAR) modelling and in the development of structure-based scoring functions for a while (Ain et al., 2015; Muratov et al., 2020).

One of the earliest ML SFs for binding affinity predictions has been developed by Deng et al. (2004). The model combines protein-ligand atom pair occurrence and distance-dependent atom pair features with a kernel partial least squares method (K-PLS) (Rännar et al., 1994, Rännar et al., 1995) to predict pK_d, demonstrating that structure-based descriptors combined with ML regression can be effective for protein-ligand binding affinity prediction on different complexes. Das et al. (2010) introduced property-encoded shape distribution signatures—descriptors encoding molecular shapes and property distributions on protein and ligand surfaces—which were used in combination with SVM to build a regression model. SVM-based regression was also used by Li et al. (2011) to develop two SFs, one based on knowledge-based potentials (SVR-KB) and another based on physicochemical properties of the protein-ligand complex (SVR-EP). Both SFs show very good performance on the CASF benchmark when compared to classical SFs.

RF models have been quite successful in the development of structure-based ML SFs. Ballester and Mitchell (2010) introduced a novel SF based on RF, called RF-Score. Protein-ligand complexes are described by a 36-dimensional feature vector storing the occurrence count of different protein-ligand atom pairs within a cutoff of 12 Å. The feature vector is used as input of a RF regression model predicting the binding affinity. Thanks to the use of the PDBbind benchmark (Cheng et al., 2009), RF-Score could be easily compared to 16 other SFs, showing that RF-Score

is a very competitive scoring function. SFCScore^{RF} improved the performance of RF-Score by using a different and larger feature vector including ligand-based features (such as the number of rotatable bonds) and interaction-specific descriptors (Zilian and Sotriffer 2013).

Gradient boosting (Mason et al., 1999; Friedman 2002)—often combined with decision trees—is another popular ML technique used in the development of SFs, also thanks to the availability of high-quality open-source implementations such as XGBoost (Chen and Guestrin 2016) and LightGBM (Ke et al., 2017). Notable scoring functions based on gradient boosting are XGB-Score (Li et al., 2019b), AGL-Score (Nguyen and Wei 2019), and OPRC-GBT (Wee and Xia 2021). Shen et al. (2021) recently developed several XGBoost-based classifiers to assess the impact of cross-docked poses on the performance on pose-prediction, highlighting the importance of cross-docked poses for training of ML SFs with a broad applicability domain and increased robustness for pose-prediction.

Instead of learning the experimental protein-ligand binding affinity directly, Wang and Zhang (2016), used a RF model learning to correct the AutoDock Vina scoring function (Trott and Olson 2009), which represent a reasonable baseline—especially for docking and virtual screening. The resulting scoring function, called Δ_{Vina}RF, retains the very good scoring performance of other ML SFs on scoring and ranking while also working well for docking and virtual screening. Δ_{Vina}RF showed great performance in the CASF-2016 benchmark (with a Pearson's r of 0.82 in the scoring task), but this superior performance can be partially attributed to the overlap between the training set and the CASF-2016 test set (Su et al., 2018).

ML-based scoring functions are still under active development both in terms of methodology and training data. For example, Boyles et al. (2019) showed that including ligand features obtained with RDKit into structure-based ML scoring functions consistently improves the performance in protein-ligand binding affinity prediction. Combining features from RF-Score (v3) with RDKit molecular descriptors improves Pearson's correlation for the CASF-2016 scoring benchmark from 0.79 to 0.84 (Boyles et al., 2019). Another example of recently developed scoring function using classical machine learning regression models for binding affinity prediction is RASPD+ (Holderbach et al., 2020).

Several other classical machine learning algorithms such as kernel ridge regression, Gaussian processes (Williams and Rasmussen 1996; Rasmussen 2003), and other methods have been used in the development of structure-based scoring functions but they are not the focus of this review. The interested reader can consult Ain et al. (2015) and (Li H. et al., 2020) for a more in-depth review of machine learning scoring functions.

4.3 Feed-Forward Neural Networks

Feed-forward neural networks (also known as multilayer perceptrons (MLPs), fully-connected neural networks, artificial neural networks (ANNs), or simply neural networks (NNs)) consist in a series of linear layers combined with point-wise non-linearities called activation functions (Bishop 2006). Originally, feed-forward neural networks were inspired by the way neurons in the brain work (McCulloch and Pitts 1943; Widrow and Hoff 1960; Rosenblatt 1962).

The basic unit of a neural network is a "neuron" (perceptron, or node) and the neurons in a neural network are clustered in different layers that are stacked. The neuron j in layer k takes an input vector $\mathbf{x} \in \mathbf{R}^N$ returns an output

$$z_j^{(k)} = g\left(\sum_i^N w_{ji}^{(k)} x_i + b_j^{(k)}\right), \tag{9}$$

where $w_{ji}^{(k)}$ (weights) and $b_j^{(k)}$ (biases) for neuron j in layer k are learnable parameters to be determined during training and where $g(\cdot)$ is a non-linear function, called activation function. Neural networks are very expressive and can be regarded as universal approximators (Hornik et al., 1989), provided a large enough number of hidden neurons and some classes of activation functions (Bishop 2006).

Initially, neural networks were composed only of a small number of neurons with a single (hidden) layer between the input layer and the output layer but thanks to the development of algorithms able to train neural networks with multiple layers in a simple and efficient way (Rumelhart et al., 1986) neural networks became deeper and deeper (now called deep neural networks, DNNs) by staking together multiple hidden layers.

The use of simple and deep NNs for the determination of quantitative structure-activity relationships (QSAR) is not new (Salt et al., 1992; Dahl et al., 2014; Ma et al., 2015). One of the first use of NNs in binding affinity prediction was published by Artemenko (2008), where a subset of physicochemical descriptors and quasi-fragmental descriptors—describing pairwise statistics of interatomic distances—were selected using multiple linear regression and used as input of a feed-forward NN. NNs have been also successfully used for classification of actives and decoys. Durrant and McCammon (2010) introduced a NN-based SF—NNScore—to distinguish between well and poorly docked ligands as well as actives from decoys. NNScore was later extended to regression of binding affinities in NNScore 2.0 (Durrant and McCammon 2011b) thus providing a direct estimation of pK_d. NNScore 2.0 uses terms from the Vina scoring function (Trott and Olson 2009)—to encode steric, hydrophobic, and hydrogen-bonding interactions–as well as BINANA features (Durrant and McCammon 2011a) as input and returns a estimate of pK_d as output.

Ashtawy and Mahapatra (2015) used a collection of NNs whose predictions are combined with the bagging (Breiman 1996)—bootstrap aggregation—or boosting (Freund and Schapire 1997; Friedman 2002) ensemble methods. The input features were obtained as a combination of classical scoring functions terms, together with features from RF-Score. Their BgN-Score and BsN-Score SFs perform significantly better on the PDBbind core set 2007 than classical SF and surpass SFs based on RFs.

Wójcikowski et al. (2018) showed that a MLP combined with their PLEC fingerprint can achieve very good performance on the CASF-2016 benchmark. However, they also show that the PLEC FPs perform equally well when using a simpler linear model

instead of a neural network, confirming that well-crafted descriptors can be extremely powerful.

More recently, Zhu et al. (2020) developed a model for pK_d prediction where pairwise contributions are computed with a fully connected NN. Trained on the PDBbind 2018, they achieve a Pearson's correlation coefficient of 0.75 and a RMSE of 1.44 on the CASF-2016 benchmark but the authors point out that there is a significant overlap between the test and training sets which might be boosting the performance of their model. Meli et al. (2021) used a collection of MLPs combined with a local representation of the atomic environment to predict protein-ligand binding affinities, reaching good performance on the CASF-2016 benchmark.

4.4 Convolutional Neural Networks

Convolutional neural networks (Fukushima 1980; Le Cun et al., 1989; Lecun et al., 1998; Krizhevsky et al., 2017) are a class of neural networks that tries to overcome some of the limitations of feed-forward neural networks, by using convolution operations instead of matrix multiplication in some of their layers (Goodfellow et al., 2016). Feed-forward neural networks use a one-dimensional vector as input which prevents the encoding of spatial relationships, and uses many parameters. CNNs are based on three main concepts (Bishop 2006): local receptive fields (inspired by the structure of the visual cortex (Hubel 1959; Hubel and Wiesel 1959)), weight sharing, and subsampling.

Local receptive fields are implemented in convolutional layers, where neurons in a layer do not receive the output of all neurons in the previous layer (as in fully-connected NNs) but only the ones in their local receptive field (Géron 2019). For two-dimensional grid-based inputs (such as images), the output of neuron at location (i, j) of feature map k of the convolutional layer l is given by (Géron 2019)

$$z_{i,j,k}^{(l)} = b_k^{(l)} + \sum_{u=1}^{f_h^{(l)}} \sum_{v=1}^{f_w^{(l)}} \sum_{k^{(l-1)}=1}^{f_n^{(l-1)}} x_{i^{(l-1)},j^{(l-1)},k^{(l-1)}} \cdot w_{u,v,k^{(l-1)},k}^{(l)}, \tag{10}$$

with

$$\begin{cases} i^{(l-1)} = us_h^{(l)} + f_h^{(l)} - 1, \\ j^{(l-1)} = us_w^{(l)} + f_w^{(l)} - 1. \end{cases} \tag{11}$$

f_h and f_w are the height and the width of the receptive field (i.e. the size of the 2D convolutional kernel) while s_h and s_w represent the strides (i.e. the size of the displacement of the receptive field). $f_n^{(l-1)}$ denotes the number of feature maps in the previous layer $(l-1)$. b_k^l is a bias term associated to feature map k while $w_{u,v,k',k}^{(l)}$ denotes the weight term associated to the connection between the input located at (u, v) in feature map $k^{(l-1)}$ (relative to the neuron's receptive field) and the neuron in feature map k of layer l. Both b_k^l and $w_{u,v,k',k}^{(l)}$ are learnable parameters to be determined during training. For clear depictions of the main building blocks of 2D CNNs we refer the reader to Dumoulin and Visin (2016).

Parameter sharing in a convolutional network comes from the fact that each weight $w_{u,v,k',k}^{(l)}$ of the kernel is used at every position of the input, avoiding the need to learn a parameter for each input element as it is the case in MLPs. Parameter sharing does not reduce the computational complexity of the forward pass, but significantly reduces the number of parameters in the network (when the size of the convolutional kernel is much smaller than the size of the input) and therefore the associated memory footprint (Goodfellow et al., 2016).

Pooling layers—such as maximum pooling (Zhou et al., 1988), and average pooling—are often inserted after (activated) convolutional layers to make the representation approximately invariant to small translations (Goodfellow et al., 2016). Additionally, they reduce the size of the input of the next layer thus increasing the computational efficiency of the CNN, and are essential for dealing with inputs of varying size (Goodfellow et al., 2016).

Convolutional neural networks have been very successfully applied to different tasks in computer vision such as image classification (Krizhevsky et al., 2017) in the ImageNet challenge (Deng et al., 2009; Russakovsky et al., 2015).

Wallach et al. (2015) introduced a structure-based deep convolutional network for bioactivity prediction (classification into two activity classes) of small drug-like molecules against a target of interest. In their architecture, denoted AtomNet, the protein-ligand binding site was converted into a 3D grid (20 Å per side at 1 Å resolution) containing values related to structural features such as the number of atom types or protein-ligand descriptors such as SPLIF (Da and Kireev 2014), SIFt (Deng et al., 2003), or APIF (Pérez-Nueno et al., 2009). They showed improved performance in the ROC-AUC compared to their baseline, provided by the SMINA docking software (Koes et al., 2013). Ragoza et al. (2017a) introduced a similar approach to distinguish good (low RMSD) from bad (high RMSD) docking poses using CNNs based on an atomic density representation of the binding site (see **Eq. 8**). This approach was later extended to include binding affinity predictions in a multitask learning framework (Sunseri et al., 2018)—both the binding affinity and the pose quality are predicted at the same time—and it was shown to provide a good correlation between experimental and predicted binding affinities for the CASF-2016 benchmark (Francoeur et al., 2020). The various pre-trained CNN scoring functions are integrated and readily available in the GNINA docking software (McNutt et al., 2021). Jiménez et al. (2018) took a similar approach—with a slightly different density representation, first introduced in DeepSite Jiménez et al. (2017)—for binding affinity prediction with their K_{deep} architecture and they achieved a very good correlation and low RMSE on the CASF-2016 benchmark. Interestingly, they analyzed the accuracy separately for the 58 different target classes of the CASF-2016 benchmark, revealing that accuracy is very sensitive to the specific protein used. Indeed, protein family-specific CNN models have been developed for virtual screening using a transfer-learning approach (Imrie et al., 2018).

Many other architectures for binding affinity predictions based on CNNs have been developed in recent years. Notable examples are Pafnucy (Stepniewska-Dziubinska et al., 2018), DeepAtom (Li Y. et al., 2019), OnionNet (Zheng et al., 2019; Wang S. et al., 2021) and OnionNet-2 (Wang Z. et al., 2021), and RoseNet (Hassan-Harrirou et al., 2020).

Pafnucy discretizes the binding site on a three-dimensional grid of 20 Å in side at 1 Å resolution and employs a set of 19 features including one-hot encoding of atom types (including selenium, halogens, and metals), hybridization state, number of bonds with heavy atoms, number of bonds with heteroatoms and a flag distinguishing protein and ligand atoms. DeepAtom uses a grid of 1 Å resolution to voxelize the binding site, with the same density representation of Jiménez et al. (2018) and using Arpeggio atom types (Jubb et al., 2017), but the architecture is inspired from ShuffleNet V2 (Ma et al., 2018). OnionNet (Zheng et al., 2019) also uses a deep convolutional neural network but the input features are based on intermolecular element-pair-specific contacts between ligand and protein atoms, which are grouped in different distance shells. Each shell is described by 64 features representing the intermolecular interactions—within the shell boundaries—between the protein and ligand for eight atoms types considered, and a total of 60 shells (of thickness 0.5 Å) is employed (Zheng et al., 2019). This idea was later extended in OnionNet-2 (Wang Z. et al., 2021), which uses protein residues types instead of protein atom types (increasing the number of features from 64 to $8 \times 21 = 168$). RoseNet (Hassan-Harrirou et al., 2020) uses an ensemble of CNNs—based on the ResNet architecture (He et al., 2016)—combining molecular mechanics energies from the Rosetta force field (Alford et al., 2017) voxelized onto a 3D grid (25 Å each side, at 1 Å resolution) and molecular descriptors—using an approach similar to K_{deep} with descriptors from AutoDock Vina (Trott and Olson 2009)—to predict absolute binding affinities.

CNNs can also be employed with lower-dimensional descriptors. For example, TopologyNet (Cang and Wei 2017) encodes the three-dimensional protein-ligand complex structure into one-dimensional element-specific fingerprints based on topological invariants. Such element-specific topological fingerprints, stacked together over multiple channels—like a one-dimensional image representation—are then used as input of a CNN, and achieve good performance on the CASF-2016 benchmark. The work was later extended to explore additional algebraic topology approaches (Cang et al., 2018).

CNNs have also been successfully applied to the related task of pose prediction. The CNN developed by Ragoza et al. (2017a) has been developed initially for pose prediction, and it was extended to binding affinity prediction on a later stage (Francoeur et al., 2020). Other notable examples are DeepBSP (Bao et al., 2021), which uses a 3D voxel representation of protein-ligand complexes to predict the RMSD between a docked ligand and its native pose—an idea previously explored by Aggarwal and Koes (2020)—and MedusaNet (Jiang H. et al., 2020), which uses CNNs to predict if a pose generated by docking is a good pose to stop the docking process earlier when k good poses are found thus reducing computational costs.

The application of CNNs in the prediction of protein-ligand binding affinities has been quite successful, as demonstrated by the methods discussed above. However, while CNNs are translational invariant they are not rotationally invariant and therefore require extensive data augmentation where the protein-ligand complex is randomly rotated before computing its associated grid representation. Data augmentation with CNNs has proven to be essential to prevent overfitting in pose prediction (Ragoza et al., 2017a), and the average over multiple random rotations can be used during inference thus reducing the variance of the predictions (Jiménez et al., 2018). Many concepts from geometric deep learning (Atz et al., 2021; Bronstein et al., 2021), such as CNNs that are equivariant to rigid body motions (Weiler et al., 2018), will spill more and more into the field of protein-ligand binding affinity prediction as well as virtual screening to overcome some of the limitations of standard CNNs by encoding relevant symmetries directly into the model.

4.5 Graph Neural Networks

Graph neural networks (GNNs) are a collection of DL architectures to work with data that can be represented as a graph (Bronstein et al., 2021). The vast majority of GNNs falls under three categories (Bronstein et al., 2021): convolutional (Defferrard et al., 2016; Kipf and Welling 2016), attentional (Monti et al., 2017; Veličković et al., 2017; Zhang et al., 2018), and message-passing (Gilmer et al., 2017; Battaglia et al., 2018). A graph $\mathcal{G}(V, E)$ is composed of a set of vertices $v_i \in V$ and a set of edges $e_{ij} \in E$ connecting the vertices. Features $\mathbf{x}$ are associated to vertices (and, optionally, edges) and such features are subsequently updated as follows:

$$\mathrm{h}_u = \phi\left(\mathrm{x}_u, \bigoplus_{v \in \mathcal{N}_u} \psi\left(\mathrm{x}_u, \mathrm{x}_v\right)\right) \tag{12}$$

where ϕ and ψ are learnable functions (often learnable affine transformations with activation functions (Bronstein et al., 2021)) and where $\oplus$ represents a permutation-invariant function allowing the aggregation of features (such as sum, mean, and maximum (Bronstein et al., 2021)) over the neighborhood $\mathcal{N}_u$ of node u. ψ is a message-passing function (which can be generalized to include edge features as well), while ϕ is a vertex update function. It is possible to learn edge features as well by introducing a hidden representation $\mathbf{h}_{uv}$ for the edges (Kearnes et al., 2016; Gilmer et al., 2017).

Since molecules can be naturally represented as graphs—with nodes in the graphs representing different atoms and edges in the graph representing the chemical bonds between such atoms—GNNs are well suited to be applied in the field of chemistry (Atz et al., 2021). Message-passing GNNs, which are the most general flavor, have been successfully applied in quantum chemistry applications (Schütt et al., 2018; Qiao et al., 2020; Schäfer et al., 2020; Christensen et al., 2021). GNNs have also been applied to several molecular property predictions (Gaudelet et al., 2021), including bioactivity and protein-ligand binding affinity.

Gomes et al. (2017), inspired by the work of Behler and Parrinello (2007), developed an atom type convolution that uses a neighbor-listed distance matrix to automatically extract features about local chemical environments and combine this information with radial pooling to downsample the output of the atom type convolution. Essentially, the atom type convolution performs a graph convolution on the nearest neighbors graph in three-dimensional space. The resulting features are then passed to a collection of fully connected layers (all with the same weights

and biases) to predict atomic contributions to the energy, which are summed together to obtain the total Gibbs free energy. To predict the binding free energy, three weight-sharing networks are used (one each for G_{complex}, G_{protein} and G_{ligand}) and the results are then combined as

$$\Delta G_{\text{complex}} = G_{\text{complex}} - G_{\text{protein}} - G_{\text{ligand}} \tag{13}$$

so that the whole architecture directly incorporates the thermodynamic cycle.

In PotentialNet (Feinberg et al., 2018) the node updates are of the form

$$h_u^{(k)} = \text{GRU}\left(h_u^{(k-1)}, \sum_e^{N_{\text{et}}} \sum_{v \in \mathcal{N}^e(v_i)} \text{NN}^e\left(h_v^{(k-1)}\right)\right) \tag{14}$$

where GRU is a gated recurrent unit (Hochreiter and Schmidhuber 1997; Cho et al., 2014; Chung et al., 2014), NN^e is a trainable NN for edge type e, and $\mathcal{N}^e(v_i)$ denotes the neighbors of edge type e for atom i. Several updates are concatenated into different stages: in the first stage information is propagated only between nodes linked by a covalent bond, in the second stage information is propagated between non-covalent and covalent bonds and finally, everything is aggregated by a ligand-based graph gather. The first step essentially produces learned (bond-based) atom types, while the second step includes both bond and spatial information between the atoms (Feinberg et al., 2018). In stage three, all learned features for the ligand atoms are summed together and the resulting vector is used as input of a fully-connected neural network to produce the final prediction.

The graphDelta architecture uses a graph-based representation for the ligand and incorporates information about the target in the node features (Karlov et al., 2020). The node features represent radial and angular Behler-Parrinello atom centered symmetry functions (ACSFs) (Behler and Parrinello 2007), combined with a message-passing neural network. With enough training epochs, they achieve a Pearson's correlation coefficient of 0.87 and a RMSE of 1.05 in the CASF-2016 benchmark for binding affinity prediction.

Li et al. (2021) developed a structure-aware interactive GNN which combines polar coordinate-inspired graph attention layers and pairwise interactive pooling. The graph attention layers leverage distances between nodes and angles between edges to iteratively update node and edge embeddings while preserving distance and angle information among atoms. The pairwise atomic type-aware pooling layer is then used to gather interactive edges to capture long-range interactions. Their model, called SIGN, achieves good results on the CASF-2016 benchmark for binding affinity prediction as well as the CSAR-NRC HiQ set.

Son and Kim (2021) developed GraphBAR, where a graph is constructed from all ligand atoms and protein atoms within 4 Å from the ligand (limited to a maximum of 200 nodes, with zero-padding of the adjacency matrix for smaller graphs). Node features consist of one-hot encoded atom types, atom hybridization states, number of neighboring atoms (heavy atoms and heteroatoms), and well as partial charges, stored in a 200, ×, 13 feature matrix. Multiple binary adjacency matrices are used to encode different interaction shells with fixed distance intervals. A graph convolution block is applied to each adjacency matrix together with the feature matrix pre-processed by a fully-connected layer. The outputs of the graph convolutional blocks are concatenated and a fully connected layer produces the final prediction. The model shows similar performance to Pafnucy (Stepniewska-Dziubinska et al., 2018), but the training time appears to be considerably shorter (Son and Kim 2021).

Jiang et al. (2021) developed InteractionGraphNet, where two independent graph convolution modules are stacked to sequentially learn intramolecular and intermolecular interactions using three molecular graphs (one for the ligand, one for the protein, and one for the protein-ligand complex). The protein-ligand bipartite graph is built using protein and ligand atoms within 8 Å of each other. At first, a series of message passing iterations is employed to update the node features in the protein and ligand graphs. Then, these learned node features are used as initial node features for the protein-ligand graph on which edge features representing non-covalent interactions are updated. The learned edge features on the protein-ligand graph, representing the non-covalent interactions between the protein and the ligand, are finally pooled together and used for downstream prediction tasks: binding affinity prediction, virtual screening and pose prediction. For binding affinity prediction, InteractionGraphNet shows good results on the CASF-2016 benchmark, although several systems were removed from the test set.

Moesser et al. (2022) recently developed a simple but effective way to include protein-ligand interactions into ligand-based graphs. Their protein-ligand interaction graphs (PLIGs) representation featurize an atom node in the molecular graph by including both atom properties and atom-atom contacts with protein atoms. Combined with the GAT architecture (Veličković et al., 2017), their model reaches a very good performance on the CASF-2016 benchmark.

Moon et al. (2022) used GNNs in a very interesting way. Instead of using standard and general architectures, Moon et al. (2022) included parametrized physics-based equations in the model architecture, to incorporate the appropriate inductive bias with the goal of improving model generalization by forcing the model to learn the underlying chemical interactions. A GNN is used to update node features across covalent bonds and intermolecular interactions, which are then used—together with pairwise distances—as input of physics-based parametrized equations describing intermolecular interactions as well as entropy loss. The parameters of the physics-informed equations are learned during training and contribute to model generalization.

GNNs have been also successfully applied for structure-based virtual screening (classification of actives and decoys) as well as pose prediction (classification of binding poses), as demonstrated by Lim et al. (2019), Morrone et al. (2020), and Stafford et al. (2022). The use of GNNs—and, more generally, geometric deep learning—in drug discovery and drug development is a very

active area of research and a recent overview on several different applications beyond the narrow scope of this review is given by Gaudelet et al. (2021).

4.6 Other Methods

Above we briefly described widely used families of deep learning architectures—MLPs, CNNs, and GNNs—and their application on the development of structure-based scoring functions. One important omission is recurrent neural networks (RNNs) (Rumelhart et al., 1986; Hochreiter and Schmidhuber 1997; Graves 2012), which are suited to learn from sequential data (such as language or time series). RNNs are also applied to protein-ligand binding affinity prediction (Karimi et al., 2019) but they usually employ unrelated representation for the protein (often the sequence of amino acids) and the ligand (SMILES strings or related representations). As mentioned above, proteochemometric or pair models (Lenselink et al., 2017; Feng et al., 2018; Öztürk et al., 2018; Shin et al., 2019; Jiang M. et al., 2020; Nguyen et al., 2020; Yang et al., 2022) are outside the scope of this review and the reader can find more information in Kimber et al. (2021).

Similarly to proteochemometric models, which combine different—often learned—representations for the protein and the ligand, protein-ligand binding affinity predictions can also benefit from the use of complementary representations of the complex. Jones et al. (2021) combine learned representations of the protein-ligand complex obtained with CNNs and GNNs using mid-level or late deep fusion (Roitberg et al., 2019).

Seo et al. (2021) recently developed BAPA, an architecture based on 1D CNNs combined with an attention layer. The protein-ligand complex is encoded into a 1D descriptor of contacts between the protein and ligand atoms and processed using a 1D CNN to obtain learned features, which are then concatenated with terms from the AutoDock Vina scoring function. The learned features are then encoded into a latent representation using a MLP. The encoded vector is then passed to an attention layer. As described by Chen et al. (2018), an attention layer computes a weighted sum of input values, where the weights are determined based on the relevance of the different input components. In BAPA, the goal of the attention layer is to extract the components of the input important for binding affinity prediction in a context vector. The encoded and context vectors are then concatenated an used by an MLP to obtain the final prediction. Wang Y. et al. (2021) also used self-attention in their PointTransformer architecture. The use of the attention mechanism (Bahdanau et al., 2014; Luong et al., 2015) in binding affinity prediction is also found in proteochemometric models (Karimi et al., 2019; Zhao et al., 2019).

A totally different approach from the data-driven ones reviewed above is to use physics-based methods for the computation of binding free energies accelerated or improved using ML and DL. Thanks to the recent developments in ML force fields (Unke et al., 2021), accurate alchemical free energy calculations based on such force fields are starting to appear (Rufa et al., 2020; Wieder et al., 2021). ML-based corrections to conventional free energy calculations will also play an important role in reaching good prediction accuracy of protein-ligand binding free energies (Dong et al., 2021). While such methods are outside the scope of this review, we believe the exploration and development of ML and DL methods in the field of free energy calculations will provide very interesting outcomes in the coming years, by getting the methodology closer to chemical accuracy while significantly reducing computational costs.

5 TRAINING AND EVALUATION

5.1 Back-Propagation, Regularization and Transfer Learning

Deep learning architectures for supervised learning are usually trained with gradient-based optimisation of a loss (or cost, or error) function that represents some measure of the prediction error (such as the mean squared difference between predicted and expected values). The weights and biases (trainable or learnable parameters) of the model are initialized from a random distribution or in a data-driven fashion (Narkhede et al., 2021), and they are iteratively adjusted by gradient-based optimisation techniques (such as stochastic gradient descent (Bottou et al., 1998)) to minimize a loss function.

Rumelhart et al. (1986) developed an algorithm called backpropagation, which allows computing the gradient of the loss function with respect to the parameters of the model (weights and biases) in an automated and efficient way. The algorithm consists of a forward pass computing the output of each component of the neural network, and the final output is used to evaluate the loss function. Then, the error is propagated backward using the chain rule of calculus to compute the gradients of the loss function with respect to each parameter of the network. The backpropagation algorithm is explained in detail in Goodfellow et al. (2016).

Modern deep learning frameworks such as PyTorch (Paszke et al., 2019; Li S. et al., 2020) and TensorFlow (Yu et al., 2018) usually require one to define only the forward pass, and gradients of the loss function can be easily and automatically computed with respect to any parameter. The availability of open-source, well-designed, and easy-to-use deep learning frameworks certainly contributed to the increased application of DL in different areas of research, including drug discovery.

Given the large number of parameters, DL architectures are often subject to the pitfalls of overfitting. To prevent overfitting, several techniques are commonly employed such as early stopping (Caruana et al., 2001), and the use of dropout layers (Srivastava et al., 2014).

Oftentimes, especially in the field of drug discovery, there is interest in models that are not completely generalizable but work well in specific cases such as specific protein families. Once a model has been trained on a general data set, it is possible to fine-tune the learned parameters to improve performance for specific tasks. Transfer learning (Bozinovski and Fulgosi 1976) methods can be subdivided in four classes (Pan and Yang 2010): instance-based, feature-based, parameter-based, and relation-based. Deep transfer learning, the combination of transfer learning and deep learning architectures (Tan et al., 2018), is commonly exploited in

drug discovery applications where learned representations are employed in different tasks (feature-based transfer learning) or where pre-trained models are fine-tuned for specific tasks (parameter-based transfer learning). The latter technique has been used successfully to develop protein family-specific models for virtual screening (Imrie et al., 2018), for example. An overview of transfer learning in drug discovery is given by Cai et al. (2020).

Multitask learning, which is closely related to transfer learning, consists in learning multiple endpoints at the same time using a shared representation (Ramsundar et al., 2017). Multitask learning can be used for the development of ML and DL SFs for both pose prediction (docking) and binding affinity prediction (scoring) (Ashtawy and Mahapatra 2017; Francoeur et al., 2020).

5.2 Evaluation

The models for protein-ligand binding affinity prediction discussed above consist of regression models, which given a protein-ligand complex as input return a real-valued estimate of the binding affinity (usually pK_d, pK_i, or pIC_{50}).

In the CASF benchmark, arguably one of the most used benchmarks for the development of scoring functions, the scoring power of a scoring function is measured in terms of correlation between experimental and predicted values. This correlation is measured quantitatively using Pearson's correlation coefficient r, defined as

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}}, \tag{15}$$

where (x_i, y_i) are the predicted and experimental values of the binding affinity, while $\bar{x}$ and $\bar{y}$ are the corresponding averages on the whole data set. A Pearson's r of 1.0 indicates perfect correlation, while a Pearson's r of 0.0 indicates no correlation. The Pearson's correlation coefficient is often accompanied by the root mean squared error

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_i^N (x_i - y_i)^2}, \tag{16}$$

or the mean absolute error (MAE)

$$\text{MAE} = \frac{1}{N}\sum_i^N |x_i - y_i|, \tag{17}$$

where N is the total number of samples in the test set.

The predicted value of the protein-ligand binding affinity can also be used to rank compounds, usually against the same target. Common metrics to evaluate the ranking power of a scoring function are rank correlation coefficients such as Spearman's ρ (Spearman 2010) and Kendall's τ (Kendall 1938). The Spearman's rank correlation coefficient is defined as (Spearman 2010)

$$\rho = \frac{\sum_i \left(r_{x_i} - \bar{r_x}\right)\left(r_{y_i} - \bar{r_y}\right)}{\sqrt{\sum_i \left(r_{x_i} - \bar{r_x}\right)^2}\sqrt{\sum_i \left(r_{y_i} - \bar{r_y}\right)^2}}, \tag{18}$$

which is similar to Pearson's r but uses the predicted and experimental ranks (r_{x_i}, r_{y_i})—and the corresponding sample averages—instead of using directly the predicted and experimental values (x_i, y_i). The other difference is that the Pearson's correlation coefficient is usually computed on the whole data set, while the Spearman's rank correlation coefficient (and other rank correlation coefficients) are often disaggregated by target. This is the case for the CASF benchmarks, for example (Su et al., 2018). Another way to quantify the ranking power of a scoring function is the predictive index (PI) introduced by Pearlman and Charifson (2001) and defined as

$$\text{PI} = \frac{\sum_i \sum_{j>i} W_{ij} C_{ij}}{\sum_i \sum_{j>i} W_{ij}} \tag{19}$$

where $W_{ij} = |y_i - y_j|$ is the absolute difference between the experimental binding data of ligands i and j and where C_{ij} is defined as (Pearlman and Charifson 2001)

$$C_{ij} = \begin{cases} 1 & \text{if } \frac{y_j - y_i}{x_j - x_i} < 0, \\ -1 & \text{if } \frac{y_j - y_i}{x_j - x_i} > 0, \\ 0 & \text{if } x_j - x_i = 0. \end{cases} \tag{20}$$

The weights W_{ij} reflect the fact that ranking incorrectly compounds with similar experimental binding affinities is less detrimental than ranking incorrectly compounds with vastly different binding affinities. As for Spearman's and Kendall's rank correlation coefficients, the PI is bound on the interval $[-1, 1]$ (with 0 indicating random predictions).

Confidence intervals for the correlation coefficients described above can be computed using bootstrapping (Efron, 1992). For the CASF-2016 benchmark this is easily done with the provided analysis scripts (Su et al., 2018). The very important topics of calculation of confidence intervals and comparison of different models are discussed at length in Nicholls (2014) and Nicholls (2016) and while we are concerned with regression models in this review, we point the reader interested in the comparison of classification models to Patrick Walters (2021).

Given that the gradient-based training described above depends on the initialization of the parameters of the model, oftentimes multiple models are trained starting from different weights and using different seeds for the random number generator (used for random weight initialization, random shuffling of examples, . . .), and the final prediction consists on a combination of the results of the different models (often an average). This ensemble approach has been shown multiple times to improve predictions of machine learning and deep learning models (Hansen and Salamon 1990; Ashtawy and Mahapatra 2015; Ericksen et al., 2017; Francoeur et al., 2020; Kwon et al., 2020; Meli et al., 2021). More generally, a consensus score amongst multiple models (also with different architectures) can be used as well (Druchok et al., 2021), and the average between different models (different architectures and/or different training data sets) has been shown to improve pose predictions with CNN scoring functions (McNutt et al., 2021). While the

average across different models is often used to estimate the performance of the ensemble, the standard deviation across predictions gives information about their stability and can be used as a diagnostic tool. Low standard deviations are expected within the domain of applicability of the models, while large standard deviations are often a symptom of poor generalizability.

Consensus scoring is not a new idea applicable only to machine learning and deep learning models; several flavors of consensus scoring have been successfully applied in combining different classical docking scoring functions for a long time (Charifson et al., 1999; Wang and Wang 2001; Clark et al., 2002). It is now commonly applied in ML and DL scoring functions to improve prediction performance.

Uncertainty quantification is an important field of machine learning and deep learning research and applications in drug discovery are a very active area of research. Some uncertainty quantification methods such as Monte Carlo dropout (Gal and Ghahramani 2016) remain under explored. Recently, evidential deep learning (Amini et al., 2020) has been applied to uncertainty quantification in DL-based QSAR (Soleimany et al., 2021). Soleimany et al. (2021) show that evidential deep learning allows to obtain predictions where uncertainty correlates with error and that uncertainty can be employed to perform sample-efficient training. Given the flexibility and scalability of the approach, which can be easily incorporated into existing architectures, this approach might contribute to the development of SFs in the near future.

5.3 Cross-Validation and Data Splitting

Very important aspects to consider when training and evaluating a new model are the size of the training set, the overlap between training and test sets, and the data set bias. These aspects need to be carefully evaluated, to properly assess the performance and generalizability of a new model.

The size of the training set affects the performance of ML and DL models and several authors noticed that including more examples in the training set—even of a lower quality, such as lower-resolution structures—improves model performance (Li et al., 2015; Francoeur et al., 2020). Learning curves, which show the prediction error as a function of the number of training examples, are commonly employed to evaluate and compare ML and DL methods in molecular properties prediction but they remain somewhat uncommon in the evaluation of structure-based models for binding affinity prediction, probably because of the much smaller size of the data sets available for training and evaluation.

The similarity between training and test sets has also a very high impact on the performance of structure-based models (Kramer and Gedeck 2010; Li and Yang 2017) and a careful model evaluation needs to take this similarity into account to avoid artificially inflated performance. Li and Yang (2017) studied the impact on ML SFs of protein structural and sequence similarity between the training and test. In their study, they remove training proteins that are highly similar to the ones in the test set, as evaluated by structural and sequence alignment. They concluded that ML SFs do not outperform classical scoring functions after removal of proteins from the training set with a high degree of similarity with the test set and therefore they attributed the higher performance of ML SFs compared to classical SFs to the existence of similarities between proteins in the training and test sets. Li et al. (2018), however, performed a similar study and concluded that the good scoring power of RF-Score is not exclusively due to a high number of similar proteins, although when sufficiently similar targets are present in both the training and test set ML scoring functions perform consistently better than classical scoring functions (Shen et al., 2020b). Additionally, ML scoring functions are able to exploit new data points as they become available, while classical scoring functions seem unable to exploit the large volumes of structural and interaction data available nowadays; incorporating a larger proportion of similar complexes to the training set does not seem to make classical SFs more accurate, according to Li et al. (2019b).

Boyles et al. (2019) and Su et al. (2020) both developed subsets of the PDBbind data set to carefully evaluate the effect of protein and ligand similarities on the performance of models trained on PDBbind and tested on the CASF data set. Boyles et al. (2019) evaluated ligand similarity using the Tanimoto similarity between Morgan fingerprints of each pair of ligands while protein similarity was evaluated with by sequence identity. Su et al. (2020) also used protein sequences to determine protein similarity, but used 3D shape similarity (Vainio et al., 2009) to evaluate the similarity between ligands. Additionally, Su et al. (2020) also evaluated binding site similarity—the binding site might be preserved, in contrast to the overall protein sequence—using structural descriptors including residue types and interatomic distances (Yeturu and Chandra 2008). Both groups confirmed the strong dependence on the similarity between the training and test set of the performance of ML scoring functions, which poses a challenge in the comparison of ML and DL SFs with classical SFs. While these considerations are very important in the development of new methods and it is important to take them into account when comparing different models, in practical applications the similarity between the training set and the system under investigation can be exploited to obtain superior predictions compared to classical SFs. For example, Li et al. (2021a) argue that the performance of ML scoring functions is underestimated due to the artificial removal of similarities between the training and tests sets and put forward a new benchmark with tries to mimic prospective binding affinity predictions. However, it is important to keep in mind that ML and DL SFs might be less effective when dealing with novel targets or small molecules (Su et al., 2020), and the applicability domain needs to be clearly defined.

Very recently, Ji et al. (2022) developed a free and open-source Python package that allows to curate dataset for benchmarking out-of-distribution (OOD) algorithms in the context of protein-ligand binding affinity predictions. The authors highlight a significant performance gap between in-distribution and OOD experiments, highlighting the need for new and domain-specific techniques allowing better OOD generalization.

Another way to elucidate the performance of ML and DL SFs in light of similarities and dissimilarities between the training test set is to use clustered cross-validation. *K*-fold cross-validation is

an established technique for the evaluation of ML and DL models (Arlot and Celisse 2010) that consists of randomly splitting the training set into K different sets and use, in turn, $K - 1$ sets for training and the remaining set for validation/testing. Francoeur et al. (2020) evaluated the performance of their CNN scoring function using cross-validation with clusters based on protein sequence and ligand fingerprint similarities (for the models trained using PDBbind) and also concluded that evaluations based on the PDBbind core set are overly-optimistic and therefore a rather poor measure of the model's ability to generalize to novel target and small molecules.

Finally, care should be taken in the presence of data set bias. One of the simplest forms of bias in current data sets is that published binding affinities tend to come from publications where potent binders were identified. Therefore, the distribution of binding affinities available for training might be skewed to potent binders and the trained model might be unable to predict binding affinities for weak binders. Bias can also be introduced in the construction of the training and test sets. For example, for the classification of actives and decoys on the DUD-E data set (Mysinger et al., 2012) it has been shown that analogue bias together with easily distinguishable decoys (decoys bias) result in CNN SFs exploiting only ligand information even when structure-based information is provided (Chen L. et al., 2019). Yang et al. (2020) also caution about the use of DUD-E to train ML and DL models to predict protein-ligand interactions but point out that the data set can still serve as an independent test set. Sieg et al. (2019) analyzed the problem of data set bias in-depth and proposed guidelines to recognize biases and develop robust models. Yang et al. (2020) suggest to evaluate the performance of ligand-only and protein-only models to better understand what ML and DL methods are learning from protein-ligand complexes.

The problem of unnoticed biases in the dataset that are exploited on learning by complex DL models is related to the infamous "black box" nature of some models.

6 EXPLAINABLE AI

As mentioned in the previous section, the "black box" nature of some models poses serious challenges in the identification of biases in the data sets and often prevents a deeper understanding of the model predictions and especially of its failures. In recent years, a lot of research effort has been devoted to model interpretability and explainable artificial intelligence (XAI) (Lipton 2018; Gunning et al., 2019; Murdoch et al., 2019).

To unpack the predictions of CNN-based scoring functions, several authors focused on feature attribution methods. For example, Stepniewska-Dziubinska et al. (2018) estimated feature importance of the different input channels by looking at the weight distributions of the convolutional filters of the first layer. Hochuli et al. (2018) also looked at the weights of the convolutional filters of the first layer, which can give some insight on how the model uses the different input atom types. Hochuli et al. (2018) used additional established methods for feature attribution—such as gradient computation, a modified version of layer-wise relevance propagation (Bach et al., 2015), and masking (Štrumbelj et al., 2009; Szegedy et al., 2013)—combined with visualization of the protein-ligand complex, showing that each method provides some insight into their CNN scoring function.

Gradient-based feature attribution methods, which allow to determine (local) feature importance, consist in computing the gradient of the prediction with respect to the input. In DL models, such gradients are readily available thanks to the automatic differentiation machinery of modern deep learning frameworks. Interestingly, the gradients of the SF with respect to atomic coordinates can be used to perform ligand pose optimization in the context of docking (Ragoza et al., 2017b). Masking, a perturbation-based feature attribution approach, consists in removing part of the input in order to measure the change in output. Masking can be performed on single atoms or fragments and whole protein residues. While masking approaches are close to chemical intuition and directly estimate feature importance of different atoms or functional groups, they are computationally expensive since they require several evaluations per input.

Hochuli et al. (2018) show that feature attribution methods are able to identify important atoms in the ligand and this information can potentially be employed to optimize protein-ligand interactions during lead optimization. However, it is not always clear why particular atoms are highlighted as important (Hochuli et al., 2018). More recently, Varela-Rial et al. (2022) applied the integrated gradient feature attribution technique (Sundararajan et al., 2017) to their K_{deep} model, confirming that the model can generally learn meaningful interactions, but that in some cases important interactions where ignored or protein residues far from the ligand were highlighted. The fact that residues far from the ligand are highlighted as important suggest that in some cases the model is exploiting protein similarity instead of important physical interactions between the protein and the ligand.

The feature attribution methods shortly described above in the context of CNN SFs can be applied to other models as well. For example, gradient-based attribution has been applied in combination with GNNs to identify pharmacophoric features involved in ligand binding (McCloskey et al., 2019), while Cho et al. (2020) applied layer-wise relevance propagation to explain the predictions of their InteractionNet model.

For GNNs, there are several XAI methods specifically tailored for such architecture (Jiménez-Luna et al., 2020) and it is currently a vibrant area of research (Baldassarre and Azizpour 2019; Yuan et al., 2020; Agarwal et al., 2021). XAI methods for graphs can be classified in two categories (Jiménez-Luna et al., 2020): sub-graph identification, and attention-based (Veličković et al., 2017) approaches. Sub-graph identification is useful to identify a compact sub-graph structure as well as a small subset of node features that contribute strongly to the model prediction (Ying et al., 2019). While GNN-based XAI has seen several applications in the prediction of molecular properties and reactivity (Ryu et al., 2018; Coley et al., 2019; Preuer et al., 2019; Jiménez-Luna et al., 2021b), its consistent application to GNN-based structure-based scoring function is still under-explored.

TABLE 2 | Non-exhaustive list of deep learning architectures for protein-ligand binding affinity prediction and their performance on the CASF-2016 scoring benchmark (if available). MLPs are included regardless of the number of hidden layers. Some methods are described in multiple publications and the ones referenced in this table are the ones where the model has been evaluated on the PDBbind Core set 2016/CASF-2016 set (or the original publication, if this evaluation is not available). The best result (the highest Pearson's r) is reported. Different publications might use slightly different custom variations of the CASF-2016 benchmark and the overlap between training and test sets might be taken into account in different ways. We refer the reader to the original publications for details, but we also report the number, N, of systems in the test set to outline possible differences. RMSEs are expressed in pK units.

Model	References	Architecture	Pearson's *r*	RMSE	*N*
—	Artemenko (2008)	MPL	—	—	—
NNScore 2.0	Durrant and McCammon (2011b)	MPL	—	—	—
BgN- & BsN-Score	Ashtawy and Mahapatra (2015)	MPL	—	—	—
DLscore	Hassan et al. (2018)	MPL	—	—	—
PLEC-NN	Wójcikowski et al. (2018)	MLP	0.82	—	290
Pair	Zhu et al. (2020)	MLP	0.75	1.44	285
AEScore	Meli et al. (2021)	MLP	0.83	1.22	285
TopologyNet	Cang and Wei (2017)	CNN	0.81	1.34	290
K_{deep}	Jiménez et al. (2018)	CNN	0.82	1.27	290
Pafnucy	Stepniewska-Dziubinska et al. (2018)	CNN	0.78	1.42	290
1D2D-CNN	Cang et al. (2018)	CNN	0.85	1.21	290
DeepAtom	Li et al. (2019c)	CNN	0.81	1.32	290
OnionNet	Zheng et al. (2019)	CNN	0.82	1.28	290
GNINA	Francoeur et al. (2020)	CNN	0.80	1.37	280
RosENet	Hassan-Harrirou et al. (2020)	CNN	0.82	1.24	
AK-Score	Kwon et al. (2020)	CNN	0.81	—	285
LigityScore1D	Azzopardi and Ebejer (2021)	CNN	0.74	1.46	285
OnionNet-2	Wang et al. (2021d)	CNN	0.86	1.16	285
SE-OnionNet	Wang et al. (2021b)	CNN	0.83	—	285
ACNN	Gomes et al. (2017)	GNN	—	—	—
PotentialNet	Feinberg et al. (2018)	GNN	—	—	—
graphDelta	Karlov et al. (2020)	GNN	0.87	1.05	285
SIGN	Li et al. (2021c)	GNN	0.80	1.32	290
InteractionGraphNet	Jiang et al. (2021)	GNN	0.84	1.22	262
GraphBAR	Son and Kim (2021)	GNN	0.78	1.41	290
PLIG/GATNet	Moesser et al. (2022)	GNN	0.84	1.22	272
PIGNet	Moon et al. (2022)	GNN	0.76	—	283
—	Berishvili et al. (2019)	CNN/ RNN	—	—	—
FAST	Jones et al. (2021)	CNN + GNN	0.81	1.31	290
BAPA	Seo et al. (2021)	CNN + ATT	0.82	1.30	285
PointTransformer	Wang et al. (2021c)	CNN + ATT	0.85	1.19	285

TABLE 3 | Performance of the models summarized in **Table 2** on the CSAR-NRC HiQ scoring benchmark. We only report evaluation results from the original reference. RMSEs are expressed in *pK* units.

Model	References	Set 1 *r*	Set 1 RMSE	Set 2 *r*	Set 2 RMSE
K_{deep}	Jiménez et al. (2018)	0.72	2.08	0.65	1.91
RosENet	Hassan-Harrirou et al. (2020)	0.83	1.78	0.80	1.44
OnionNet-2	Wang et al. (2021d)	0.89	1.50	0.87	1.21
graphDelta	Karlov et al. (2020)	0.74	1.59	0.71	1.52
GraphBAR	Son and Kim (2021)	0.75	1.59	0.65	1.56
PIGNet	Moon et al. (2022)	0.77	—	0.80	—
BAPA	Seo et al. (2021)	0.83	1.06	0.75	0.98

Uncertainty quantification, briefly discussed above in the context of model evaluation, is also an important XAI technique with the goal of quantifying the reliability of a prediction. Ensemble approaches are currently employed in most applications but probabilistic approaches such as evidential deep learning (Amini et al., 2020; Soleimany et al., 2021) will play a major role in the future.

Model interpretability is also important for classical ML methods—such as RFs, and SVMs—and QSAR models (Riniker and Landrum 2013b; Marchese Robinson et al., 2017; Rodríguez-Pérez and Bajorath 2019; Sheridan 2019), that are not the focus of this review. Several XAI methods are model-agnostic and therefore work with several ML and DL methods. However, it is worth mentioning that the heavily pre-processed

TABLE 4 | Performance of the models summarized in **Table 2** on the Astex Diverse Set scoring benchmark. We only report evaluation results from the original reference. RMSEs are expressed in *pK* units.

Model	References	Pearson's *r*	RMSE
Pafnucy	Stepniewska-Dziubinska et al (2018)	0.57	1.37
DeepAtom	Li et al. (2019c)	0.77	1.03
RosENet	Hassan-Harrirou et al. (2020)	0.48	1.65

features—such as interaction fingerprints discussed above—often used in combination with classical ML methods might render the models less interpretable than complex DL methods (Lipton 2018).

XAI approaches have the potential to transform the application of DL in real drug discovery applications. Being able to explain why a particular prediction is relevant and interesting would facilitate the adaptation of computational models in experimental pipelines. However, several limitations of XAI remain. For example, XAI approaches are still under active development and research, and often the methods need to be carefully tailored to the problem at hand. Additionally, as pointed out by Jiménez-Luna et al. (2020), there is no method that combines all desirable features of XAI—transparency, justification, informativeness, and uncertainty estimation—and therefore current applications often rely on consensus approaches between methods possessing different desirable features.

A recent, extensive, and very accessible review of XAI applications in drug discovery is given by Jiménez-Luna et al. (2020), which also outline recent advances in the field of XAI that are yet to be applied to chemistry or drug discovery. However, the field is moving at a fast pace and some of the methods without any reported application in drug discovery in Jiménez-Luna et al. (2020)—such as instance-based methods—are now starting to be applied successfully (Wellawatte et al., 2022).

7 DISCUSSION AND CONCLUSION

In this review, we focused on structure-based scoring functions for binding affinity prediction based on deep learning, many of which have been developed in recent years. The large number of recently developed SFs (see **Table 2** for a non-exhaustive list) is a testament to this rapid and fast-moving field. Li H. et al. (2020) recently reviewed ML and DL scoring functions for structure-based lead optimization developed between 2015 and 2019, but several new DL SFs have been developed and published in the last 2 years. Another example is the review of Shen et al. (2019), where only one GNN-based scoring function—PotentialNet (Feinberg et al., 2018)—was identified; most GNN scoring functions in **Table 2** are from 2020 and later.

Table 2 reports the scoring performance of several deep learning SFs mostly based on MLPs, CNNs, and GNNs on the CASF-2016 benchmark (whenever available in the primary reference). **Tables 3**, **4** report the scoring performance (Pearson's correlation coefficient) for the CSAR-NRC HiQ sets and the Astex Diverse Set for the same methods outlined in **Table 2**. The significantly lower number of methods tested on the CSAR-NRC HiQ sets and the Astex Diverse Set shows that the CASF benchmark is the *de facto* standard for the assessment of novel ML and DL scoring functions. Going forward, it would be interesting to see the other benchmarks gaining more traction in order to obtain more information about scoring function performance.

Despite the standardized benchmarks, some methods required the removal of some systems—leading to parametrization problems or outside the applicability domain—, but it is clear that most methods achieve similar performance on this benchmark. Additionally, the comparison between different methods on the same benchmark remains challenging due to possible differences in the training set—and the possible overlap between training and test sets. Finally, most methods are only tested on the CASF benchmark, despite other benchmark sets being widely available. These observations call for an in-depth comparison of the different methods trained and tested on exactly the same data sets, and using all available high-quality test sets.

The performance on CASF-2016 of the DL methods reviewed here is much higher than the performance of classical SFs on the same benchmark (Su et al., 2018). However, deep learning scoring functions do not always perform better or significantly better than scoring functions based on classical ML algorithms (Li H. et al., 2020). For example, it was shown that deep NNs and shallow regularized NNs perform similarly in QSAR applications when using the same set of descriptors (Winkler and Le 2016), and RF-based methods can achieve state-of-the-art performance when combined with suitable descriptors (Boyles et al., 2019). This is in stark contrast with other fields such as computer vision and natural language processing, where DL has quickly taken over classical ML algorithms. Additionally, while most ML and DL SFs for binding affinity prediction are trained and tested on crystal structures, their performance deteriorates when trained and tested on docked poses (Boyles et al., 2021), but it is worth noting that augmenting structure-based features obtained from docked structures with ligand-based features can recover the performance of structure-based models trained on crystal structures.

Another problem identified with ML and DL stricture-based SFs for binding affinity prediction is that while they perform significantly better than classical SFs for scoring (better correlation of the score with experimental binding affinities), they often perform poorly in virtual screening tasks (Gabel et al., 2014). Gabel et al. (2014) suggest that the development of novel ML and DL scoring function for binding affinity predictions should be accompanied by analysis of ligand pose sensitivity and enrichment capabilities in structure-based virtual screening. A more recent study by Shen et al. (2020a) confirms that ML scoring functions trained on PDBbind do not work well for virtual screening, especially on novel targets or targets with unconventional binding pockets. Multitask learning for binding affinity prediction and pose prediction trained using docked poses instead of crystallographic structures is effective to increase pose sensitivity in the context of CNN scoring functions (Francoeur et al., 2020). In the context of virtual screening, data

augmentation techniques can also increase pose sensitivity by forcing the model to rely less on ligand information (Scantlebury et al., 2020).

It is well known that the maximum achievable performance of ML and DL models for binding affinity predictions is limited by experimental errors and uncertainties (Kramer et al., 2012). This explains the similar performance of the best performing models on CASF-2016, which are likely close to the theoretical limit. Ventures like the Critical Assessment of Computational Hit-finding Experiments (CACHE) (Müller et al., 2022) will play an important role to validate computational methods in the future and generate a larger corpus of very high-quality data.

Going forward, it is important to evaluate ML and DL scoring functions as part of the docking pipeline. Most SFs discussed here are applied as a post-processing step of docking—or they are only applied to crystal structures—and only a few SFs seem to have been incorporated into readily available docking software. One such example is GNINA, where the CNN scoring function can be employed within the docking pipeline to re-score or locally optimize the ligand poses after fast Monte Carlo search (McNutt et al., 2021).

In this review, we have focussed mainly on methods for the prediction of protein-ligand binding affinity, and scoring functions evaluated on scoring tasks. However, ranking different compounds against the same target of interest is extremely useful in drug discovery applications. This is the case for lead optimization, where a lead compound against the target of interest has been identified and the goal is to increase potency while improving pharmacokinetic and pharmacodynamic properties. With binding affinity predictions computing such rankings is trivial. However, it remains unclear if the performance of ML and DL methods developed for scoring work equally well for ranking, especially in real drug discovery applications. For example some methods trained to predict binding affinities performed poorly on the different task of predicting the differences in binding affinity upon protein mutation (Aldeghi et al., 2018b). DL methods specifically designed for ranking—computing relative binding affinities—have been developed (Jiménez-Luna et al., 2019) and are an active area of research (McNutt and Koes 2022).

Given that the performance of a DL SFs varies widely from the target under consideration (Jiménez et al., 2018; Hassan-Harrirou et al., 2020; Meli et al., 2021), there is a lot of room for improvement in the development of target-specific scoring functions (Ross et al., 2013; Nogueira and Koch 2019). ML and DL algorithms are very good at exploiting similarities between inputs to perform predictions—as demonstrated by the performance drop when similarities between the training and test sets are removed (Boyles et al., 2019; Su et al., 2020)—and therefore family-specific scoring function will play an increasing role in early stages of drug discovery, when a particular target has been identified. However, it is still unclear if family-specific structure-based SFs consistently outperform ligand-based methods (Shen et al., 2020a).

Finally, given the ultimate goal of lowering the high attrition rate at later stages of drug discovery, the use of ADME/Tox predictions will also play an increasingly important role (Bhhatarai et al., 2019) alongside SFs to identify potent compounds against the target of interest and prioritize compounds for further experimental validation.

While the application of deep learning has not yet provided a step-changing improvement in the performance of binding affinity prediction compared to classical ML methods, further research into novel architectures, combined with the ever-increasing size and quality of data sets of protein-ligand complexes might change the tide in the future. Physics-based ML and DL will probably take over purely data-driven models in the long term, combining the best of both worlds. It is however important to remain realistic on the capabilities of DL SFs and it will be interesting to see how they actually perform in real-world drug discovery applications. Schneider et al. (2019) suggest a "curious but cautious approach" to the application of DL in the drug discovery process. XAI methods will certainly play a central role in the application of DL scoring functions to real drug discovery programs because knowing the reason behind a given prediction and understanding well the failure modes of the developed models will help to guide the next steps in the drug discovery process.

AUTHOR CONTRIBUTIONS

RM, GM, and PB conceptualised the review. RM wrote the initial draft. RM, GM, and PB edited, reviewed, and expanded the initial draft. All authors read and approved the final manuscript.

FUNDING

This work was supported by funding from the Biotechnology and Biological Sciences Research Council (BBSRC) [BB/MO11224/1] National Productivity Investment Fund (NPIF) [BB/S50760X/1] and Evotec (UK) via the Interdisciplinary Biosciences DTP at the University of Oxford.

REFERENCES

Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B., et al. (2015). GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX* 1-2, 19–25. doi:10.1016/j.softx.2015.06.001

Adcock, S. A., and McCammon, J. A. (2006). Molecular Dynamics: Survey of Methods for Simulating the Activity of Proteins. *Chem. Rev.* 106, 1589–1615. doi:10.1002/chin.20063029710.1021/cr040426m

Agarwal, C., Zitnik, M., and Lakkaraju, H. (2021). *Towards a Rigorous Theoretical Analysis and Evaluation of GNN Explanations. arXiv preprint arXiv:2106.09078.*

Aggarwal, R., Gupta, A., Chelur, V., Jawahar, C. V., and Priyakumar, U. D. (2021). DeepPocket: Ligand Binding Site Detection and Segmentation Using 3D Convolutional Neural Networks. *J. Chem. Inf. Model.* doi:10.1021/acs.jcim.1c00799

Aggarwal, R., and Koes, D. R. (2020). Learning Rmsd to Improve Protein-Ligand Scoring and Pose Selection. *ChemRxiv*. doi:10.26434/chemrxiv.11910870.v2

Ahmed, A., Smith, R. D., Clark, J. J., Dunbar, J. B., and Carlson, H. A. (2014). Recent Improvements to Binding MOAD: a Resource for Protein-Ligand Binding Affinities and Structures. *Nucleic Acids Res.* 43, D465–D469. doi:10.1093/nar/gku1088

Ain, Q. U., Aleksandrova, A., Roessler, F. D., and Ballester, P. J. (2015). Machine-learning Scoring Functions to Improve Structure-Based Binding Affinity Prediction and Virtual Screening. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 5, 405–424. doi:10.1002/wcms.1225

Aldeghi, M., Gapsys, V., and de Groot, B. L. (2018b). Accurate Estimation of Ligand Binding Affinity Changes upon Protein Mutation. *ACS Cent. Sci.* 4, 1708–1718. doi:10.1021/acscentsci.8b00717

Aldeghi, M., Heifetz, A., Bodkin, M. J., Knapp, S., and Biggin, P. C. (2016). Accurate Calculation of the Absolute Free Energy of Binding for Drug Molecules. *Chem. Sci.* 7, 207–218. doi:10.1039/c5sc02678d

Aldeghi, M., Heifetz, A., Bodkin, M. J., Knapp, S., and Biggin, P. C. (2017). Predictions of Ligand Selectivity from Absolute Binding Free Energy Calculations. *J. Am. Chem. Soc.* 139, 946–957. doi:10.1021/jacs.6b11467

Aldeghi, M., Bluck, J. P., and Biggin, P. C. (2018a). "Absolute Alchemical Free Energy Calculations for Ligand Binding: A Beginner's Guide," in *Methods in Molecular Biology* (New York: Springer), 199–232. doi:10.1007/978-1-4939-7756-7_11

Alford, R. F., Leaver-Fay, A., Jeliazkov, J. R., O'Meara, M. J., DiMaio, F. P., Park, H., et al. (2017). The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J. Chem. Theory Comput.* 13, 3031–3048. doi:10.1021/acs.jctc.7b00125

Alibay, I., Magarkar, A., Seeliger, D., and Biggin, P. (2022). Evaluating the Use of Absolute Binding Free Energy in the Fragment Optimization Process. *ChemRxiv*. doi:10.26434/chemrxiv-2022-cw2kq

Allen, W. J., Balius, T. E., Mukherjee, S., Brozell, S. R., Moustakas, D. T., Lang, P. T., et al. (2015). DOCK 6: Impact of New Features and Current Docking Performance. *J. Comput. Chem.* 36, 1132–1156. doi:10.1002/jcc.23905

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic Local Alignment Search Tool. *J. Mol. Biol.* 215, 403–410. doi:10.1016/s0022-2836(05)80360-2

Amini, A., Schwarting, W., Soleimany, A., and Rus, D. (2020). Deep Evidential Regression. *Adv. Neural Inf. Process. Syst.* 33, 14927–14937.

Åqvist, J., Medina, C., and Samuelsson, J. E. (1994). A New Method for Predicting Binding Affinity in Computer-Aided Drug Design. *Protein Eng.* 7, 385–391. doi:10.1093/protein/7.3.385

Arlot, S., and Celisse, A. (2010). A Survey of Cross-Validation Procedures for Model Selection. *Stat. Surv.* 4, 40–79. doi:10.1214/09-ss054

Artemenko, N. (2008). Distance Dependent Scoring Function for Describing Protein-Ligand Intermolecular Interactions. *J. Chem. Inf. Model.* 48, 569–574. doi:10.1021/ci700224e

Ashtawy, H. M., and Mahapatra, N. R. (2015). BgN-score and BsN-Score: Bagging and Boosting Based Ensemble Neural Networks Scoring Functions for Accurate Binding Affinity Prediction of Protein-Ligand Complexes. *BMC Bioinforma.* 16 Suppl 4, S8. doi:10.1186/1471-2105-16-s4-s8

Ashtawy, H. M., and Mahapatra, N. R. (2017). Task-specific Scoring Functions for Predicting Ligand Binding Poses and Affinity and for Screening Enrichment. *J. Chem. Inf. Model.* 58, 119–133. doi:10.1021/acs.jcim.7b00309

Atz, K., Grisoni, F., and Schneider, G. (2021). Geometric Deep Learning on Molecular Representations. *Nat. Mach. Intell.* 3, 1023–1032. doi:10.1038/s42256-021-00418-8

Azzopardi, J., and Ebejer, J. (2021). LigityScore: Convolutional Neural Network for Binding-Affinity Predictions. *Proc. 14th Int. Jt. Conf. Biomed. Eng. Syst. Technol.*, 38–49. doi:10.5220/0010228300380049

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K. R., and Samek, W. (2015). On Pixel-wise Explanations for Non-linear Classifier Decisions by Layer-wise Relevance Propagation. *PLoS One* 10, e0130140. doi:10.1371/journal.pone.0130140

Bahdanau, D., Cho, K., and Bengio, Y. (2014). *Neural Machine Translation by Jointly Learning to Align and Translate. arXiv preprint arXiv:1409.0473*.

Baldassarre, F., and Azizpour, H. (2019). *Explainability Techniques for Graph Convolutional Networks. arXiv preprint arXiv:1905.13686*.

Baldi, P. (2021). *Deep Learning in Science*. Cambridge University Press. doi:10.1017/9781108955652Deep Learning in Science

Ballester, P. J., and Mitchell, J. B. (2010). A Machine Learning Approach to Predicting Protein-Ligand Binding Affinity with Applications to Molecular Docking. *Bioinformatics* 26, 1169–1175. doi:10.1093/bioinformatics/btq112

Ballester, P. J., Schreyer, A., and Blundell, T. L. (2014). Does a More Precise Chemical Description of Protein-Ligand Complexes Lead to More Accurate Prediction of Binding Affinity? *J. Chem. Inf. Model.* 54, 944–955. doi:10.1021/ci500091r

Bao, J., He, X., and Zhang, J. Z. H. (2021). DeepBSP-a Machine Learning Method for Accurate Prediction of Protein-Ligand Docking Structures. *J. Chem. Inf. Model.* 61, 2231–2240. doi:10.1021/acs.jcim.1c00334

Bartók, A. P., De, S., Poelking, C., Bernstein, N., Kermode, J. R., Csányi, G., et al. (2017). Machine Learning Unifies the Modeling of Materials and Molecules. *Sci. Adv.* 3, e1701816. doi:10.1126/sciadv.1701816

Bartók, A. P., Kondor, R., and Csányi, G. (2013). On Representing Chemical Environments. *Phys. Rev. B* 87, 184115. doi:10.1103/physrevb.87.184115

Bash, P. A., Singh, U. C., Brown, F. K., Langridge, R., and Kollman, P. A. (1987). Calculation of the Relative Change in Binding Free Energy of a Protein-Inhibitor Complex. *Science* 235, 574–576. doi:10.1126/science.3810157

Baskin, I. I. (2020). The Power of Deep Learning to Ligand-Based Novel Drug Discovery. *Expert Opin. Drug Discov.* 15, 755–764. doi:10.1080/17460441.2020.1745183

Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., et al. (2018). *Relational Inductive Biases, Deep Learning, and Graph Networks. arXiv preprint arXiv:1806.01261*.

Behler, J., and Parrinello, M. (2007). Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.* 98, 146401. doi:10.1103/physrevlett.98.146401

Benson, M. L., Smith, R. D., Khazanov, N. A., Dimcheff, B., Beaver, J., Dresslar, P., et al. (2007). Binding MOAD, a High-Quality Protein-Ligand Database. *Nucleic Acids Res.* 36, D674–D678. doi:10.1093/nar/gkm911

Bento, A. P., Gaulton, A., Hersey, A., Bellis, L. J., Chambers, J., Davies, M., et al. (2013). The ChEMBL Bioactivity Database: An Update. *Nucleic Acids Res.* 42, D1083–D1090. doi:10.1093/nar/gkt1031

Berendsen, H. J. C., van der Spoel, D., and van Drunen, R. (1995). GROMACS: A Message-Passing Parallel Molecular Dynamics Implementation. *Comput. Phys. Commun.* 91, 43–56. doi:10.1016/0010-4655(95)00042-e

Berishvili, V. P., Perkin, V. O., Voronkov, A. E., Radchenko, E. V., Syed, R., Venkata Ramana Reddy, C., et al. (2019). Time-domain Analysis of Molecular Dynamics Trajectories Using Deep Neural Networks: Application to Activity Ranking of Tankyrase Inhibitors. *J. Chem. Inf. Model.* 59, 3519–3532. doi:10.1021/acs.jcim.9b00135

Berman, H. M., Bhat, T. N., Bourne, P. E., Feng, Z., Gilliland, G., Weissig, H., et al. (2000). The Protein Data Bank and the Challenge of Structural Genomics. *Nat. Struct. Biol.* 7 Suppl, 957–959. doi:10.1038/80734

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Brice, M. D., Rodgers, J. R., et al. (1977). The Protein Data Bank. A Computer-Based Archival File for Macromolecular Structures. *Eur. J. Biochem.* 80, 319–324. doi:10.1016/s0022-2836(77)80200-310.1111/j.1432-1033.1977.tb11885.x

Bhhatarai, B., Walters, W. P., Hop, C. E. C. A., Lanza, G., and Ekins, S. (2019). Opportunities and Challenges Using Artificial Intelligence in ADME/Tox. *Nat. Mat.* 18, 418–422. doi:10.1038/s41563-019-0332-5

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer New York. doi:10.1007/978-0-387-45528-0Pattern Recognition and Machine Learning

Böhm, H.-J., and Stahl, M. (2002). "The Use of Scoring Functions in Drug Discovery Applications," in *Reviews in Computational Chemistry* (John Wiley & Sons), 18, 41–87. chap. 2. doi:10.1002/0471433519.ch2

Böhm, H. J. (1992). Ludi: Rule-Based Automatic Design of New Substituents for Enzyme Inhibitor Leads. *J. Comput. Aided Mol. Des.* 6, 593–606. doi:10.1007/BF00126217

Böhm, H. J. (1994). The Development of a Simple Empirical Scoring Function to Estimate the Binding Constant for a Protein-Ligand Complex of Known Three-Dimensional Structure. *J. Comput. Aided Mol. Des.* 8, 243–256. doi:10.1007/bf00126743

Boresch, S., Tettinger, F., Leitgeb, M., and Karplus, M. (2003). Absolute Binding Free Energies: A Quantitative Approach for Their Calculation. *J. Phys. Chem. B* 107, 9535–9551. doi:10.1021/jp0217839

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). "A Training Algorithm for Optimal Margin Classifiers," in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory - COLT '92* (New York, NY, USA: Association for Computing Machinery), 144–152. doi:10.1145/130385.130401

Bottou, L. (1998). Online Learning and Stochastic Approximations. *On-line Learn. neural Netw.* 17 (9), 142.

Boyles, F., Deane, C. M., and Morris, G. M. (2019). Learning from the Ligand: Using Ligand-Based Features to Improve Binding Affinity Prediction. *Bioinformatics* 36, 758–764. doi:10.1093/bioinformatics/btz665

Boyles, F., Deane, C. M., and Morris, G. M. (2021). Learning from Docked Ligands: Ligand-Based Features Rescue Structure-Based Scoring Functions when Trained on Docked Poses. *J. Chem. Inf. Model.* doi:10.1021/acs.jcim.1c00096

Bozinovski, S., and Fulgosi, A. (1976). The Influence of Pattern Similarity and Transfer Learning upon Training of a Base Perceptron B2. *Proc. Symposium Inf.*, 3–121.

Breiman, L. (1996). Bagging Predictors. *Mach. Learn.* 24, 123–140. doi:10.1007/bf00058655

Breiman, L. (2001). Random Forests. *Mach. Learn.* 45, 5–32. doi:10.1023/a:1010933404324

Bronstein, M. M., Bruna, J., Cohen, T., and Veličković, P. (2021). *Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges. arXiv preprint arXiv:2104.13478*.

Brown, D. G., and Wobst, H. J. (2021). A Decade of FDA-Approved Drugs (2010-2019): Trends and Future Directions. *J. Med. Chem.* 64, 2312–2338. doi:10.1021/acs.jmedchem.0c01516

N. Brown (Editor) (2020). *Artificial Intelligence in Drug Discovery. Drug Discovery* (London, UK: Royal Society of Chemistry). doi:10.1039/9781788016841

Cai, C., Wang, S., Xu, Y., Zhang, W., Tang, K., Ouyang, Q., et al. (2020). Transfer Learning for Drug Discovery. *J. Med. Chem.* 63, 8683–8694. doi:10.1021/acs.jmedchem.9b02147

Cang, Z., Mu, L., and Wei, G. W. (2018). Representability of Algebraic Topology for Biomolecules in Machine Learning Based Scoring and Virtual Screening. *PLOS Comput. Biol.* 14, e1005929. doi:10.1371/journal.pcbi.1005929

Cang, Z., and Wei, G. W. (2017). TopologyNet: Topology Based Deep Convolutional and Multi-Task Neural Networks for Biomolecular Property Predictions. *PLOS Comput. Biol.* 13, e1005690. doi:10.1371/journal.pcbi.1005690

Carlson, H. A., Smith, R. D., Damm-Ganamet, K. L., Stuckey, J. A., Ahmed, A., Convery, M. A., et al. (2016). CSAR 2014: A Benchmark Exercise Using Unpublished Data from Pharma. *J. Chem. Inf. Model.* 56, 1063–1077. doi:10.1021/acs.jcim.5b00523

Caruana, R., Lawrence, S., and Giles, L. (2001). Overfitting in Neural Nets: Backpropagation, Conjugate Gradient, and Early Stopping. *Adv. Neural Inf. Process Syst.*, 402–408.

Chang, C. E., Chen, W., and Gilson, M. K. (2007). Ligand Configurational Entropy and Protein Binding. *Proc. Natl. Acad. Sci. U. S. A.* 104, 1534–1539. doi:10.1073/pnas.0610494104

Charifson, P. S., Corkery, J. J., Murcko, M. A., and Walters, W. P. (1999). Consensus Scoring: A Method for Obtaining Improved Hit Rates from Docking Databases of Three-Dimensional Structures into Proteins. *J. Med. Chem.* 42, 5100–5109. doi:10.1021/jm990352k

Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., and Blaschke, T. (2018a). The Rise of Deep Learning in Drug Discovery. *Drug Discov. Today* 23, 1241–1250. doi:10.1016/j.drudis.2018.01.039

Chen, L., Cruz, A., Ramsey, S., Dickson, C. J., Duca, J. S., Hornak, V., et al. (2019a). Hidden Bias in the DUD-E Dataset Leads to Misleading Performance of Deep Learning in Structure-Based Virtual Screening. *PLoS One* 14, e0220113. doi:10.1371/journal.pone.0220113

Chen, M. X., Firat, O., Bapna, A., Johnson, M., Macherey, W., Foster, G., et al. (2018b). "The Best of Both Worlds: Combining Recent Advances in Neural Machine Translation," in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 76–86, Melbourne, Australia. Association for Computational Linguistics

Chen, P., Ke, Y., Lu, Y., Du, Y., Li, J., Yan, H., et al. (2019b). DLIGAND2: an Improved Knowledge-Based Energy Function for Protein-Ligand Interactions Using the Distance-Scaled, Finite, Ideal-Gas Reference State. *J. Cheminform* 11, 52–11. doi:10.1186/s13321-019-0373-4

Chen, T., and Guestrin, C. (2016). "XGBoost," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (ACM), 785–794. doi:10.1145/2939672.2939785

Chen, X., Liu, M., and Gilson, M. K. (2001). BindingDB: A Web-Accessible Molecular Recognition Database. *Comb. Chem. High. Throughput Screen* 4, 719–725. doi:10.2174/1386207013330670

Cheng, T., Li, X., Li, Y., Liu, Z., and Wang, R. (2009). Comparative Assessment of Scoring Functions on a Diverse Test Set. *J. Chem. Inf. Model.* 49, 1079–1093. doi:10.1021/ci9000053

Cho, H., Lee, E. K., and Choi, I. S. (2020). Layer-wise Relevance Propagation of InteractionNet Explains Protein-Ligand Interactions at the Atom Level. *Sci. Rep.* 10, 21155–21211. doi:10.1038/s41598-020-78169-6

Cho, K., Van Merriënboer, B., Bahdanau, D., and Bengio, Y. (2014). *On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. arXiv preprint arXiv:1409.1259*.

Christensen, A. S., Sirumalla, S. K., Qiao, Z., O'Connor, M. B., Smith, D. G. A., Ding, F., et al. (2021). OrbNet Denali: A Machine Learning Potential for Biological and Organic Chemistry with Semi-empirical Cost and DFT Accuracy. *J. Chem. Phys.* 155, 204103. doi:10.1063/5.0061990

Chuang, K. V., Gunsalus, L. M., and Keiser, M. J. (2020). Learning Molecular Representations for Medicinal Chemistry. *J. Med. Chem.* 63, 8705–8722. doi:10.1021/acs.jmedchem.0c00385

Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014). *Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. arXiv preprint arXiv:1412.3555*.

Chupakhin, V., Marcou, G., Gaspar, H., and Varnek, A. (2014). Simple Ligand-Receptor Interaction Descriptor (SILIRID) for Alignment-free Binding Site Comparison. *Comput. Struct. Biotechnol. J.* 10, 33–37. doi:10.1016/j.csbj.2014.05.004

Clark, J. J., Benson, M. L., Smith, R. D., and Carlson, H. A. (2019). Inherent versus Induced Protein Flexibility: Comparisons within and between Apo and Holo Structures. *PLOS Comput. Biol.* 15, e1006705. doi:10.1371/journal.pcbi.1006705

Clark, J. J., Orban, Z. J., and Carlson, H. A. (2020). Predicting Binding Sites from Unbound versus Bound Protein Structures. *Sci. Rep.* 10, 15856–15918. doi:10.1038/s41598-020-72906-7

Clark, R. D., Strizhev, A., Leonard, J. M., Blake, J. F., and Matthew, J. B. (2002). Consensus Scoring for Ligand/protein Interactions. *J. Mol. Graph Model.* 20, 281–295. doi:10.1016/s1093-3263(01)00125-5

Coley, C. W., Jin, W., Rogers, L., Jamison, T. F., Jaakkola, T. S., Green, W. H., et al. (2019). A Graph-Convolutional Neural Network Model for the Prediction of Chemical Reactivity. *Chem. Sci.* 10, 370–377. doi:10.1039/c8sc04228d

Cortes, C., and Vapnik, V. (1995). Support-vector Networks. *Mach. Learn.* 20, 273–297. doi:10.1007/bf00994018

Cournia, Z., Allen, B., and Sherman, W. (2017). Relative Binding Free Energy Calculations in Drug Discovery: Recent Advances and Practical Considerations. *J. Chem. Inf. Model.* 57, 2911–2937. doi:10.1021/acs.jcim.7b00564

Da, C., and Kireev, D. (2014). Structural Protein-Ligand Interaction Fingerprints (SPLIF) for Structure-Based Virtual Screening: Method and Benchmark Study. *J. Chem. Inf. Model.* 54, 2555–2561. doi:10.1021/ci500319f

Dahl, G. E., Jaitly, N., and Salakhutdinov, R. (2014). Multi-task Neural Networks for QSAR Predictions. *arXiv preprint arXiv:1406.1231*.

Damm-Ganamet, K. L., Smith, R. D., Dunbar, J. B., Stuckey, J. A., and Carlson, H. A. (2013). CSAR Benchmark Exercise 2011-2012: Evaluation of Results from Docking and Relative Ranking of Blinded Congeneric Series. *J. Chem. Inf. Model.* 53, 1853–1870. doi:10.1021/ci400025f

Darby, J. F., Hopkins, A. P., Shimizu, S., Roberts, S. M., Brannigan, J. A., Turkenburg, J. P., et al. (2019). Water Networks Can Determine the Affinity of Ligand Binding to Proteins. *J. Am. Chem. Soc.* 141, 15818–15826. doi:10.1021/jacs.9b06275

Das, S., Krein, M. P., and Breneman, C. M. (2010). Binding Affinity Prediction with Property-Encoded Shape Distribution Signatures. *J. Chem. Inf. Model.* 50, 298–308. doi:10.1021/ci9004139

David, L., Thakkar, A., Mercado, R., and Engkvist, O. (2020). Molecular Representations in AI-Driven Drug Discovery: A Review and Practical Guide. *J. Cheminform* 12, 1–22. doi:10.1186/s13321-020-00460-5

de Magalhães, C. S., Almeida, D. M., Barbosa, H. J. C., and Dardenne, L. E. (2014). A Dynamic Niching Genetic Algorithm Strategy for Docking Highly Flexible Ligands. *Inf. Sci.* 289, 206–224. doi:10.1016/j.ins.2014.08.002

De, S., Bartók, A. P., Csányi, G., and Ceriotti, M. (2016). Comparing Molecules and Solids across Structural and Alchemical Space. *Phys. Chem. Chem. Phys.* 18, 13754–13769. doi:10.1039/c6cp00415f

Debroise, T., Shakhnovich, E. I., and Chéron, N. (2017a). A Hybrid Knowledge-Based and Empirical Scoring Function for Protein-Ligand Interaction: SMoG2016. *J. Chem. Inf. Model.* 57, 584–593. doi:10.1021/acs.jcim.6b00610

Debroise, T., Shakhnovich, E. I., and Chéron, N. (2017b). A Hybrid Knowledge-Based and Empirical Scoring Function for Protein-Ligand Interaction: SMoG2016. *J. Chem. Inf. Model.* 57, 584–593. doi:10.1021/acs.jcim.6b00610

Defferrard, M., Bresson, X., and Vandergheynst, P. (2016). Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering. *Adv. Neural Inf. Process Syst.* 29, 3844–3852.

Deng, J., Dong, W., Socher, R., Li, L.-J., Kai Li, K., and Li Fei-Fei, L. (2009). "ImageNet: A Large-Scale Hierarchical Image Database," in 2009 IEEE Conference on Computer Vision and Pattern Recognition (IEEE), 248–255. doi:10.1109/cvpr.2009.5206848

Deng, W., Breneman, C., and Embrechts, M. J. (2004). Predicting Protein-Ligand Binding Affinities Using Novel Geometrical Descriptors and Machine-Learning Methods. *J. Chem. Inf. Comput. Sci.* 44, 699–703. doi:10.1021/ci034246+

Deng, Z., Chuaqui, C., and Singh, J. (2003). Structural Interaction Fingerprint (SIFt): a Novel Method for Analyzing Three-Dimensional Protein-Ligand Binding Interactions. *J. Med. Chem.* 47, 337–344. doi:10.1021/jm030331x

DesJarlais, R. L., Sheridan, R. P., Seibel, G. L., Dixon, J. S., Kuntz, I. D., and Venkataraghavan, R. (1988). Using Shape Complementarity as an Initial Screen in Designing Ligands for a Receptor Binding Site of Known Three-Dimensional Structure. *J. Med. Chem.* 31, 722–729. doi:10.1021/jm00399a006

DeWitte, R. S., Ishchenko, A. V., and Shakhnovich, E. I. (1997). SMoG: De Novo Design Method Based on Simple, Fast, and Accurate Free Energy Estimates. 2. Case Studies in Molecular Design. *J. Am. Chem. Soc.* 119, 4608–4617. doi:10.1021/ja963689+

DeWitte, R. S., and Shakhnovich, E. I. (1996). SMoG: De Novo Design Method Based on Simple, Fast, and Accurate Free Energy Estimates. 1. Methodology and Supporting Evidence. *J. Am. Chem. Soc.* 118, 11733–11744. doi:10.1021/ja960751u

Dickson, M., and Gagnon, J. P. (2004). Key Factors in the Rising Cost of New Drug Discovery and Development. *Nat. Rev. Drug Discov.* 3, 417–429. doi:10.1038/nrd1382

DiMasi, J. A., Grabowski, H. G., and Hansen, R. W. (2016). Innovation in the Pharmaceutical Industry: New Estimates of R&D Costs. *J. Health Econ.* 47, 20–33. doi:10.1016/j.jhealeco.2016.01.012

Dittrich, J., Schmidt, D., Pfleger, C., and Gohlke, H. (2018). Converging a Knowledge-Based Scoring Function: DrugScore2018. *J. Chem. Inf. Model.* 59, 509–521. doi:10.1021/acs.jcim.8b00582

Dong, L., Qu, X., Zhao, Y., and Wang, B. (2021). Prediction of Binding Free Energy of Protein-Ligand Complexes with a Hybrid Molecular Mechanics/Generalized Born Surface Area and Machine Learning Method. *ACS Omega* 6, 32938–32947. doi:10.1021/acsomega.1c04996

Drews, J. (2000). Drug Discovery: A Historical Perspective. *Science* 287, 1960–1964. doi:10.1126/science.287.5460.1960

Druchok, M., Yarish, D., Garkot, S., Nikolaienko, T., and Gurbych, O. (2021). Ensembling Machine Learning Models to Boost Molecular Affinity Prediction. *Comput. Biol. Chem.* 93, 107529. doi:10.1016/j.compbiolchem.2021.107529

Dumoulin, V., and Visin, F. (2016). *A Guide to Convolution Arithmetic for Deep Learning. arXiv preprint arXiv:1603.07285*.

Dunbar, J. B., Smith, R. D., Damm-Ganamet, K. L., Ahmed, A., Esposito, E. X., Delproposto, J., et al. (2013). CSAR Data Set Release 2012: Ligands, Affinities, Complexes, and Docking Decoys. *J. Chem. Inf. Model.* 53, 1842–1852. doi:10.1021/ci4000486

Dunbar, J. B., Smith, R. D., Yang, C. Y., Ung, P. M., Lexa, K. W., Khazanov, N. A., et al. (2011). CSAR Benchmark Exercise of 2010: Selection of the Protein-Ligand Complexes. *J. Chem. Inf. Model.* 51, 2036–2046. doi:10.1021/ci200082t

Durrant, J. D., and McCammon, J. A. (2011a). BINANA: A Novel Algorithm for Ligand-Binding Characterization. *J. Mol. Graph Model.* 29, 888–893. doi:10.1016/j.jmgm.2011.01.004

Durrant, J. D., and McCammon, J. A. (2011b). NNScore 2.0: a Neural-Network Receptor-Ligand Scoring Function. *J. Chem. Inf. Model.* 51, 2897–2903. doi:10.1021/ci2003889

Durrant, J. D., and McCammon, J. A. (2010). NNScore: a Neural-Network-Based Scoring Function for the Characterization of Protein-Ligand Complexes. *J. Chem. Inf. Model.* 50, 1865–1871. doi:10.1021/ci100244v

Efron, B. (1992). "Bootstrap Methods: Another Look at the Jackknife," in *Springer Series in Statistics* (New York: Springer), 569–593. doi:10.1007/978-1-4612-4380-9_41

Eldridge, M. D., Murray, C. W., Auton, T. R., Paolini, G. V., and Mee, R. P. (1997). 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.* 11, 425–445. doi:10.1023/a:1007996124545

Ericksen, S. S., Wu, H., Zhang, H., Michael, L. A., Newton, M. A., Hoffmann, F. M., et al. (2017). Machine Learning Consensus Scoring Improves Performance across Targets in Structure-Based Virtual Screening. *J. Chem. Inf. Model.* 57, 1579–1590. doi:10.1021/acs.jcim.7b00153

Ewing, T. J., Makino, S., Skillman, A. G., and Kuntz, I. D. (2001). DOCK 4.0: Search Strategies for Automated Molecular Docking of Flexible Molecule Databases. *J. Comput. Aided Mol. Des.* 15, 411–428. doi:10.1023/a:1011115820450

Feinberg, E. N., Sur, D., Wu, Z., Husic, B. E., Mai, H., Li, Y., et al. (2018). PotentialNet for Molecular Property Prediction. *ACS Cent. Sci.* 4, 1520–1530. doi:10.1021/acscentsci.8b00507

Feng, Q., Dueva, E., Cherkasov, A., and Ester, M. (2018). *Padme: A Deep Learning-Based Framework for Drug-Target Interaction Prediction. arXiv preprint arXiv:1807.09741*.

Francoeur, P. G., Masuda, T., Sunseri, J., Jia, A., Iovanisci, R. B., Snyder, I., et al. (2020). Three-dimensional Convolutional Neural Networks and a Cross-Docked Data Set for Structure-Based Drug Design. *J. Chem. Inf. Model.* 60, 4200–4215. doi:10.1021/acs.jcim.0c00411

Freund, Y., and Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *J. Comput. Syst. Sci.* 55, 119–139. doi:10.1006/jcss.1997.1504

Friedman, J. H. (2002). Stochastic Gradient Boosting. *Comput. Statistics Data Analysis* 38, 367–378. doi:10.1016/s0167-9473(01)00065-2

Friesner, R. A., Banks, J. L., Murphy, R. B., Halgren, T. A., Klicic, J. J., Mainz, D. T., et al. (2004). Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* 47, 1739–1749. doi:10.1021/jm0306430

Friesner, R. A., Murphy, R. B., Repasky, M. P., Frye, L. L., Greenwood, J. R., Halgren, T. A., et al. (2006). Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein-Ligand Complexes. *J. Med. Chem.* 49, 6177–6196. doi:10.1021/jm051256o

Fukushima, K. (1980). Neocognitron: a Self Organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position. *Biol. Cybern.* 36, 193–202. doi:10.1007/bf00344251

Gabel, J., Desaphy, J., and Rognan, D. (2014). Beware of Machine Learning-Based Scoring Functions-On the Danger of Developing Black Boxes. *J. Chem. Inf. Model.* 54, 2807–2815. doi:10.1021/ci500406k

Gaieb, Z., Liu, S., Gathiaka, S., Chiu, M., Yang, H., Shao, C., et al. (2017). D3R Grand Challenge 2: Blind Prediction of Protein-Ligand Poses, Affinity Rankings, and Relative Binding Free Energies. *J. Comput. Aided Mol. Des.* 32, 1–20. doi:10.1007/s10822-017-0088-4

Gaieb, Z., Parks, C. D., Chiu, M., Yang, H., Shao, C., Walters, W. P., et al. (2019). D3R Grand Challenge 3: Blind Prediction of Protein-Ligand Poses and Affinity Rankings. *J. Comput. Aided Mol. Des.* 33, 1–18. doi:10.1007/s10822-018-0180-4

Gal, Y., and Ghahramani, Z. (2016). "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning," in Proceedings of The 33rd International Conference on Machine Learning, Proceedings of Machine Learning Research 48:1050-1059. Available from https://proceedings.mlr.press/v48/gal16.html..

Gao, X., Ramezanghorbani, F., Isayev, O., Smith, J. S., and Roitberg, A. E. (2020). TorchANI: A Free and Open Source PyTorch-Based Deep Learning Implementation of the ANI Neural Network Potentials. *J. Chem. Inf. Model.* 60, 3408–3415. doi:10.1021/acs.jcim.0c00451

Gathiaka, S., Liu, S., Chiu, M., Yang, H., Stuckey, J. A., Kang, Y. N., et al. (2016). D3R Grand Challenge 2015: Evaluation of Protein-Ligand Pose and Affinity Predictions. *J. Comput. Aided Mol. Des.* 30, 651–668. doi:10.1007/s10822-016-9946-8

Gaudelet, T., Day, B., Jamasb, A. R., Soman, J., Regep, C., Liu, G., et al. (2021). Utilizing Graph Machine Learning within Drug Discovery and Development. *Brief. Bioinform.* 22. doi:10.1093/bib/bbab159

Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., et al. (2011). ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* 40, D1100–D1107. doi:10.1093/nar/gkr777

Genheden, S., and Ryde, U. (2015). The MM/PBSA and MM/GBSA Methods to Estimate Ligand-Binding Affinities. *Expert Opin. Drug Discov.* 10, 449–461. doi:10.1517/17460441.2015.1032936

Géron, A. (2019). *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. Sebastapol, CA: O'Reilly Media.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. (2017). "Neural Message Passing for Quantum Chemistry," in International conference on machine learning (PMLR), 1263–1272.

Gilson, M. K., and Honig, B. H. (1986). The Dielectric Constant of a Folded Protein. *Biopolymers* 25, 2097–2119. doi:10.1002/bip.360251106

Gohlke, H., Hendlich, M., and Klebe, G. (2000). Knowledge-based Scoring Function to Predict Protein-Ligand Interactions. *J. Mol. Biol.* 295, 337–356. doi:10.1006/jmbi.1999.3371

Gomes, J., Ramsundar, B., Feinberg, E. N., and Pande, V. S. (2017). *Atomic Convolutional Networks for Predicting Protein-Ligand Binding Affinity*. *arXiv preprint arXiv:1703.10603*.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT press.

Goodsell, D. S., Autin, L., and Olson, A. J. (2019a). Illustrate: Software for Biomolecular Illustration. *Structure* 27, 1716–e1. doi:10.1016/j.str.2019.08.011

Goodsell, D. S., and Olson, A. J. (1990). Automated Docking of Substrates to Proteins by Simulated Annealing. *Proteins* 8, 195–202. doi:10.1002/prot.340080302

Goodsell, D. S., Zardecki, C., Di Costanzo, L., Duarte, J. M., Hudson, B. P., Persikova, I., et al. (2019b). RCSB Protein Data Bank: Enabling Biomedical Research and Drug Discovery. *Protein Sci.* 29, 52–65. doi:10.1002/pro.3730

Graves, A. (2012). *Supervised Sequence Labelling*. Springer Berlin Heidelberg, 5–13. chap. Supervised Sequence Labelling. doi:10.1007/978-3-642-24797-2_2Supervised Sequence Labelling

Guedes, I. A., Pereira, F. S. S., and Dardenne, L. E. (2018). Empirical Scoring Functions for Structure-Based Virtual Screening: Applications, Critical Aspects, and Challenges. *Front. Pharmacol.* 9, 1089. doi:10.3389/fphar.2018.01089

Guedes, I. A., Barreto, A., Marinho, D., Krempser, E., Kuenemann, M. A., Sperandio, O., et al. (2021a). New Machine Learning and Physics-Based Scoring Functions for Drug Discovery. *Sci. Rep.* 11, 1–19. doi:10.1038/s41598-021-82410-1

Guedes, I. A., Barreto, A. M. S., Marinho, D., Krempser, E., Kuenemann, M. A., Sperandio, O., et al. (2021b). New Machine Learning and Physics-Based Scoring Functions for Drug Discovery. *Sci. Rep.* 11, 1–19. doi:10.1038/s41598-021-82410-1

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., and Yang, G. Z. (2019). XAI-explainable Artificial Intelligence. *Sci. Robot.* 4. doi:10.1126/scirobotics.aay7120

Hahn, D. F., Bayly, C. I., Macdonald, H. E. B., Chodera, J. D., Mey, A. S., Mobley, D. L., et al. (2021). *Best Practices for Constructing, Preparing, and Evaluating Protein-Ligand Binding Affinity Benchmarks*. *arXiv preprint arXiv:2105.06222*.

Hansen, L. K., and Salamon, P. (1990). Neural Network Ensembles. *IEEE Trans. Pattern Anal. Mach. Intell.* 12, 993–1001. doi:10.1109/34.58871

Hartshorn, M. J., Verdonk, M. L., Chessari, G., Brewerton, S. C., Mooij, W. T., Mortenson, P. N., et al. (2007). Diverse, High-Quality Test Set for the Validation of Protein-Ligand Docking Performance. *J. Med. Chem.* 50, 726–741. doi:10.1021/jm061277y

Hassan, M., Mogollon, D. C., Fuentes, O., and sirimulla, s. (2018). DLSCORE: A Deep Learning Model for Predicting Protein-Ligand Binding Affinities. *ChemRxiv*. doi:10.26434/chemrxiv.6159143.v1

Hassan-Harrirou, H., Zhang, C., and Lemmin, T. (2020). RosENet: Improving Binding Affinity Prediction by Leveraging Molecular Mechanics Energies with an Ensemble of 3D Convolutional Neural Networks. *J. Chem. Inf. Model.* 60, 2791–2802. doi:10.1021/acs.jcim.0c00075

Hauser, K., Negron, C., Albanese, S. K., Ray, S., Steinbrecher, T., Abel, R., et al. (2018). Predicting Resistance of Clinical Abl Mutations to Targeted Kinase Inhibitors Using Alchemical Free-Energy Calculations. *Commun. Biol.* 1, 70–14. doi:10.1038/s42003-018-0075-x

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep Residual Learning for Image Recognition," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (IEEE), 770–778. doi:10.1109/cvpr.2016.90

Hingerty, B. E., Ritchie, R. H., Ferrell, T. L., and Turner, J. E. (1985). Dielectric Effects in Biopolymers: The Theory of Ionic Saturation Revisited. *Biopolymers* 24, 427–439. doi:10.1002/bip.360240302

Hochreiter, S., and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural comput.* 9, 1735–1780. doi:10.1162/neco.1997.9.8.1735

Hochuli, J., Helbling, A., Skaist, T., Ragoza, M., and Koes, D. R. (2018). Visualizing Convolutional Neural Network Protein-Ligand Scoring. *J. Mol. Graph Model.* 84, 96–108. doi:10.1016/j.jmgm.2018.06.005

Holderbach, S., Adam, L., Jayaram, B., Wade, R. C., and Mukherjee, G. (2020). RASPD+: Fast Protein-Ligand Binding Free Energy Prediction Using Simplified Physicochemical Features. *Front. Mol. Biosci.* 7, 601065. doi:10.3389/fmolb.2020.601065

Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer Feedforward Networks Are Universal Approximators. *Neural Netw.* 2, 359–366. doi:10.1016/0893-6080(89)90020-8

Hu, L., Benson, M. L., Smith, R. D., Lerner, M. G., and Carlson, H. A. (2005). Binding MOAD (Mother of All Databases). *Proteins* 60, 333–340. doi:10.1002/prot.20512

Huang, D. Z., Baber, J. C., and Bahmanyar, S. S. (2021a). The Challenges of Generalizability in Artificial Intelligence for ADME/Tox Endpoint and Activity Prediction. *Expert Opin. Drug Discov.* 16, 1045–1056. doi:10.1080/17460441.2021.1901685

Huang, K., Fu, T., Gao, W., Zhao, Y., Roohani, Y., Leskovec, J., et al. (2021b). *Therapeutics Data Commons: Machine Learning Datasets and Tasks for Therapeutics*. *arXiv preprint arXiv:2102.09548*.

Huang, S. Y., Grinter, S. Z., and Zou, X. (2010). Scoring Functions and Their Evaluation Methods for Protein-Ligand Docking: Recent Advances and Future Directions. *Phys. Chem. Chem. Phys.* 12, 12899–12908. doi:10.1039/c0cp00151a

Huang, S. Y., and Zou, X. (2006a). An Iterative Knowledge-Based Scoring Function to Predict Protein-Ligand Interactions: I. Derivation of Interaction Potentials. *J. Comput. Chem.* 27, 1866–1875. doi:10.1002/jcc.20504

Huang, S. Y., and Zou, X. (2006b). An Iterative Knowledge-Based Scoring Function to Predict Protein-Ligand Interactions: II. Validation of the Scoring Function. *J. Comput. Chem.* 27, 1876–1882. doi:10.1002/jcc.20505

Huang, S. Y., and Zou, X. (2010). Inclusion of Solvation and Entropy in the Knowledge-Based Scoring Function for Protein-Ligand Interactions. *J. Chem. Inf. Model.* 50, 262–273. doi:10.1021/ci9002987

Hubel, D. H. (1959). Single Unit Activity in Striate Cortex of Unrestrained Cats. *J. Physiol.* 147, 226–238. doi:10.1113/jphysiol.1959.sp006238

Hubel, D. H., and Wiesel, T. N. (1959). Receptive Fields of Single Neurones in the Cat's Striate Cortex. *J. Physiol.* 148, 574–591. doi:10.1113/jphysiol.1959.sp006308

Huey, R., Morris, G. M., Olson, A. J., and Goodsell, D. S. (2007). A Semiempirical Free Energy Force Field with Charge-Based Desolvation. *J. Comput. Chem.* 28, 1145–1152. doi:10.1002/jcc.20634

Imrie, F., Bradley, A. R., van der Schaar, M., and Deane, C. M. (2018). Protein Family-specific Models Using Deep Neural Networks and Transfer Learning Improve Virtual Screening and Highlight the Need for More Data. *J. Chem. Inf. Model.* 58, 2319–2330. doi:10.1021/acs.jcim.8b00350

Jasper, J. B., Humbeck, L., Brinkjost, T., and Koch, O. (2018). A Novel Interaction Fingerprint Derived from Per Atom Score Contributions: Exhaustive Evaluation of Interaction Fingerprint Performance in Docking Based Virtual Screening. *J. Cheminform* 10, 15–13. doi:10.1186/s13321-018-0264-0

Ji, Y., Zhang, L., Wu, J., Wu, B., Huang, L.-K., Xu, T., et al. (2022). *DrugOOD: Out-Of-Distribution (OOD) Dataset Curator and Benchmark for AI-Aided Drug Discovery–A Focus on Affinity Prediction Problems with Noise Annotations*. *arXiv preprint arXiv:2201.09637*.

Jiang, D., Hsieh, C. Y., Wu, Z., Kang, Y., Wang, J., Wang, E., et al. (2021). InteractionGraphNet: A Novel and Efficient Deep Graph Representation Learning Framework for Accurate Protein-Ligand Interaction Predictions. *J. Med. Chem.* 64, 18209–18232. doi:10.1021/acs.jmedchem.1c01830

Jiang, H., Fan, M., Wang, J., Sarma, A., Mohanty, S., Dokholyan, N. V., et al. (2020a). Guiding Conventional Protein-Ligand Docking Software with

Convolutional Neural Networks. *J. Chem. Inf. Model.* 60, 4594–4602. doi:10.1021/acs.jcim.0c00542

Jiang, M., Li, Z., Zhang, S., Wang, S., Wang, X., Yuan, Q., et al. (2020b). Drug-target Affinity Prediction Using Graph Neural Network and Contact Maps. *RSC Adv.* 10, 20701–20712. doi:10.1039/d0ra02297g

Jiménez, J., Doerr, S., Martínez-Rosell, G., Rose, A. S., and De Fabritiis, G. (2017). DeepSite: Protein-Binding Site Predictor Using 3D-Convolutional Neural Networks. *Bioinformatics* 33, 3036–3042. doi:10.1093/bioinformatics/btx350

Jiménez, J., Škalič, M., Martínez-Rosell, G., and De Fabritiis, G. (2018). KDEEP: Protein-Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks. *J. Chem. Inf. Model.* 58, 287–296. doi:10.1021/acs.jcim.7b00650

Jiménez-Luna, J., Pérez-Benito, L., Martínez-Rosell, G., Sciabola, S., Torella, R., Tresadern, G., et al. (2019). DeltaDelta Neural Networks for Lead Optimization of Small Molecule Potency. *Chem. Sci.* 10, 10911–10918. doi:10.1039/c9sc04606b

Jiménez-Luna, J., Skalic, M., Weskamp, N., and Schneider, G. (2021b). Coloring Molecules with Explainable Artificial Intelligence for Preclinical Relevance Assessment. *J. Chem. Inf. Model.* 61, 1083–1094. doi:10.1021/acs.jcim.0c01344

Jiménez-Luna, J., Grisoni, F., and Schneider, G. (2020). Drug Discovery with Explainable Artificial Intelligence. *Nat. Mach. Intell.* 2, 573–584. doi:10.1038/s42256-020-00236-4

Jiménez-Luna, J., Grisoni, F., Weskamp, N., and Schneider, G. (2021a). Artificial Intelligence in Drug Discovery: Recent Advances and Future Perspectives. *Expert Opin. Drug Discov.* 16, 949–959. doi:10.1080/17460441.2021.1909567

Jing, Y., Bian, Y., Hu, Z., Wang, L., and Xie, X. Q. (2018). Deep Learning for Drug Design: An Artificial Intelligence Paradigm for Drug Discovery in the Big Data Era. *AAPS J.* 20, 58–10. doi:10.1208/s12248-018-0210-0

Jones, D., Kim, H., Zhang, X., Zemla, A., Stevenson, G., Bennett, W. F. D., et al. (2021). Improved Protein-Ligand Binding Affinity Prediction with Structure-Based Deep Fusion Inference. *J. Chem. Inf. Model.* 61, 1583–1592. doi:10.1021/acs.jcim.0c01306

Jones, G., Willett, P., Glen, R. C., Leach, A. R., and Taylor, R. (1997). Development and Validation of a Genetic Algorithm for Flexible Docking. *J. Mol. Biol.* 267, 727–748. doi:10.1006/jmbi.1996.0897

Jones, G., Willett, P., and Glen, R. C. (1995). Molecular Recognition of Receptor Sites Using a Genetic Algorithm with a Description of Desolvation. *J. Mol. Biol.* 245, 43–53. doi:10.1016/s0022-2836(95)80037-9

Jones-Hertzog, D. K., and Jorgensen, W. L. (1997). Binding Affinities for Sulfonamide Inhibitors with Human Thrombin Using Monte Carlo Simulations with a Linear Response Method. *J. Med. Chem.* 40, 1539–1549. doi:10.1021/jm960684e

Jubb, H. C., Higueruelo, A. P., Ochoa-Montaño, B., Pitt, W. R., Ascher, D. B., and Blundell, T. L. (2017). Arpeggio: A Web Server for Calculating and Visualising Interatomic Interactions in Protein Structures. *J. Mol. Biol.* 429, 365–371. doi:10.1016/j.jmb.2016.12.004

Kadukova, M., and Grudinin, S. (2017). Convex-pl: a Novel Knowledge-Based Potential for Protein-Ligand Interactions Deduced from Structural Databases Using Convex Optimization. *J. Comput. Aided Mol. Des.* 31, 943–958. doi:10.1007/s10822-017-0068-8

Kadukova, M., Machado, K. D. S., Chacón, P., and Grudinin, S. (2021). KORP-PL: a Coarse-Grained Knowledge-Based Scoring Function for Protein-Ligand Interactions. *Bioinformatics* 37, 943–950. doi:10.1093/bioinformatics/btaa748

Karimi, M., Wu, D., Wang, Z., and Shen, Y. (2019). DeepAffinity: Interpretable Deep Learning of Compound-Protein Affinity through Unified Recurrent and Convolutional Neural Networks. *Bioinformatics* 35, 3329–3338. doi:10.1093/bioinformatics/btz111

Karlov, D. S., Sosnin, S., Fedorov, M. V., and Popov, P. (2020). graphDelta: MPNN Scoring Function for the Affinity Prediction of Protein-Ligand Complexes. *ACS Omega* 5, 5150–5159. doi:10.1021/acsomega.9b04162

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., et al. (2017). Lightgbm: A Highly Efficient Gradient Boosting Decision Tree. *Adv. Neural Inf. Process Syst.* 30.

Kearnes, S., McCloskey, K., Berndl, M., Pande, V., and Riley, P. (2016). Molecular Graph Convolutions: Moving beyond Fingerprints. *J. Comput. Aided Mol. Des.* 30, 595–608. doi:10.1007/s10822-016-9938-8

Kendall, M. G. (1938). A New Measure of Rank Correlation. *Biometrika* 30, 81–93. doi:10.1093/biomet/30.1-2.81

Kimber, T. B., Chen, Y., and Volkamer, A. (2021). Deep Learning in Virtual Screening: Recent Applications and Developments. *Int. J. Mol. Sci.* 22, 4435. doi:10.3390/ijms22094435

Kipf, T. N., and Welling, M. (2016). *Semi-supervised Classification with Graph Convolutional Networks. arXiv preprint arXiv:1609.02907*.

Koes, D. R., Baumgartner, M. P., and Camacho, C. J. (2013). Lessons Learned in Empirical Scoring with Smina from the CSAR 2011 Benchmarking Exercise. *J. Chem. Inf. Model.* 53, 1893–1904. doi:10.1021/ci300604z

Kramer, C., and Gedeck, P. (2010). Leave-cluster-out Cross-Validation Is Appropriate for Scoring Functions Derived from Diverse Protein Data Sets. *J. Chem. Inf. Model.* 50, 1961–1969. doi:10.1021/ci100264e

Kramer, C., Kalliokoski, T., Gedeck, P., and Vulpetti, A. (2012). The Experimental Uncertainty of Heterogeneous Public K(i) Data. *J. Med. Chem.* 55, 5165–5173. doi:10.1021/jm300131x

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* 60, 84–90. doi:10.1145/3065386

Kuzminykh, D., Polykovskiy, D., Kadurin, A., Zhebrak, A., Baskov, I., Nikolenko, S., et al. (2018). 3D Molecular Representations Based on the Wave Transform for Convolutional Neural Networks. *Mol. Pharm.* 15, 4378–4385. doi:10.1021/acs.molpharmaceut.7b01134

Kwon, Y., Shin, W. H., Ko, J., and Lee, J. (2020). AK-score: Accurate Protein-Ligand Binding Affinity Prediction Using an Ensemble of 3D-Convolutional Neural Networks. *Int. J. Mol. Sci.* 21, 8424. doi:10.3390/ijms21228424

Le Cun, Y., Jackel, L. D., Boser, B., Denker, J. S., Graf, H. P., Guyon, I., et al. (1989). Handwritten Digit Recognition: Applications of Neural Network Chips and Automatic Learning. *IEEE Commun. Mag.* 27, 41–46. doi:10.1109/35.41400

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based Learning Applied to Document Recognition. *Proc. IEEE* 86, 2278–2324. doi:10.1109/5.726791

Lenselink, E. B., ten Dijke, N., Bongers, B., Papadatos, G., van Vlijmen, H. W. T., Kowalczyk, W., et al. (2017). Beyond the Hype: Deep Neural Networks Outperform Established Methods Using a ChEMBL Bioactivity Benchmark Set. *J. Cheminform* 9, 45–14. doi:10.1186/s13321-017-0232-0

Li, H., Leung, K. S., Wong, M. H., and Ballester, P. J. (2015). Low-quality Structural and Interaction Data Improves Binding Affinity Prediction via Random Forest. *Molecules* 20, 10947–10962. doi:10.3390/molecules200610947

Li, H., Lu, G., Sze, K. H., Su, X., Chan, W. Y., and Leung, K. S. (2021a). Machine-learning Scoring Functions Trained on Complexes Dissimilar to the Test Set Already Outperform Classical Counterparts on a Blind Benchmark. *Brief. Bioinform.* 22, bbab225. doi:10.1093/bib/bbab225

Li, H., Peng, J., Leung, Y., Leung, K. S., Wong, M. H., Lu, G., et al. (2018). The Impact of Protein Structure and Sequence Similarity on the Accuracy of Machine-Learning Scoring Functions for Binding Affinity Prediction. *Biomolecules* 8, 12. doi:10.3390/biom8010012

Li, H., Peng, J., Sidorov, P., Leung, Y., Leung, K. S., Wong, M. H., et al. (2019a). Classical Scoring Functions for Docking Are Unable to Exploit Large Volumes of Structural and Interaction Data. *Bioinformatics* 35, 3989–3995. doi:10.1093/bioinformatics/btz183

Li, H., Peng, J., Sidorov, P., Leung, Y., Leung, K. S., Wong, M. H., et al. (2019b). Classical Scoring Functions for Docking Are Unable to Exploit Large Volumes of Structural and Interaction Data. *Bioinformatics* 35, 3989–3995. doi:10.1093/bioinformatics/btz183

Li, H., Sze, K.-H., Lu, G., and Ballester, P. J. (2020a). Machine-learning Scoring Functions for Structure-Based Drug Lead Optimization. *WIREs Comput. Mol. Sci.* 10, e1465. doi:10.1002/wcms.1465

Li, H., Sze, K.-H., Lu, G., and Ballester, P. J. (2021b). Machine-learning Scoring Functions for Structure-Based Virtual Screening. *WIREs Comput. Mol. Sci.* 11, e1478. doi:10.1002/wcms.1478

Li, L., Wang, B., and Meroueh, S. O. (2011). Support Vector Regression Scoring of Receptor-Ligand Complexes for Rank-Ordering and Virtual Screening of Chemical Libraries. *J. Chem. Inf. Model.* 51, 2132–2138. doi:10.1021/ci200078f

Li, S., Zhao, Y., Varma, R., Salpekar, O., Noordhuis, P., Li, T., et al. (2020b). PyTorch Distributed. *Proc. VLDB Endow.* 13, 3005–3018. doi:10.14778/3415478.3415530

Li, S., Zhou, J., Xu, T., Huang, L., Wang, F., Xiong, H., et al. (2021c). "Structure-aware Interactive Graph Neural Networks for the Prediction of Protein-Ligand Binding Affinity," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (ACM)*, 975–985. doi:10.1145/3447548.3467311

Li, Y., Han, L., Liu, Z., and Wang, R. (2014a). Comparative Assessment of Scoring Functions on an Updated Benchmark: 2. Evaluation Methods and General Results. *J. Chem. Inf. Model.* 54, 1717–1736. doi:10.1021/ci500081m

Li, Y., Liu, Z., Li, J., Han, L., Liu, J., Zhao, Z., et al. (2014b). Comparative Assessment of Scoring Functions on an Updated Benchmark: 1. Compilation of the Test Set. *J. Chem. Inf. Model.* 54, 1700–1716. doi:10.1021/ci500080q

Li, Y., and Yang, J. (2017). Structural and Sequence Similarity Makes a Significant Impact on Machine-Learning-Based Scoring Functions for Protein-Ligand Interactions. *J. Chem. Inf. Model.* 57, 1007–1012. doi:10.1021/acs.jcim.7b00049

Li, Y., Rezaei, M. A., Li, C., and Li, X. (2019c). "DeepAtom: A Framework for Protein-Ligand Binding Affinity Prediction," in *2019 IEEE International Conference on Bioinformatics and Biomedicine* (IEEE), 303–310. doi:10.1109/bibm47256.2019.8982964

Lim, J., Ryu, S., Park, K., Choe, Y. J., Ham, J., and Kim, W. Y. (2019). Predicting Drug-Target Interaction Using a Novel Graph Neural Network with 3D Structure-Embedded Graph Representation. *J. Chem. Inf. Model.* 59, 3981–3988. doi:10.1021/acs.jcim.9b00387

Limongelli, V., Marinelli, L., Cosconati, S., La Motta, C., Sartini, S., Mugnaini, L., et al. (2012). Sampling Protein Motion and Solvent Effect during Ligand Binding. *Proc. Natl. Acad. Sci. U. S. A.* 109, 1467–1472. doi:10.1073/pnas.1112181108

Lipton, Z. C. (2018). The Mythos of Model Interpretability. *Queue* 16, 31–57. doi:10.1145/3236386.3241340

Liu, J., and Wang, R. (2015). Classification of Current Scoring Functions. *J. Chem. Inf. Model.* 55, 475–482. doi:10.1021/ci500731a

Liu, T., Lin, Y., Wen, X., Jorissen, R. N., and Gilson, M. K. (2007). BindingDB: A Web-Accessible Database of Experimentally Determined Protein-Ligand Binding Affinities. *Nucleic Acids Res.* 35, D198–D201. doi:10.1093/nar/gkl999

Liu, Z., Li, Y., Han, L., Li, J., Liu, J., Zhao, Z., et al. (2014). PDB-wide Collection of Binding Data: Current Status of the PDBbind Database. *Bioinformatics* 31, 405–412. doi:10.1093/bioinformatics/btu626

Liu, Z., Su, M., Han, L., Liu, J., Yang, Q., Li, Y., et al. (2017). Forging the Basis for Developing Protein-Ligand Interaction Scoring Functions. *Acc. Chem. Res.* 50, 302–309. doi:10.1021/acs.accounts.6b00491

Lo, Y. C., Rensi, S. E., Torng, W., and Altman, R. B. (2018). Machine Learning in Chemoinformatics and Drug Discovery. *Drug Discov. Today* 23, 1538–1546. doi:10.1016/j.drudis.2018.05.010

Luong, M.-T., Pham, H., and Manning, C. D. (2015). *Effective Approaches to Attention-Based Neural Machine Translation. arXiv preprint arXiv:1508.04025*.

Ma, J., Sheridan, R. P., Liaw, A., Dahl, G. E., and Svetnik, V. (2015). Deep Neural Nets as a Method for Quantitative Structure-Activity Relationships. *J. Chem. Inf. Model.* 55, 263–274. doi:10.1021/ci500747n

Ma, N., Zhang, X., Zheng, H.-T., and Sun, J. (2018). "Shufflenet V2: Practical Guidelines for Efficient Cnn Architecture Design," in *Proceedings of the European Conference on Computer Vision*(Cham, Switzerland: ECCV), 116–131. doi:10.1007/978-3-030-01264-9_8

Macarron, R., Banks, M. N., Bojanic, D., Burns, D. J., Cirovic, D. A., Garyantes, T., et al. (2011). Impact of High-Throughput Screening in Biomedical Research. *Nat. Rev. Drug Discov.* 10, 188–195. doi:10.1038/nrd3368

Marchese Robinson, R. L., Palczewska, A., Palczewski, J., and Kidley, N. (2017). Comparison of the Predictive Performance and Interpretability of Random Forest and Linear Models on Benchmark Data Sets. *J. Chem. Inf. Model.* 57, 1773–1792. doi:10.1021/acs.jcim.6b00753

Mason, L., Baxter, J., Bartlett, P., and Frean, M. (1999). Boosting Algorithms as Gradient Descent in Function Space. *Proc. NIPS.* 12, 512–518.

Mayr, L. M., and Bojanic, D. (2009). Novel Trends in High-Throughput Screening. *Curr. Opin. Pharmacol.* 9, 580–588. doi:10.1016/j.coph.2009.08.004

McCloskey, K., Taly, A., Monti, F., Brenner, M. P., and Colwell, L. J. (2019). Using Attribution to Decode Binding Mechanism in Neural Network Models for Chemistry. *Proc. Natl. Acad. Sci. U. S. A.* 116, 11624–11629. doi:10.1073/pnas.1820657116

McCorkindale, W., Poelking, C., and Lee, A. A. (2020). *Investigating 3D Atomic Environments for Enhanced QSAR. arXiv preprint arXiv:2010.12857*.

McCulloch, W. S., and Pitts, W. (1943). A Logical Calculus of the Ideas Immanent in Nervous Activity. *Bull. Math. Biophysics* 5, 115–133. doi:10.1007/bf02478259

McNutt, A. T., Francoeur, P., Aggarwal, R., Masuda, T., Meli, R., Ragoza, M., et al. (2021). GNINA 1.0: Molecular Docking with Deep Learning. *J. Cheminform* 13, 1–20. doi:10.1186/s13321-021-00522-2

McNutt, A. T., and Koes, D. R. (2022). Improving ΔΔG Predictions with a Multitask Convolutional Siamese Network. *J. Chem. Inf. Model.* 62, 1819–1829. doi:10.1021/acs.jcim.1c01497

Meli, R., Anighoro, A., Bodkin, M. J., Morris, G. M., and Biggin, P. C. (2021). Learning Protein-Ligand Binding Affinity with Atomic Environment Vectors. *J. Cheminform* 13, 1–19. doi:10.1186/s13321-021-00536-w

Mendez, D., Gaulton, A., Bento, A. P., Chambers, J., De Veij, M., Félix, E., et al. (2018). ChEMBL: Towards Direct Deposition of Bioassay Data. *Nucleic Acids Res.* 47, D930–D940. doi:10.1093/nar/gky1075

Meng, E. C., Shoichet, B. K., and Kuntz, I. D. (1992). Automated Docking with Grid-Based Energy Evaluation. *J. Comput. Chem.* 13, 505–524. doi:10.1002/jcc.540130412

Menke, J., and Koch, O. (2021). Using Domain-specific Fingerprints Generated through Neural Networks to Enhance Ligand-Based Virtual Screening. *J. Chem. Inf. Model.* 61, 664–675. doi:10.1021/acs.jcim.0c01208

Mey, A. S., Allen, B., Macdonald, H. E. B., Chodera, J. D., Kuhn, M., Michel, J., et al. (2020). *Best Practices for Alchemical Free Energy Calculations. arXiv preprint arXiv:2008.03067*.

Meyers, J., Fabian, B., and Brown, N. (2021). De Novo molecular Design and Generative Models. *Drug Discov. Today* 26, 2707–2715. doi:10.1016/j.drudis.2021.05.019

Mobley, D. L., Graves, A. P., Chodera, J. D., McReynolds, A. C., Shoichet, B. K., and Dill, K. A. (2007). Predicting Absolute Ligand Binding Free Energies to a Simple Model Site. *J. Mol. Biol.* 371, 1118–1134. doi:10.1016/j.jmb.2007.06.002

Moesser, M. A., Klein, D., Boyles, F., Deane, C. M., Baxter, A., and Morris, G. M. (2022). Protein-ligand Interaction Graphs: Learning from Ligand-Shaped 3d Interaction Graphs to Improve Binding Affinity Prediction. *bioRxiv*.

Monti, F., Boscaini, D., Masci, J., Rodola, E., Svoboda, J., and Bronstein, M. M. (2017). "Geometric Deep Learning on Graphs and Manifolds Using Mixture Model CNNs," in *2017 IEEE Conference on Computer Vision and Pattern Recognition* (New York: IEEE), 5115. –5124. doi:10.1109/cvpr.2017.576

Monticelli, L., and Tieleman, D. P. (2012). "Force Fields for Classical Molecular Dynamics," in *Methods in Molecular Biology* (Totowa, NJ: Humana Press), 197–213. doi:10.1007/978-1-62703-017-5_8

Moon, S., Zhung, W., Yang, S., Lim, J., and Kim, W. Y. (2022). PIGNet: a Physics-Informed Deep Learning Model toward Generalized Drug-Target Interaction Predictions. *Chem. Sci.* 13, 3661–3673. doi:10.1039/d1sc06946b

Morris, G. M., Goodsell, D. S., Huey, R., and Olson, A. J. (1996). Distributed Automated Docking of Flexible Ligands to Proteins: Parallel Applications of AutoDock 2.4. *J. Comput. Aided Mol. Des.* 10, 293–304. doi:10.1007/bf00124499

Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S., et al. (2009). AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *J. Comput. Chem.* 30, 2785–2791. doi:10.1002/jcc.21256

Morris, G. M., Goodsell, D. S., Halliday, R. S., Huey, R., Hart, W. E., Belew, R. K., et al. (1998). Automated Docking Using a Lamarckian Genetic Algorithm and an Empirical Binding Free Energy Function. *J. Comput. Chem.* 19, 1639–1662. doi:10.1002/(sici)1096-987x(19981115)19:14<1639::aid-jcc10>3.0.co;2-b

Morrone, J. A., Weber, J. K., Huynh, T., Luo, H., and Cornell, W. D. (2020). Combining Docking Pose Rank and Structure with Deep Learning Improves Protein-Ligand Binding Mode Prediction over a Baseline Docking Approach. *J. Chem. Inf. Model.* 60, 4170–4179. doi:10.1021/acs.jcim.9b00927

Moustakas, D. T., Lang, P. T., Pegg, S., Pettersen, E., Kuntz, I. D., Brooijmans, N., et al. (2006). Development and Validation of a Modular, Extensible Docking Program: Dock 5. *J. Comput. Aided Mol. Des.* 20, 601–619. doi:10.1007/s10822-006-9060-4

Muegge, I., and Martin, Y. C. (1999). A General and Fast Scoring Function for Protein-Ligand Interactions: a Simplified Potential Approach. *J. Med. Chem.* 42, 791–804. doi:10.1021/jm980536j

Muegge, I. (2000). A Knowledge-Based Scoring Function for Protein-Ligand Interactions: Probing the Reference State. *Perspect. Drug Discov.* 20, 99–114. doi:10.1023/a:1008729005958

Muegge, I. (2001). Effect of Ligand Volume Correction on PMF Scoring. *J. Comput. Chem.* 22, 418–425. doi:10.1002/1096-987x(200103)22:4<418:aid-jcc1012>3.0.co;2-3

Muegge, I., and Rarey, M. (2001). "Small Molecule Docking and Scoring," in *Reviews in Computational Chemistry* (John Wiley & Sons), 1–60. doi:10.1002/0471224413.ch1

Müller, S., Ackloo, S., Al Chawaf, A., Al-Lazikani, B., Antolin, A., Baell, J. B., et al. (2022). Target 2035 - Update on the Quest for a Probe for Every Protein. *RSC Med. Chem.* 13, 13–21. doi:10.1039/d1md00228g

Muratov, E. N., Bajorath, J., Sheridan, R. P., Tetko, I. V., Filimonov, D., Poroikov, V., et al. (2020). QSAR without Borders. *Chem. Soc. Rev.* 49, 3525–3564. doi:10.1039/d0cs00098a

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., and Yu, B. (2019). Definitions, Methods, and Applications in Interpretable Machine Learning. *Proc. Natl. Acad. Sci. U. S. A.* 116, 22071–22080. doi:10.1073/pnas.1900654116

Musil, F., Grisafi, A., Bartók, A. P., Ortner, C., Csányi, G., and Ceriotti, M. (2021). Physics-inspired Structural Representations for Molecules and Materials. *Chem. Rev.* 121, 9759–9815. doi:10.1021/acs.chemrev.1c00021

Mysinger, M. M., Carchia, M., Irwin, J. J., and Shoichet, B. K. (2012). Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *J. Med. Chem.* 55, 6582–6594. doi:10.1021/jm300687e

Narkhede, M. V., Bartakke, P. P., and Sutaone, M. S. (2021). A Review on Weight Initialization Strategies for Neural Networks. *Artif. Intell. Rev.* 55, 291–322. doi:10.1007/s10462-021-10033-z

Neudert, G., and Klebe, G. (2011). DSX: a Knowledge-Based Scoring Function for the Assessment of Protein-Ligand Complexes. *J. Chem. Inf. Model.* 51, 2731–2745. doi:10.1021/ci200274q

Nguyen, D. D., and Wei, G. W. (2019). AGL-score: Algebraic Graph Learning Score for Protein-Ligand Binding Scoring, Ranking, Docking, and Screening. *J. Chem. Inf. Model.* 59, 3291–3304. doi:10.1021/acs.jcim.9b00334

Nguyen, T., Le, H., Quinn, T. P., Nguyen, T., Le, T. D., and Venkatesh, S. (2020). GraphDTA: Predicting Drug-Target Binding Affinity with Graph Neural Networks. *Bioinformatics* 37, 1140–1147. doi:10.1093/bioinformatics/btaa921

Nicholls, A. (2014). Confidence Limits, Error Bars and Method Comparison in Molecular Modeling. Part 1: The Calculation of Confidence Intervals. *J. Comput. Aided Mol. Des.* 28, 887–918. doi:10.1007/s10822-014-9753-z

Nicholls, A. (2016). Confidence Limits, Error Bars and Method Comparison in Molecular Modeling. Part 2: Comparing Methods. *J. Comput. Aided Mol. Des.* 30, 103–126. doi:10.1007/s10822-016-9904-5

Nogueira, M. S., and Koch, O. (2019). The Development of Target-specific Machine Learning Models as Scoring Functions for Docking-Based Target Prediction. *J. Chem. Inf. Model.* 59, 1238–1252. doi:10.1021/acs.jcim.8b00773

Öztürk, H., Özgür, A., and Ozkirimli, E. (2018). DeepDTA: Deep Drug-Target Binding Affinity Prediction. *Bioinformatics* 34, i821–i829. doi:10.1093/bioinformatics/bty593

Palazzesi, F., and Pozzan, A. (2022). Deep Learning Applied to Ligand-Based. *Artif. Intell. Drug Des.*, 273–299. doi:10.1007/978-1-0716-1787-8_12

Pan, S. J., and Yang, Q. (2010). A Survey on Transfer Learning. *IEEE Trans. Knowl. Data Eng.* 22, 1345–1359. doi:10.1109/tkde.2009.191

Pan, X., Wang, H., Zhang, Y., Wang, X., Li, C., Ji, C., et al. (2022). Aa-score: a New Scoring Function Based on Amino Acid-specific Interaction for Molecular Docking. *J. Chem. Inf. Model.* doi:10.1021/acs.jcim.1c01537

Parks, C. D., Gaieb, Z., Chiu, M., Yang, H., Shao, C., Walters, W. P., et al. (2020). D3R Grand Challenge 4: Blind Prediction of Protein-Ligand Poses, Affinity Rankings, and Relative Binding Free Energies. *J. Comput. Aided Mol. Des.* 34, 99–119. doi:10.1007/s10822-020-00289-y

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). Pytorch: An Imperative Style, High-Performance Deep Learning Library. *Adv. Neural Inf. Process Syst.* 32, 8026–8037.

Patrick Walters, W. (2021). Comparing Classification Models-A Practical Tutorial. *J. Comput. Aided Mol. Des.* 1. doi:10.1007/s10822-021-00417-2

Pearlman, D. A., and Charifson, P. S. (2001). Are Free Energy Calculations Useful in Practice? a Comparison with Rapid Scoring Functions for the P38 MAP Kinase Protein System. *J. Med. Chem.* 44, 3417–3423. doi:10.1021/jm0100279

Pérez-Nueno, V. I., Rabal, O., Borrell, J. I., and Teixidó, J. (2009). APIF: A New Interaction Fingerprint Based on Atom Pairs and its Application to Virtual Screening. *J. Chem. Inf. Model.* 49, 1245–1260. doi:10.1021/ci900043r

Preuer, K., Klambauer, G., Rippmann, F., Hochreiter, S., and Unterthiner, T. (2019). "Interpretable Deep Learning in Drug Discovery," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (Springer International Publishing), 331–345. doi:10.1007/978-3-030-28954-6_18

Pu, L., Govindaraj, R. G., Lemoine, J. M., Wu, H. C., and Brylinski, M. (2019). DeepDrug3D: Classification of Ligand-Binding Pockets in Proteins with a Convolutional Neural Network. *PLOS Comput. Biol.* 15, e1006718. doi:10.1371/journal.pcbi.1006718

Qiao, Z., Welborn, M., Anandkumar, A., Manby, F. R., and Miller, T. F. (2020). OrbNet: Deep Learning for Quantum Chemistry Using Symmetry-Adapted Atomic-Orbital Features. *J. Chem. Phys.* 153, 124111. doi:10.1063/5.0021955

Qiu, T., Qiu, J., Feng, J., Wu, D., Yang, Y., Tang, K., et al. (2016). The Recent Progress in Proteochemometric Modelling: Focusing on Target Descriptors, Cross-Term Descriptors and Application Scope. *Brief. Bioinform.* 18, 125–136. doi:10.1093/bib/bbw004

Quiroga, R., and Villarreal, M. A. (2016). Vinardo: A Scoring Function Based on Autodock Vina Improves Scoring, Docking, and Virtual Screening. *PloS one* 11, e0155183. doi:10.1371/journal.pone.0155183

Radifar, M., Yuniarti, N., and Istyastono, E. P. (2013). PyPLIF: Python-Based Protein-Ligand Interaction Fingerprinting. *Bioinformation* 9, 325–328. doi:10.6026/97320630009325

Ragoza, M., Hochuli, J., Idrobo, E., Sunseri, J., and Koes, D. R. (2017a). Protein-Ligand Scoring with Convolutional Neural Networks. *J. Chem. Inf. Model.* 57, 942–957. doi:10.1021/acs.jcim.6b00740

Ragoza, M., Turner, L., and Koes, D. R. (2017b). *Ligand Pose Optimization with Atomic Grid-Based Convolutional Neural Networks. arXiv preprint arXiv: 1710.07400*.

Ramsundar, B., Liu, B., Wu, Z., Verras, A., Tudor, M., Sheridan, R. P., et al. (2017). Is Multitask Deep Learning Practical for Pharma? *J. Chem. Inf. Model.* 57, 2068–2076. doi:10.1021/acs.jcim.7b00146

Rännar, S., Geladi, P., Lindgren, F., and Wold, S. (1995). A PLS Kernel Algorithm for Data Sets with Many Variables and Few Objects. Part II: Cross-Validation, Missing Data and Examples. *J. Chemom.* 9, 459–470. doi:10.1002/cem.1180090604

Rännar, S., Lindgren, F., Geladi, P., and Wold, S. (1994). A PLS Kernel Algorithm for Data Sets with Many Variables and Fewer Objects. Part 1: Theory and Algorithm. *J. Chemom.* 8, 111–125. doi:10.1002/cem.1180080204

Rasmussen, C. E. (2003). "Gaussian Processes in Machine Learning," in *Summer School on Machine Learning* (Springer), 63–71.

Reymond, J.-L., van Deursen, R., Blum, L. C., and Ruddigkeit, L. (2010). Chemical Space as a Source for New Drugs. *Med. Chem. Commun.* 1, 30. doi:10.1039/c0md00020e

Rifaioglu, A. S., Atas, H., Martin, M. J., Cetin-Atalay, R., Atalay, V., and Doğan, T. (2019). Recent Applications of Deep Learning and Machine Intelligence on In Silico Drug Discovery: Methods, Tools and Databases. *Brief. Bioinform* 20, 1878–1912. doi:10.1093/bib/bby061

Riniker, S., and Landrum, G. A. (2013a). Open-source Platform to Benchmark Fingerprints for Ligand-Based Virtual Screening. *J. Cheminform* 5, 26–17. doi:10.1186/1758-2946-5-26

Riniker, S., and Landrum, G. A. (2013b). Similarity Maps - a Visualization Strategy for Molecular Fingerprints and Machine-Learning Methods. *J. Cheminform* 5, 43–47. doi:10.1186/1758-2946-5-43

Rodríguez-Pérez, R., and Bajorath, J. (2019). Interpretation of Compound Activity Predictions from Complex Machine Learning Models Using Local Approximations and Shapley Values. *J. Med. Chem.* 63, 8761–8777. doi:10.1021/acs.jmedchem.9b01101

Rogers, D., and Hahn, M. (2010). Extended-connectivity Fingerprints. *J. Chem. Inf. Model.* 50, 742–754. doi:10.1021/ci100050t

Roitberg, A., Pollert, T., Haurilet, M., Martin, M., and Stiefelhagen, R. (2019). "Analysis of Deep Fusion Strategies for Multi-Modal Gesture Recognition," in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (IEEE). 0–0. doi:10.1109/cvprw.2019.00029

Rosenblatt, F. (1962). *Perceptions and the Theory of Brain Mechanisms*. Spartan books.

Ross, G. A., Morris, G. M., and Biggin, P. C. (2013). One Size Does Not Fit All: The Limits of Structure-Based Models in Drug Discovery. *J. Chem. Theory Comput.* 9, 4266–4274. doi:10.1021/ct4004228

Ross, G. A., Morris, G. M., and Biggin, P. C. (2012). Rapid and Accurate Prediction and Scoring of Water Molecules in Protein Binding Sites. *PLoS ONE* 7, e32036. doi:10.1371/journal.pone.0032036

Rufa, D. A., Macdonald, H. E. B., Fass, J., Wieder, M., Grinaway, P. B., Roitberg, A. E., et al. (2020). Towards Chemical Accuracy for Alchemical Free Energy Calculations with Hybrid Physics-Based Machine Learning/molecular Mechanics Potentials. *BioRxiv*.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning Representations by Back-Propagating Errors. *Nature* 323, 533–536. doi:10.1038/323533a0

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* 115, 211–252. doi:10.1007/s11263-015-0816-y

Ryu, S., Lim, J., Hong, S. H., and Kim, W. Y. (2018). *Deeply Learning Molecular Structure-Property Relationships Using Attention-And Gate-Augmented Graph Convolutional Network*. *arXiv preprint arXiv:1805.10988*.

Salt, D. W., Yildiz, N., Livingstone, D. J., and Tinsley, C. J. (1992). The Use of Artificial Neural Networks in QSAR. *Pestic. Sci.* 36, 161–170. doi:10.1002/ps.2780360212

Scantlebury, J., Brown, N., Von Delft, F., and Deane, C. M. (2020). 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.* 60, 3722–3730. doi:10.1021/acs.jcim.0c00263

Schäfer, M., Oeing, C. U., Rohm, M., Baysal-Temel, E., Lehmann, L. H., Bauer, R., et al. (2020). 'Corrigendum to "Ataxin-10 Is Part of a Cachexokine Cocktail Triggering Cardiac Metabolic Dysfunction in Cancer Cachexia" [Molecular Metabolism 5 (2) (2015) 67-78]'. *Mol. Metab.* 35, 100970. doi:10.1016/j.molmet.2020.02.013

Schneider, G., and Clark, D. E. (2019). Automated De Novo Drug Design: Are We Nearly There yet? *Angew. Chem. Int. Ed. Engl.* 58, 10792–10803. doi:10.1002/anie.201814681

Schneider, P., and Schneider, G. (2016). De Novo design at the Edge of Chaos. *J. Med. Chem.* 59, 4077–4086. doi:10.1021/acs.jmedchem.5b01849

Schneider, P., Walters, W. P., Plowright, A. T., Sieroka, N., Listgarten, J., Goodnow, R. A., et al. (2019). Rethinking Drug Design in the Artificial Intelligence Era. *Nat. Rev. Drug Discov.* 19, 353–364. doi:10.1038/s41573-019-0050-3

Schütt, K. T., Sauceda, H. E., Kindermans, P. J., Tkatchenko, A., and Müller, K. R. (2018). SchNet - A Deep Learning Architecture for Molecules and Materials. *J. Chem. Phys.* 148, 241722. doi:10.1063/1.5019779

Seo, S., Choi, J., Park, S., and Ahn, J. (2021). Binding Affinity Prediction for Protein-Ligand Complex Using Deep Attention Mechanism Based on Intermolecular Interactions. *BMC Bioinforma.* 22. doi:10.1186/s12859-021-04466-0

Shen, C., Hu, Y., Wang, Z., Zhang, X., Pang, J., Wang, G., et al. (2020a). Beware of the Generic Machine Learning-Based Scoring Functions in Structure-Based Virtual Screening. *Brief. Bioinform.* 22, bbaa070. doi:10.1093/bib/bbaa070

Shen, C., Hu, Y., Wang, Z., Zhang, X., Zhong, H., Wang, G., et al. (2020b). Can Machine Learning Consistently Improve the Scoring Power of Classical Scoring Functions? Insights into the Role of Machine Learning in Scoring Functions. *Brief. Bioinform.* 22, 497–514. doi:10.1093/bib/bbz173

Shen, C., Ding, J., Wang, Z., Cao, D., Ding, X., and Hou, T. (2019). From Machine Learning to Deep Learning: Advances in Scoring Functions for Protein-Ligand Docking. *WIREs Comput. Mol. Sci.* 10, e1429. doi:10.1002/wcms.1429

Shen, C., Hu, X., Gao, J., Zhang, X., Zhong, H., Wang, Z., et al. (2021). The Impact of Cross-Docked Poses on Performance of Machine Learning Classifier for Protein-Ligand Binding Pose Prediction. *J. Cheminform* 13, 1–18. doi:10.1186/s13321-021-00560-w

Sheridan, R. P. (2019). Interpretation of QSAR Models by Coloring Atoms According to Changes in Predicted Activity: How Robust Is it? *J. Chem. Inf. Model.* 59, 1324–1337. doi:10.1021/acs.jcim.8b00825

Shin, B., Park, S., Kang, K., and Ho, J. C. (2019). "Self-attention Based Molecule Representation for Predicting Drug-Target Interaction," in *Proceedings of the 4th Machine Learning for Healthcare Conference. Proceedings of Machine Learning Research* (PMLR 106), 230–248.

Shin, W. H., Kim, J. K., Kim, D. S., and Seok, C. (2013). GalaxyDock2: Protein-Ligand Docking Using Beta-Complex and Global Optimization. *J. Comput. Chem.* 34, 2647–2656. doi:10.1002/jcc.23438

Shin, W. H., and Seok, C. (2012). GalaxyDock: Protein-Ligand Docking with Flexible Protein Side-Chains. *J. Chem. Inf. Model.* 52, 3225–3232. doi:10.1021/ci300342z

Shoichet, B. K., Kuntz, I. D., and Bodian, D. L. (1992). Molecular Docking Using Shape Descriptors. *J. Comput. Chem.* 13, 380–397. doi:10.1002/jcc.540130311

Sieg, J., Flachsenberg, F., and Rarey, M. (2019). In Need of Bias Control: Evaluating Chemical Data for Machine Learning in Structure-Based Virtual Screening. *J. Chem. Inf. Model.* 59, 947–961. doi:10.1021/acs.jcim.8b00712

Sliwoski, G., Kothiwale, S., Meiler, J., and Lowe, E. W. (2013). Computational Methods in Drug Discovery. *Pharmacol. Rev.* 66, 334–395. doi:10.1124/pr.112.007336

Smith, J. S., Isayev, O., and Roitberg, A. E. (2017). ANI-1: An Extensible Neural Network Potential with DFT Accuracy at Force Field Computational Cost. *Chem. Sci.* 8, 3192–3203. doi:10.1039/c6sc05720a

Smith, R. D., Clark, J. J., Ahmed, A., Orban, Z. J., Dunbar, J. B., and Carlson, H. A. (2019). Updates to Binding MOAD (Mother of All Databases): Polypharmacology Tools and Their Utility in Drug Repurposing. *J. Mol. Biol.* 431, 2423–2433. doi:10.1016/j.jmb.2019.05.024

Smith, R. D., Damm-Ganamet, K. L., Dunbar, J. B., Ahmed, A., Chinnaswamy, K., Delproposto, J. E., et al. (2015). CSAR Benchmark Exercise 2013: Evaluation of Results from a Combined Computational Protein Design, Docking, and Scoring/Ranking Challenge. *J. Chem. Inf. Model.* 56, 1022–1031. doi:10.1021/acs.jcim.5b00387

Smith, R. D., Dunbar, J. B., Ung, P. M., Esposito, E. X., Yang, C. Y., Wang, S., et al. (2011). CSAR Benchmark Exercise of 2010: Combined Evaluation across All Submitted Scoring Functions. *J. Chem. Inf. Model.* 51, 2115–2131. doi:10.1021/ci200269q

Soleimany, A. P., Amini, A., Goldman, S., Rus, D., Bhatia, S. N., and Coley, C. W. (2021). Evidential Deep Learning for Guided Molecular Property Prediction and Discovery. *ACS Cent. Sci.* 7, 1356–1367. doi:10.1021/acscentsci.1c00546

Son, J., and Kim, D. (2021). Development of a Graph Convolutional Neural Network Model for Efficient Prediction of Protein-Ligand Binding Affinities. *PLoS One* 16, e0249404. doi:10.1371/journal.pone.0249404

Sotriffer, C. A., Sanschagrin, P., Matter, H., and Klebe, G. (2008). SFCscore: Scoring Functions for Affinity Prediction of Protein-Ligand Complexes. *Proteins* 73, 395–419. doi:10.1002/prot.22058

Spearman, C. (2010). The Proof and Measurement of Association between Two Things. *Int. J. Epidemiol.* 39, 1137–1150. doi:10.1093/ije/dyq191

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* 15, 1929–1958.

Stafford, K. A., Anderson, B. M., Sorenson, J., and van den Bedem, H. (2022). AtomNet PoseRanker: Enriching Ligand Pose Quality for Dynamic Proteins in Virtual High-Throughput Screens. *J. Chem. Inf. Model.* 62, 1178–1189. doi:10.1021/acs.jcim.1c01250

Stepniewska-Dziubinska, M. M., Zielenkiewicz, P., and Siedlecki, P. (2018). Development and Evaluation of a Deep Learning Model for Protein-Ligand Binding Affinity Prediction. *Bioinformatics* 34, 3666–3674. doi:10.1093/bioinformatics/bty374

Štrumbelj, E., Kononenko, I., and Robnik Šikonja, M. (2009). Explaining Instance Classifications with Interactions of Subsets of Feature Values. *Data & Knowl. Eng.* 68, 886–904. doi:10.1016/j.datak.2009.01.004

Su, M., Feng, G., Liu, Z., Li, Y., and Wang, R. (2020). Tapping on the Black Box: How Is the Scoring Power of a Machine-Learning Scoring Function Dependent on the Training Set? *J. Chem. Inf. Model.* 60, 1122–1136. doi:10.1021/acs.jcim.9b00714

Su, M., Yang, Q., Du, Y., Feng, G., Liu, Z., Li, Y., et al. (2018). Comparative Assessment of Scoring Functions: The CASF-2016 Update. *J. Chem. Inf. Model.* 59, 895–913. doi:10.1021/acs.jcim.8b00545

Sundararajan, M., Taly, A., and Yan, Q. (2017). "Axiomatic Attribution for Deep Networks," in *Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research* (PMLR), 3319–3328.

Sunseri, J., King, J. E., Francoeur, P. G., and Koes, D. R. (2018). Convolutional Neural Network Scoring and Minimization in the D3R 2017 Community Challenge. *J. Comput. Aided Mol. Des.* 33, 19–34. doi:10.1007/s10822-018-0133-y

Sunseri, J., and Koes, D. R. (2020). Libmolgrid: Graphics Processing Unit Accelerated Molecular Gridding for Deep Learning Applications. *J. Chem. Inf. Model.* 60, 1079–1084. doi:10.1021/acs.jcim.9b01145

Szegedy, C., Toshev, A., and Erhan, D. (2013). *Deep Neural Networks for Object Detection*.

Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., and Liu, C. (2018). "A Survey on Deep Transfer Learning," in *International Conference on Artificial Neural Networks* (Springer), 270–279. doi:10.1007/978-3-030-01424-7_27

Tin Kam Ho, T. K. (1995). "Random Decision Forests," in Proceedings of 3rd International Conference on Document Analysis and Recognition (IEEE Comput. Soc. Press), 1, 278–282. doi:10.1109/icdar.1995.598994

Tin Kam Ho, T. K. (1998). The Random Subspace Method for Constructing Decision Forests. *IEEE Trans. Pattern Anal. Mach. Intell.* 20, 832–844. doi:10.1109/34.709601

Tropsha, A. (2010). Best Practices for QSAR Model Development, Validation, and Exploitation. *Mol. Inf.* 29, 476–488. doi:10.1002/minf.201000061

Trott, O., and Olson, A. J. (2009). AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *J. Comput. Chem.* 31, 455–461. NA–NA. doi:10.1002/jcc.21334

Unke, O. T., Chmiela, S., Sauceda, H. E., Gastegger, M., Poltavsky, I., Schütt, K. T., et al. (2021). Machine Learning Force Fields. *Chem. Rev.* 121, 10142–10186. doi:10.1021/acs.chemrev.0c01111

Unterthiner, T., Mayr, A., Klambauer, G., Steijaert, M., Wegner, J. K., Ceulemans, H., et al. (2014). Deep Learning as an Opportunity in Virtual Screening. *Proc. deep Learn. workshop A. T. NIPS* 27, 1–9.

Vainio, M. J., Puranen, J. S., and Johnson, M. S. (2009). ShaEP: Molecular Overlay Based on Shape and Electrostatic Potential. *J. Chem. Inf. Model.* 49, 492–502. doi:10.1021/ci800315d

Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., et al. (2019). Applications of Machine Learning in Drug Discovery and Development. *Nat. Rev. Drug Discov.* 18, 463–477. doi:10.1038/s41573-019-0024-5

van Westen, G. J., Wegner, J. K., Geluykens, P., Kwanten, L., Vereycken, I., Peeters, A., et al. (2011). Which Compound to Select in Lead Optimization? Prospectively Validated Proteochemometric Models Guide Preclinical Development. *PLoS ONE* 6, e27518. doi:10.1371/journal.pone.0027518

Varela-Rial, A., Maryanow, I., Majewski, M., Doerr, S., Schapin, N., Jiménez-Luna, J., et al. (2022). PlayMolecule Glimpse: Understanding Protein-Ligand Property Predictions with Interpretable Neural Networks. *J. Chem. Inf. Model.* 62, 225–231. doi:10.1021/acs.jcim.1c00691

Velec, H. F., Gohlke, H., and Klebe, G. (2005). 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.* 48, 6296–6303. doi:10.1021/jm050436v

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., and Bengio, Y. (2017). *Graph Attention Networks. arXiv preprint arXiv:1710.10903*.

Verdonk, M. L., Cole, J. C., Hartshorn, M. J., Murray, C. W., and Taylor, R. D. (2003). Improved Protein-Ligand Docking Using GOLD. *Proteins* 52, 609–623. doi:10.1002/prot.10465

Voulodimos, A., Doulamis, N., Doulamis, A., and Protopapadakis, E. (2018). Deep Learning for Computer Vision: A Brief Review. *Comput. Intell. Neurosci.* 2018, 7068349–7068413. doi:10.1155/2018/7068349

Wallach, I., Dzamba, M., and Heifets, A. (2015). *AtomNet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-Based Drug Discovery. arXiv preprint arXiv:1510.02855*, 1-13.

Wang, C., and Zhang, Y. (2016). Improving Scoring-Docking-Screening Powers of Protein-Ligand Scoring Functions Using Random Forest. *J. Comput. Chem.* 38, 169–177. doi:10.1002/jcc.24667

Wang, R., Fang, X., Lu, Y., and Wang, S. (2004). The PDBbind Database: Collection of Binding Affinities for Protein-Ligand Complexes with Known Three-Dimensional Structures. *J. Med. Chem.* 47, 2977–2980. doi:10.1021/jm030580l

Wang, R., Fang, X., Lu, Y., Yang, C. Y., and Wang, S. (2005). The PDBbind Database: Methodologies and Updates. *J. Med. Chem.* 48, 4111–4119. doi:10.1021/jm048957q

Wang, R., Lai, L., and Wang, S. (2002). Further Development and Validation of Empirical Scoring Functions for Structure-Based Binding Affinity Prediction. *J. Comput. Aided Mol. Des.* 16, 11–26. doi:10.1023/a:1016357811882

Wang, R., and Wang, S. (2001). How Does Consensus Scoring Work for Virtual Library Screening? an Idealized Computer Experiment. *J. Chem. Inf. Comput. Sci.* 41, 1422–1426. doi:10.1021/ci010025x

Wang, S., Liu, D., Ding, M., Du, Z., Zhong, Y., Song, T., et al. (2021b1805). Seonionnet: a Convolution Neural Network for Protein–Ligand Binding Affinity Prediction *Front. Genet.*

Wang, S., Liu, D., Ding, M., Du, Z., Zhong, Y., Song, T., et al. (2021a). SE-OnionNet: A Convolution Neural Network for Protein-Ligand Binding Affinity Prediction. *Front. Genet.* 11, 1805. doi:10.3389/fgene.2020.607824

Wang, S., and Riniker, S. (2020). Chapter 9. Machine Learning in the Area of Molecular Dynamics Simulations. *Artif. Intell. Drug Discov.* 75, 184–214. doi:10.1039/9781788016841-00184

Wang, Y., Xiao, J., Suzek, T. O., Zhang, J., Wang, J., and Bryant, S. H. (2009). PubChem: A Public Information System for Analyzing Bioactivities of Small Molecules. *Nucleic Acids Res.* 37, W623–W633. doi:10.1093/nar/gkp456

Wang, Y., Xiao, J., Suzek, T. O., Zhang, J., Wang, J., Zhou, Z., et al. (2011). PubChem's BioAssay Database. *Nucleic Acids Res.* 40, D400–D412. doi:10.1093/nar/gkr1132

Wang, Y., Wu, S., Duan, Y., and Huang, Y. (2021c). A Point Cloud-Based Deep Learning Strategy for Protein-Ligand Binding Affinity Prediction. *Brief. Bioinform.* 23. doi:10.1093/bib/bbab474

Wang, Z., Zheng, L., Liu, Y., Qu, Y., Li, Y.-Q., Zhao, M., et al. (2021d). OnionNet-2: A Convolutional Neural Network Model for Predicting Protein-Ligand Binding Affinity Based on Residue-Atom Contacting Shells. *Front. Chem.* 9, 913. doi:10.3389/fchem.2021.753002

Wee, J., and Xia, K. (2021). Ollivier Persistent Ricci Curvature-Based Machine Learning for the Protein-Ligand Binding Affinity Prediction. *J. Chem. Inf. Model.* 61, 1617–1626. doi:10.1021/acs.jcim.0c01415

Weiler, M., Geiger, M., Welling, M., Boomsma, W., and Cohen, T. (2018). *3d Steerable Cnns: Learning Rotationally Equivariant Features in Volumetric Data. arXiv preprint arXiv:1807.02547*.

Wellawatte, G. P., Seshadri, A., and White, A. D. (2022). Model Agnostic Generation of Counterfactual Explanations for Molecules. *Chem. Sci.* doi:10.1039/d1sc05259d

Widrow, B., and Hoff, M. E. (1960). *Adaptive Switching Circuits. Tech. Rep.* Stanford Univ Ca Stanford Electronics Labs. doi:10.21236/ad0241531

Wieder, M., Fass, J., and Chodera, J. D. (2021). *Teaching Free Energy Calculations to Learn from Experimental Data*. bioRxiv.

Williams, C. K., and Rasmussen, C. E. (1996). *Gaussian Processes for Regression*.

Winkler, D. A., and Le, T. C. (2016). Performance of Deep and Shallow Neural Networks, the Universal Approximation Theorem, Activity Cliffs, and QSAR. *Mol. Inf.* 36, 1600118. doi:10.1002/minf.201600118

Wójcikowski, M., Kukiełka, M., Stepniewska-Dziubinska, M. M., and Siedlecki, P. (2018). Development of a Protein-Ligand Extended Connectivity (PLEC) Fingerprint and its Application for Binding Affinity Predictions. *Bioinformatics* 35, 1334–1341. doi:10.1093/bioinformatics/bty757

Wójcikowski, M., Zielenkiewicz, P., and Siedlecki, P. (2015). Open Drug Discovery Toolkit (ODDT): A New Open-Source Player in the Drug Discovery Field. *J. Cheminform* 7, 26–6. doi:10.1186/s13321-015-0078-2

Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., et al. (2018). MoleculeNet: A Benchmark for Molecular Machine Learning. *Chem. Sci.* 9, 513–530. doi:10.1039/c7sc02664a

Xavier, M. M., Heck, G. S., Avila, M. B., Levin, N. M. B., Pintro, V. O., Carvalho, N. L., et al. (2016). SAnDReS a Computational Tool for Statistical Analysis of Docking Results and Development of Scoring Functions. *Comb. Chem. High. Throughput Screen* 19, 801–812. doi:10.2174/1386207319666160927111347

Xiong, G., Shen, C., Yang, Z., Jiang, D., Liu, S., Lu, A., et al. (2021). Featurization Strategies for Protein-Ligand Interactions and Their Applications in Scoring Function Development. *WIREs Comput. Mol. Sci.* 12, e1567. doi:10.1002/wcms.1567

Xu, Y., Verma, D., Sheridan, R. P., Liaw, A., Ma, J., Marshall, N. M., et al. (2020). Deep Dive into Machine Learning Models for Protein Engineering. *J. Chem. Inf. Model.* 60, 2773–2790. doi:10.1021/acs.jcim.0c00073

Yakovenko, O., and Jones, S. J. M. (2017). Modern Drug Design: The Implication of Using Artificial Neuronal Networks and Multiple Molecular Dynamic Simulations. *J. Comput. Aided Mol. Des.* 32, 299–311. doi:10.1007/s10822-017-0085-7

Yang, C., and Zhang, Y. (2021). Lin_F9: A Linear Empirical Scoring Function for Protein-Ligand Docking. *J. Chem. Inf. Model.* 61, 4630–4644. doi:10.1021/acs.jcim.1c00737

Yang, C. Y., Wang, R., and Wang, S. (2005). M-score: A Knowledge-Based Potential Scoring Function Accounting for Protein Atom Mobility. *J. Med. Chem.* 49, 5903–5911. doi:10.1021/jm050043w

Yang, J., Shen, C., and Huang, N. (2020). Predicting or Pretending: Artificial Intelligence for Protein-Ligand Interactions Lack of Sufficiently Large and Unbiased Datasets. *Front. Pharmacol.* 11, 69. doi:10.3389/fphar.2020.00069

Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., et al. (2019). Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* 59, 3370–3388. doi:10.1021/acs.jcim.9b00237

Yang, Z., Zhong, W., Zhao, L., and Yu-Chian Chen, C. (2022). MGraphDTA: Deep Multiscale Graph Neural Network for Explainable Drug-Target Binding Affinity Prediction. *Chem. Sci.* 13, 816–833. doi:10.1039/d1sc05180f

Yeturu, K., and Chandra, N. (2008). PocketMatch: A New Algorithm to Compare Binding Sites in Protein Structures. *BMC Bioinforma.* 9, 1–17. doi:10.1186/1471-2105-9-543

Ying, R., Bourgeois, D., You, J., Zitnik, M., and Leskovec, J. (2019). Gnnexplainer: Generating Explanations for Graph Neural Networks. *Adv. Neural Inf. Process Syst.* 32, 9240–9251.

Young, T., Hazarika, D., Poria, S., and Cambria, E. (2018). Recent Trends in Deep Learning Based Natural Language Processing [Review Article]. *IEEE Comput. Intell. Mag.* 13, 55–75. doi:10.1109/mci.2018.2840738

Yu, Y., Abadi, M., Barham, P., Brevdo, E., Burrows, M., Davis, A., et al. (2018). Dynamic Control Flow in Large-Scale Machine Learning. In Proceedings of the Thirteenth EuroSys Conference (ACM), 265–283. doi:10.1145/3190508.3190551

Yuan, H., Yu, H., Gui, S., and Ji, S. (2020). *Explainability in Graph Neural Networks: A Taxonomic Survey*. *arXiv preprint arXiv:2012.15445*.

Zhang, J., Shi, X., Xie, J., Ma, H., King, I., and Yeung, D.-Y. (2018). *Gaan: Gated Attention Networks for Learning on Large and Spatiotemporal Graphs*. *arXiv preprint arXiv:1803.07294*.

Zhao, Q., Xiao, F., Yang, M., Li, Y., and Wang, J. (2019). "AttentionDTA: Prediction of Drug-Target Binding Affinity Using Attention Model," in *2019 IEEE International Conference on Bioinformatics and Biomedicine* (IEEE), 64–69. doi:10.1109/bibm47256.2019.8983125

Zheng, L., Fan, J., and Mu, Y. (2019). OnionNet: a Multiple-Layer Intermolecular-Contact-Based Convolutional Neural Network for Protein-Ligand Binding Affinity Prediction. *ACS Omega* 4, 15956–15965. doi:10.1021/acsomega.9b01997

Zheng, Z., and Merz, K. M. (2013). Development of the Knowledge-Based and Empirical Combined Scoring Algorithm (KECSA) to Score Protein-Ligand Interactions. *J. Chem. Inf. Model.* 53, 1073–1083. doi:10.1021/ci300619x

Zhou, Y.-T., Chellappa, R., Vaid, A., and Jenkins, B. K. (1988). Image Restoration Using a Neural Network. *IEEE Trans. Acoust. Speech, Signal Process.* 36, 1141–1151. doi:10.1109/29.1641

Zhu, F., Zhang, X., Allen, J. E., Jones, D., and Lightstone, F. C. (2020). Binding Affinity Prediction by Pairwise Function Based on Neural Network. *J. Chem. Inf. Model.* 60, 2766–2772. doi:10.1021/acs.jcim.0c00026

Zilian, D., and Sotriffer, C. A. (2013). SFCscore(RF): a Random Forest-Based Scoring Function for Improved Affinity Prediction of Protein-Ligand Complexes. *J. Chem. Inf. Model.* 53, 1923–1933. doi:10.1021/ci400120b

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.