Received: 22 November 2019 | Revised: 4 January 2020 | Accepted: 7 January 2020

DOI: 10.1002/wcms.1465

ADVANCED REVIEW

WIREs COMPUTATIONAL MOLECULAR SCIENCE WILEY

Machine-learning scoring functions for structure-based drug lead optimization

Hongjian Li[1] | **Kam-Heung Sze**[1] | **Gang Lu**[1] | **Pedro J. Ballester**[2]

[1]CUHK-SDU Joint Laboratory on Reproductive Genetics, School of Biomedical Sciences, Chinese University of Hong Kong, Shatin, Hong Kong

[2]Cancer Research Center of Marseille (INSERM U1068, Institut Paoli-Calmettes, Aix-Marseille Université UM105, CNRS UMR7258), Marseille, France

Correspondence
Pedro J. Ballester, Cancer Research Center of Marseille (INSERM U1068, Institut Paoli-Calmettes, Aix-Marseille Université UM105, CNRS UMR7258), Marseille, France.
Email: pedro.ballester@inserm.fr

Funding information
Agence Nationale de la Recherche, Grant/Award Number: ANR-17-ERC2-0003-01

Abstract

Molecular docking can be used to predict how strongly small-molecule binders and their chemical derivatives bind to a macromolecular target using its available three-dimensional structures. Scoring functions (SFs) are employed to rank these molecules by their predicted binding affinity (potency). A classical SF assumes a predetermined theory-inspired functional form for the relationship between the features characterizing the structure of the protein–ligand complex and its predicted binding affinity (this relationship is almost always assumed to be linear). Recent years have seen the prosperity of machine-learning SFs, which are fast regression models built instead with contemporary supervised learning algorithms. In this review, we analyzed machine-learning SFs for drug lead optimization in the 2015–2019 period. The performance gap between classical and machine-learning SFs was large and has now broadened owing to methodological improvements and the availability of more training data. Against the expectations of many experts, SFs employing deep learning techniques were not always more predictive than those based on more established machine learning techniques and, when they were, the performance gain was small. More codes and webservers are available and ready to be applied to prospective structure-based drug lead optimization studies. These have exhibited excellent predictive accuracy in compelling retrospective tests, outperforming in some cases much more computationally demanding molecular simulation-based methods. A discussion of future work completes this review.

This article is categorized under:

Computer and Information Science > Chemoinformatics

KEYWORDS

binding affinity prediction, lead optimization, machine learning, molecular docking, scoring function, Structural bioinformatics

1 | INTRODUCTION

Molecular docking is an important method in the domain of computer-aided drug design. It can be utilized to generate three-dimensional (3D) conformations of small-organic molecules as bound to a macromolecular target (pose generation[1]) or to predict which of these conformations are closer to that of a cocrystallized molecule (native pose prediction[2]). The target is often a protein, although other macromolecules like DNA can also be targeted by small molecules.[3,4] When the modulation of the target function exerted by the bound molecule triggers a medicinal effect in the organism,[5] the target is called therapeutic.

Docking is often adapted to predicting the binding affinities of protein–ligand complexes from their X-ray crystal structures,[6] a problem known as binding affinity prediction (BAP). While the latter permits reducing confounding factors to a minimum (e.g., pose generation error does not have to be considered in crystal structures), it has the disadvantage of being suboptimal to discriminate nonbinding molecules (nonbinders). The latter is crucial for structure-based virtual screening (SBVS),[7] where the goal is to identify binders within compound libraries containing a far larger proportion of nonbinders. It is also important for structure-based lead optimization (SBLO),[8] where the objective is typically to identify the most potent molecules among those with similar chemical structure to a drug lead. In either case, a scoring function (SF) is required to estimate the binding strength of a putative protein–ligand complex, as a surrogate of ligand bioactivity (e.g., inhibition of the molecular function of the target by the ligand). The accuracies of SFs arguably remain the major limitation for the reliability of these docking applications.

SFs can be categorized into two classes: classical SFs and machine-learning SFs.[9] A classical SF assumes a predetermined theory-inspired functional form for the relationship between the features characterizing the protein–ligand complex and its predicted binding affinity. This relationship is almost always assumed to be linear. However, many complexes will not conform to this strong modeling assumption and hence less accurate predictions will be obtained in those cases. In contrast, approaches based on machine learning (ML) circumvent the limitation of imposing a fixed functional form for the SF, which is learnt instead from training data. ML-based SFs are thus capable of implicitly capturing intermolecular binding interactions that are hard to model explicitly.[10] A broader introduction to ML-based SFs can be found in a previous review[9] (see for instance Figures 1–3 there).

2 | WHY IS THIS REVIEW TIMELY?

Predicting the binding strength of small molecules to a macromolecular target is one of the most challenging open problems in computational molecular science. It is thus exciting to see the latest progress made in the past 4 years since our last review in this area.[9] ML-based SFs are now sometimes called artificial intelligence (AI)-based SFs, following the widespread switch of nomenclature from ML to AI. This rebranding was motivated by the breakthroughs of deep learning (DL),[11] a subfield of ML, on problems from other disciplines. DL techniques such as deep convolutional neural network (CNN) achieved outstanding accuracy at image recognition[12] and deep recurrent neural network (RNN) were also found to be very powerful in speech recognition.[13] This review will analyze the level of improvement introduced by DL in BAP in comparison with more established ML methods.

Progress also includes the application of recent ML techniques such as eXtreme gradient boosting (XGBoost),[14] more suitable featurization schemes and studies investigating how well ML-based SFs perform at SBLO. The number of targets for which a ML-based SF has demonstrated top performance is continuously growing. Last but not least, the work of those gathering and curating protein structure and bioactivity data (Table 1) cannot be praised enough. Larger datasets are beneficial for ML training and result in better predictions across targets, even without any further algorithmic or featurization improvements.

Excellent reviews analyzing the success of ML-based SFs have been recently presented.[15–20] For example, Hou and coworkers[19] reviewed ML-based SFs by the employed ML algorithm, with particular attention to DL approaches. Another example is the particularly comprehensive review on the application of AI to computer-assisted drug discovery presented by Schneider, Yang and coworkers.[15] Reviews have also been presented about the application of ML to quantitative structure–activity relationship (QSAR) modeling[21] as well as proteochemometric modeling.[22] While these are different research topics, they share important commonalities with structure-based ML-based SFs that warrant following their developments.

TABLE 1 Popular benchmarks for evaluating the performance of scoring functions (SFs) at binding affinity prediction (BAP) or structure-based lead optimization (SBLO)

Benchmark	Data sources	Tasks	Availability
CASF-2007[23] **CASF-2013**[24] **CASF-2016**[25]	PDB	BAP	http://www.pdbbind-cn.org/casf.asp
D3R GC 2015[26] **D3R GC 2**[27] **D3R GC 3**[28]	Various	BAP, SBLO	https://drugdesigndata.org/about/datasets
BindingDB[29]	Various	BAP, SBLO	https://www.bindingdb.org/bind/index.jsp
CSAR[30]	Various	BAP	http://www.csardock.org/
BindingMOAD[31]	PDB	BAP	http://bindingmoad.org/
MoleculeNet[32]	Various	BAP	https://github.com/deepchem/deepchem/

Our review is instead organized by BAP application, not by ML algorithm, to clarify the progress introduced by ML-based SFs on each application. We paid particular attention to SBLO, the most important application of SFs for BAP. By contrast, SFs for SBVS are out of the scope of this review. The latter research topic has also experienced strong growth and thus requires a dedicated review to analyze its differential circumstances (the use of ML classifiers instead ML regressors, the high proportions of nonbinders, the unreliability of established benchmarks, the much higher size, and chemical diversity of test sets). Some of the specific questions that we will address in this review are:

- To which extent are ML-based SFs better than classical SFs on multitarget test sets?
- To which extent are ML-based SFs better than classical SFs on single-target test sets?
- How have these SFs been applied to SBLO?
- Which codes are freely available for each of these applications?
- Predictive target-specific ML-based SFs have been achieved: for which targets and requiring what training set sizes?
- There were very high expectations about the application of DL algorithms: are the resulting SFs consistently better than SFs using other ML algorithms?

The rest of the article is organized as follows. Section "Generic machine-learning SFs for binding affinity prediction" reviews generic ML-based SFs to predict binding affinity for diverse protein–ligand complexes spanning multiple targets. Section "Family-specific machine-learning SFs for binding affinity prediction" explores the application of generic SFs as well as development of ML-based SFs tailored to a protein family of interest (or a particular target within a family), with particular attention to how these can be used for SBLO prospective applications. Last, section "Conclusions" summarizes the current state and future prospects of this research topic.

3 | GENERIC MACHINE-LEARNING SFS FOR BAP

This section overviews studies predicting the binding affinities of protein–ligand complexes from multiple targets using their X-ray crystal structures. A recent review has grouped the performance of these SFs by the employed test benchmark.[19] Given that there are subtle differences between studies regarding training set composition and SF implementation, we focus instead on direct comparisons of different SFs within the same study.

Table 1 compiles common benchmarks for assessing the performance of SFs at various tasks. Generic SFs are usually benchmarked on PDBbind datasets. The first benchmark using PDBbind data was initially unnamed,[23] it was thus named the PDBbind benchmark in its first application to assess ML-based SFs.[10] Later, PDBbind authors renamed it CASF-2007 while they presented CASF-2013.[24] To date, three versions of comparative assessment of scoring functions (CASF) have been constructed: CASF-2007,[23] CASF-2013,[24] and CASF-2016[25] (these use the core sets of PDBbind v2007, v2013, and v2016 as test sets, respectively). Evaluating the performance of a generic SF at BAP corresponds to the scoring power test of CASF. The most widely used performance metric is the Pearson correlation coefficient (R_p) between the predicted and the measured binding affinities, so we will primarily report this metric. If R_p is not reported, we will use Spearman correlation coefficient (R_s) instead.

As expected in 2015,[9] there have been applications of deep neural network (DNN) for the development of SFs. For instance, Sirimulla and coworkers developed a DNN-based SF, DLSCORE,[33] by using an ensemble of fully connected neural

networks (NNs). DLSCORE was trained and tested on a partition of the PDBbind v2016 refined set. It employed 348 descriptors generated by the BINANA[34] software, which identifies ligand and protein atoms within a distance of 2.5–4.0 Å between them, as well as electrostatic interactions, binding pocket flexibility, hydrogen bonds, salt bridges, rotatable bonds, π interactions, among others. A 10-fold cross validation was performed. The best 10 networks based on R_p^2 (the square of the Pearson's correlation coefficient) on the validation set were aggregated to construct DLSCORE, which yielded a R_p^2 of 0.82 on the test set. On the same set, NNScore 2.0[35] and Vina[36] obtained a R_p^2 of just 0.21 and 0.15, respectively.

A salient characteristic of DNN is that it builds a sophisticated neural network with a large number of hidden layers and neurons. However, like other ML approaches such as random forest (RF) and support vector machine (SVM), DNN still relies on feature engineering, where expert knowledge is required to describe or represent molecules with fixed-length feature vectors. The use of CNN, on the other hand, has made it possible to generate features directly from the crystal structure of a complex, thus permitting automatic extraction of features that are not readily encoded in simplified potentials, such as hydrophobic enclosure or surface area dependent terms, and features that have not been acknowledged as informative or relevant by existing SFs. CNN arranges its neurons spatially, and only connects locally to the output of the previous layer. Therefore, a CNN is particularly suited to exploit data instances whose components are arranged spatially (e.g., protein and ligand atoms in 3D space).

Wei and coworkers combined the element-specific persistent homology (ESPH) method and CNN to develop a multichannel topological NN, TopologyNet.[37] This was the basis of TNet-BP, a deep CNN-based SF for BAP. The 3D complex geometry was represented by topological invariants, which helped to reduce the dimensionality of 3D biomolecular data. The element-specific persistent barcodes were transformed to a one-dimensional (1D) image-like representation with multiple channels. A CNN comprising a few 1D convolution layers and some fully connected layers was used to obtain higher level features from topological images which were used for regression. TNet-BP achieved $R_p = 0.826$ on the PDBbind v2007 core set and $R_p = 0.810$ on the v2016 core set.

Jiménez et al. utilized 3D CNN to devise K_{DEEP}.[38] A 3D voxel representation of both the protein and the ligand was generated using a van der Waals radius for each atom type, which in turns gets assigned to one of the eight pharmacophoric-like property channels (hydrophobic, hydrogen bond donor or acceptor, aromatic, positive or negative ionizable, metallic, and total excluded volume). The descriptors were computed on a fixed 24 $Å^3$ subgrid centered at the ligand's centroid, thereby capturing a neighborhood of the binding site in practice. Evaluated on the PDBbind v2016 core set, K_{DEEP} obtained $R_p = 0.82$. Nevertheless, when evaluated on four CSAR datasets, it obtained an average R_p of just 0.59, lower than that of RF-Score v3[39] ($R_p = 0.7$) and X-Score[40] ($R_p = 0.64$). When evaluated on several congeneric series of target-bound molecules, K_{DEEP} still performed worse than RF-Score v3 on four of the nine blind targets (MCL1, p38, Tyk2, and TRK-Kinase), although K_{DEEP} yielded an average R_p of 0.38, higher than that of RF-Score v3 ($R_p = 0.28$).

Siedlecki and coworkers developed a CNN-based SF called Pafnucy,[41] in which the protein–ligand complex is represented by a 3D grid for the DNN model to utilize a 3D convolution to produce a feature map. The complex was cropped to a defined size of 20 Å cubic box centered at the ligand's centroid. The positions of heavy atoms were discretized using a 3D grid with 1 Å resolution. In this way, each point was represented by a four-dimensional (4D) tensor where the first three dimensions define the coordinates of the atom and the last dimension defines a vector of 19 features used to describe an atom, including atom type, hybridization, valence, hydrophobic and aromatic properties, and partial charges. To process the input, which is a molecular complex represented as a 4D tensor, the neural network consisted in three 3D convolutional layers followed by max pooling layers and three fully connected layers. Trained on 11,906 complexes from the PDBbind v2016 general set, Pafnucy achieved $R_p = 0.70$ on CASF-2013, $R_p = 0.78$ on CASF-2016, and $R_p = 0.57$ on the independent Astex Diverse Set. These results were better than those of X-Score,[24] which obtained $R_p = 0.61$ and $R_p = 0.52$ on the CASF-2013 scoring power benchmark and the Astex Diverse Set, respectively.

Pande and colleagues proposed PotentialNet,[42] a graph CNN (GCN)-based SF. The staged PotentialNet comprises three main steps (covalent-only propagation, dual noncovalent and covalent propagation, and ligand-based graph) to achieve feature learning. The input descriptors covered atom types, bonds, and spatial distances of atoms. Evaluated on CASF-2007, PotentialNet obtained $R_p = 0.822$, better than RF-Score v1[10] ($R_p = 0.783$). Note that less old versions of RF-Score obtain higher performance on the same test set ($R_p = 0.803$),[43] which provides a better appraisal of the improvement introduced by PotentialNet (+0.019 R_p). By contrast, the performances of classical SFs range from 0.22 to 0.64 on the same test set.[23] Incidentally, on the agglomerative sequence cross validation split, PotentialNet obtained an R_p of 0.700, worse than that of RF-Score v1 ($R_p = 0.732$). This is a reminder that the relative performance of SFs can substantially depends not only on the dataset, but also on how it is split.

Ashtawy and Mahapatra[44] conducted a comprehensive assessment of the scoring power of classical and ML-based SFs across both diverse and protein-family-specific benchmark test sets using a common diverse set of features. Physicochemical and geometrical features used by X-Score, AffiScore[45] and RF-Score v1 were extracted and combined. Six regression techniques were utilized, including multiple linear regression (MLR), multivariate adaptive regression splines (MARS), k-nearest neighbors (kNN), SVM, RF, and boosted regression trees (BRT). The best ML-based SF turned out to be RF::XR, which obtained $R_p = 0.806$ on the PDBbind v2007 core set, whereas $R_p = 0.644$ was obtained by the best classical SF X-Score.[23]

Expanding from their previous work, Ashtawy and Mahapatra developed two new SFs based on bagging (BgN-Score) and boosting (BsN-Score)[46] ensembles of NN models. These SFs used combinations of the terms from X-Score, AffiScore,[45] GOLD,[47] and RF-Score v1 as features. Evaluated on the PDBbind v2007 core set, BgN-Score led to $R_p = 0.804$ and BsN-Score produced $R_p = 0.816$. These ensemble SFs were also more accurate than SFs based on a single nondeep NN ($R_p = 0.675$).

Khamis and Gomaa proposed 12 ML-based SFs and evaluated them on the PDBbind v2013 core set.[48] They employed a range of ML algorithms, including RF, BRT, kNN, NN, or SVM. The features were initially 108 terms from RF-Score, BALL,[49] X-Score,[40] and SLIDE.[50] Principal component analysis (PCA) was performed to reduce the dimensionality of the input set of features to just 17 principal components. The resultant SFs (with @ML suffix) were assessed and compared to 20 classical SFs. In the scoring power test, RF@ML, BRT@ML, and kNN@ML obtained a R_p of 0.704, 0.694, and 0.672, respectively, versus 0.614 achieved by the best classical SF (X-Score[24]).

Pires and Ascher developed CSM-lig,[51] a web server for protein–ligand BAP that encompasses protein and ligand complementarity in terms of shape and chemistry via a class of graph-based structural signatures called cutoff scanning matrix (CSM) to describe the 3D environment of proteins and small ligands. Atoms were labeled with eight pharmacophore types based on their physicochemical characteristics. A cumulative distribution of distances between atoms per pharmacophore pair was generated. Complementary small molecule properties were also considered. With these two sets of information, Gaussian processes (GP) was employed to train CSM-Lig, which achieved $R_p = 0.751$ on the PDBbind v2007 core set, $R_p = 0.80$ on the v2013 core set, and $R_p = 0.71$ on the v2014 core set.

Wei and coworkers developed a topology-based binding prediction method, T-Bind,[52] by combining the ESPH method and gradient boosting decision tree (GBDT) regression. ESPH retains important chemical and biological information while dramatically reducing biomolecular complexity. The features generated from element-specific topological fingerprints using binned barcode representation were fed to GBDT. T-Bind achieved $R_p = 0.818$ on the PDBbind v2007 core set, $R_p = 0.767$ on the v2013 core set, and $R_p = 0.775$ on the v2015 core set. It was suggested that protein–ligand hydrophobic interactions are extended to 40 Å away from the binding site.

De Azevedo and colleagues developed SAnDReS,[53] a computational tool for statistical analysis of docking results and development of SFs. SAnDReS offers several ML regression methods, including least absolute shrinkage and selection operator (LASSO), Ridge, Elastic Net, their cross validation version, Ordinary Linear Regression, and stochastic gradient descent (SGD). With this in-house tool, Bitencourt-Ferreira and de Azevedo[54] compiled a dataset of 48 high-resolution crystallographic structures for which binding data were available. The energy terms of SFs from AutoDock Vina,[36] AutoDock4,[55] and MolDock[56] were employed to build a series of ML-based SFs using SAnDReS.[53] Results showed that a polynomial equation with coefficients determined by elastic net with cross validation obtained the best performance with $R_s = 0.886$ versus $R_s = 0.746$ obtained by Vina.

Using RF-Score v3 features, Li et al. generated XGB-Score,[57] the first SF employing XGBoost, which is an implementation of GBDT designed for increased speed and performance. They investigated how the accuracy of XGB-Score varies with training set size, and observed that like RF-Score v3, XGB-Score also improves with training set size while outperforming classical SFs. XGB-Score,[57] RF-Score v3,[39] X-Score, Vina, and Cyscore[58] produced an average R_p of 0.806, 0.800, 0.643, 0.596, and 0.657, respectively, on CASF-2007 (the maximum test R_p was 0.815 by XGB-Score).

Li et al. revised RF-Score v1 to RF-Score v3 by expanding the set of features to include energy terms from Vina.[36] By following a ML approach, the accuracy of Vina was strongly improved. The factors responsible for this improvement and their generality were analyzed. Importantly, with the help of a proposed time-stamped benchmark, this improvement was demonstrated to grow larger as more data becomes available for training RF models, as regression models implying additive functional forms did not improve with more training data. Additional studies have shown how other classical SFs improve by substituting their predetermined functional form with a ML method.[57,59–61] For instance, Afifi and Al-Sadek[59] attempted to improve the prediction performance of four SFs (X-Score, Vina, AutoDock4, and RF-Score v1) by both replacing the linear regression model with RF and combining SFs into hybrid ones. Evaluated on the PDBbind v2016 core set, RF-Score v1 achieved the best performance with $R_p = 0.808$, followed by X-Score HP, Vina,

and AutoDock, which obtained $R_p = 0.656$, $R_p = 0.646$, $R_p = 0.573$, respectively. After substituting RF, the latter three SFs led to increased performance with $R_p = 0.672$, $R_p = 0.711$, $R_p = 0.672$, respectively. AutoDock and Vina benefited considerably from such algorithmic substitution, whereas X-Score had only a slight improvement. Six hybrid RF-based SFs combining the features of two individual SFs were also evaluated, among which the hybrid SF combining AutoDock and RF-Score v1 performed the best with $R_p = 0.824$.

Wang and Zhang[62] observed that those ML-based SFs designed for BAP (scoring) did not perform so well on other tasks (docking, screening). They further showed that it is possible to build a ML-based SF that excels at all three tasks: $\Delta_{vina}RF_{20}$, which employs RF and 10 features related to pharmacophore-based solvent-accessible surface area plus the 10 terms from AutoDock Vina score. To better estimate binding affinities for structures having weak binding, the initial training set of 3,336 crystal structures with weak binding affinities was supplemented with 3,322 computationally generated structures (synthetic data). As a result, $\Delta_{vina}RF_{20}$ obtained $R_p = 0.686$ and $R_p = 0.732$ for the CASF-2013 and CASF-2007 benchmarks, respectively. Subsequently, these authors realized two limitations of $\Delta_{vina}RF_{20}$: (a) receptor-bound water molecules were not considered, despite the analysis by a previous study[63] that over 85% protein–ligand complex structures have at least one bridging water molecule in the ligand binding site; (b) the change of internal ligand conformational energy was not considered either, which, in addition to the change of intermolecular interactions and solvation, can influence the process for a ligand molecule to adopt a receptor-bound conformation. To address these limitations, an XGB-based SF termed $\Delta_{vina}XGB$[64] was developed by exploring new features characterizing explicit mediating water molecules and ligand conformation stability. The training set was judiciously enlarged from 6,658 complexes used by $\Delta_{vina}RF_{20}$ to 14,406 complexes used by $\Delta_{vina}XGB$. The feature set was also expanded from 20 features used by $\Delta_{vina}RF_{20}$ to 94 features used by $\Delta_{vina}XGB$, including 58 Vina features, 30 bSASA features, plus three features related to water effect, two features related to ligand stability, and one feature related to ions. Tested on CASF-2016, $\Delta_{vina}XGB$ obtained a higher R_p than $\Delta_{vina}RF_{20}$ and Vina (0.796 vs. 0.732 and 0.604, respectively).

Wei and coworkers proposed a feature functional theory-binding predictor (FFT-BP),[65] which uses six categories of microscopic features derived from physical models, including Poisson Boltzmann theory, nonpolar solvation models, and components in molecular mechanics poisson-boltzmann surface area and quantum models. Multiple additive regression tree (MART), also named GBDT, was used for ranking the nearest neighbors via microscopic features. FFT-BP obtained $R_p = 0.80$ on the PDBbind v2007 core set, and $R_p = 0.78$ on the v2015 core set. In another study, these authors postulated that ligand-binding-induced reduction of protein flexibility, or rigidity strengthening, plays a unique role in protein−ligand binding. They considered nothing but protein rigidity change upon ligand binding, quantified by the element-specific rigidity indices calculated from interatomic distances. These rigidity indices were used as features and combined with RF in model development to generate RI-Score,[66] which achieved $R_p = 0.803$ on the PDBbind v2007 core set, $R_p = 0.782$ on the v2013 core set, and $R_p = 0.815$ on the v2016 core set. They concluded that flexibility reduction or rigidity enhancement is a mechanism in protein−ligand binding, with contributing interactions from the nearest four layers of residues.

As Wang and Zhang,[62] Ashtawy and Mahapatra[67] pointed out the limited predictive accuracies of generic binding affinity-based SFs when applied to docking and screening. Hence they employed a task-specific strategy and developed specific ML-based SFs for each of the three tasks. For the task of BAP, they developed BT-Score, an ensemble of 4,000 BRT looking at 2714 features and trained with 3,000 complexes. This large set of descriptors comes from several popular SFs such as AffiScore,[45] Vina, Cyscore,[58] DSX,[68] RF-Score v1, X-Score, and others. BT-Score reproduced binding affinity of out-of-sample test complexes with $R_p = 0.827$ on the PDBbind v2014 core set, while RF-Score v1 obtained $R_p = 0.725$, and X-Score obtained $R_p = 0.627$. Moreover, a novel multitask DNN (MT-Net) was proposed to simultaneously tackle the three tasks. Its performance was shown to be superior to classical SFs and on a par with or better than models based on single task neural networks. More recently, Nguyen and Wei[69] have presented a novel algebraic graph learning score, AGL-Score, which was design to excel at multiple tasks. AGL-Score employs multiscale weighted colored subgraphs to describe crucial molecular and biomolecular interactions in terms of graph invariants derived from graph Laplacian, its pseudoinverse, and adjacency matrices. The eigenvalues and eigenvectors computed from these matrices were used as features to characterize the biological and physical interactions of molecules. Coupled with GBDT, AGL-Score achieved some of the best performances on scoring power benchmarks ($R_p = 0.830$ on CASF-2007, $R_p = 0.792$ on CASF-2013, and $R_p = 0.833$ on CASF-2016).

Wei and coworkers also investigated the impact of featurization (i.e., translating the 3D structures of biomolecules to features) on SF performance. Despite the powerful capability of DL for automatic extraction of features from original inputs such as images, DL-based SFs taking biomolecules as inputs are not as competitive as some ML-based SFs with carefully designed features, due to the intrinsic complexity of biomolecules. To this end, they introduced a number of

algebraic topology approaches to characterize biomolecular complexes. With topological fingerprints generated from multicomponent persistent homology, multilevel persistent homology, and electrostatic persistence, they employed GBDT to build TopBP-ML(complex) and CNN to generate TopBP-DL(complex), and their consensus TopBP(complex).[70] Evaluated on the PDBbind v2007, v2013, v2015, and v2016 core sets, TopBP(Complex) exhibited the best performance with R_p = 0.827, 0.808, 0.812, and 0.861, respectively. TopBP-DL(complex) did not outperform TopBP-ML(complex) on v2007 and v2013 core sets, showing again that DL-based SFs do not necessarily outperform SFs based on other ML algorithms.

Nguyen and Wei introduced differential geometry-based geometric learning (DG-GL) as a representation of biomolecular structures and their interactions.[71] The idea is to encode chemical, biological, and physical information contained in high-dimensional data into differentiable low-dimensional manifolds, from which latent mathematical representations of the original dataset are then constructed using differential geometry tools. Their DG-GL strategy required only atomic coordinates and element types as its essential input data, without molecular force fields in general. Element interactive curvatures (EICs) generated from the manifolds were used as differential geometry features to GBDT. The resulting SF, EIC-Score,[71] achieved $R_p = 0.817$ on the PDBbind v2007 core set, $R_p = 0.774$ on the v2013 core set, and $R_p = 0.825$ on the v2016 core set.

Boyles et al.[72] demonstrated that the inclusion of diverse ligand-based features in ML-based SFs improves their scoring power across multiple targets. A RF-based SF combining RF-Score v3 features with 183 RDKit molecular descriptors achieved R_p = 0.836, 0.780, and 0.821 on the PDBbind 2007, 2013, and 2016 core sets, respectively, compared to R_p = 0.790, 0.746, and 0.814 when exclusively using the features of RF-Score v3. Excluding proteins or ligands that are similar to those in the test sets from the training set had a deleterious effect on scoring power, but did not remove the predictive power of ligand-based features. This result is contrary to that by Schneider et al.,[73] who found that ligand-based features have lower predictive power than structure-based features, and their combination led to a moderate intermediate accuracy when an RF-based SF was applied on the estrogen receptor alpha (ERα) target.

In the above studies, the developed SFs were exclusively evaluated on crystal structures of protein–ligand complexes. This approach has the advantage of circumventing the introduction of confounding factors such as pose generation error. Thus, SFs developed in this way should perform well at predicting binding affinities from crystal structures. On the other hand, when a crystal structure with the ligand cocrystallized is unavailable, which is a common scenario, docking has to be performed to generate putative binding poses. There are only a few studies analyzing the prediction of binding affinities from docked poses in the presence of pose generation error. One of these studies is by Li et al.,[2] who systematically analyzed the influence of this error, measured as the difference between the geometry of the docked pose and that of the same molecule cocrystallized with the considered protein, on BAP across diverse protein–ligand complexes. Against commonly held views, they found that the impact of pose generation error on the scoring power of SFs is generally small. This observation applies to not only ML-based SFs such as RF::VinaElem, but also classical SFs such as MLR::Vina. It was also shown that a substantial part of this error can be corrected if the SF is calibrated on docked poses, rather than on crystal poses as usual. In this way, the relationship between the poses generated by docking software and their binding affinities is directly learned. After this error-correcting procedure, the SF performance becomes pretty close to that of predicting the binding affinity without pose generation error (i.e., on crystal structures). Furthermore, several strategies were assessed, among which those using a single docked pose per ligand obtained better results than those using multiple docked poses per ligand. The SF implementing this error-correcting procedure was termed RF-Score v4.[2]

Recent controversy over the impact of different ways to partition data into training and test sets has arisen. Li and Yang[74] measured the training-test set similarity by their constituting protein structures and sequences. Through controlling the similarity cutoff, the original full training set was split to form a series of nested sets, with only training complexes whose proteins are highly dissimilar to those in the test set initially, and subsequently expanded gradually to incorporate similar proteins as well. These nested training sets were also be sorted in the opposite direction, that is, from small sets of highly similar proteins to large sets that also include highly dissimilar proteins, to better understand to what extent such most relevant data could contribute to the performance of SFs. They showed that ML-based SFs failed to outperform classical SFs after removal of training complexes with proteins highly similar to the test proteins, resulting in the conclusion that the remarkable scoring power of ML-based SFs is exclusively credited to the presence of training data most relevant to the test set.[74] Nonetheless, a subsequent but expanded reanalysis by Li et al.[75] found issues with this analysis. ML-based SFs were shown to outperform classical SFs even when trained with a moderate percent of dissimilar proteins, suggesting that ML-based SFs owe a considerable part of their remarkable performance to training on complexes whose proteins are dissimilar to those in the test set.[75] By generating additional nested training

sets with even fewer training complexes, these authors highlighted that classical SFs are unable to exploit large sizes of structural and interaction data, as including a larger proportion of similar complexes to the training set did not make them more accurate.[57] On the contrary, ML-based SFs, regardless of employing either RF or XGB, managed to keep learning from more data and improving performance.

To sum up, Table 2 compiles a list of the above-reviewed ML-based SFs for BAP, along with their ML algorithm, features and benchmarks in use, and their availability. In addition, Figure 1 shows how the best performance on CASF-2007 has been improving with new SFs since this benchmark was presented about 10 years ago.[23] We focus on CASF-2007, instead of more recent benchmarks like CASF-2013,[76] because otherwise some early competitive SFs such as CScore[77] would be missed.

Furthermore, there are some papers that facilitate to start research in this area. Wang and coworkers have presented a protocol to carry out the CASF benchmark.[76] This protocol specifically refers to the CASF-2013 benchmark, enabling evaluation of not only the so-called scoring power (to which this section is dedicated), but also other tasks ("ranking power," "docking power," and "screening power"). Evaluation results of classical SFs implemented in several commercial software packages (including Schrödinger, MOE, Discovery Studio, SYBYL, and GOLD) are provided as reference. In this protocol, the authors provide detailed descriptions of the data files included in the CASF-2013 package and step-by-step instructions on how to conduct the performance tests with the provided ready-to-use computer scripts. A complementary protocol was presented by Wójcikowski et al.,[6] a step-by-step explication of how to build and evaluate the original version (v1) of RF-Score using the CASF-2007 benchmark. This paper also points out how to use different data, features, and regression models using either R or Python programming languages.

Last, we have carried out several experiments to show how the future availability of more data is expected to affect the performance of ML-based and classical SFs. Figure 2 shows how ML-based SFs increase performance given more data for training, whereas classical SFs do not. This was carried out by multiple time-wise splits of the PDBbind data, each split sharing the same test set (318 complexes from the v2018 refined set not already included in the v2017 refined set, that is, refined2018\refined2017). Interestingly, this test set is completely new, that is, none of these 318 complexes is included in any of the v2007, v2013, v2014, v2015, v2016, and v2017 refined sets, hence it represents future unseen data. Therefore, the five training sets of increasing data size are exactly the refined sets themselves: refined2007 ($N = 1{,}300$), refined2013 ($N = 2{,}959$), refined2014 ($N = 3{,}446$), refined2015 ($N = 3{,}706$), and refined2017 ($N = 4{,}154$). Five comparing models are Vina, MLR::Vina, RF::Vina, RF::Elem (analogous to RF-Score v1[10]), and RF::VinaElem (analogous to RF-Score v3[39]). Results show that time-wise splits are somewhat harder than CASF splits for all the tested SFs. Even without improving features or ML algorithms, exploiting more training data results in more accurate BAP of test set complexes.

Beyond protein–ligand complexes, a ML approach to predict protein–protein interactions (PPIs) has also been taken. Here the challenge is to predict the binding strength of two macromolecules. For instance, Li et al.[78] designed an SVM for regression ensemble for protein–protein BAP, which was reported to be substantially more predictive than popular knowledge-based SFs for PPIs. While there are notable differences between this problem and that of protein–ligand BAP (e.g., possibility of using residue–residue features, more flexibility in the ligand protein or much less data available), many concepts and methodologies are transferable. This is explored in a recent review.[79] Furthermore, Han et al.[80] created a new benchmark, named CASF–PPI, specifically for assessing the SFs applicable to protein–protein docking tasks.[80] A high-quality dataset of 273 protein–protein complexes was compiled and employed in both tests. This is based on a larger, nonredundant set of protein–protein complexes with carefully examined 3D structures and experimental binding data. Four SFs were evaluated to demonstrate how the CASF–PPI benchmark may be applied.

4 | FAMILY-SPECIFIC MACHINE-LEARNING SFS FOR BAP

Tailoring the ML-based SF to the characteristics of a target or a family of targets represents a promising route to improving its performance. The SF can be tailored by training on only those complexes with the considered target. Another way to tailor SFs to a target is identifying the most predictive features for that target, instead of using a generic set of features in all targets as in the slides of the previous section. For example, the predictive performance of family-specific SFs is in principle expected to improve when the presence of a given metal ion coordinating ligand binding is part of the employed features.

Utilizing their in-house program SAnDReS,[53] de Azevedo and colleagues built ML-based SFs to investigate four popular protein targets: Cyclin-dependent kinase 2 (CDK2),[81] HIV protease (HIV PR),[82] CDK,[83] and 3-dehydroquinate dehydratase (DHQD).[84] In the study of CDK2,[81] these authors compared ML-based SFs to classical SFs (PLANTS and

TABLE 2 Machine-learning scoring functions for binding affinity prediction (BAP), sorted by their publication date with the most recent at the end

SF	Machine learning (ML) method	Features or descriptors	Benchmark	Availability
RF::XR[44] **BRT::XAR**[44] **SVM::XAR**[44]	kNN, SVM, RF, BRT	Combinations of the terms from X-Score, AffiScore and RF-Score	CASF-2007	N/A
RF-Score v3[39]	RF	Terms from RF-Score v1 and Vina	CASF-2007 and PDBbind v2013 blind benchmark	http://ballester.marseille.inserm.fr/rf-score-3.tgz
BgN-Score[46] **BsN-Score**[46]	NN	Combinations of the terms from X-Score, AffiScore, GOLD and RF-Score v1	CASF-2007	N/A
RF@ML[48] **BRT@ML**[48] **kNN@ML**[48]	RF, BRT, kNN, NN, SVM, etc.	Terms from RF-Score, BALL, X-Score and SLIDE	CASF-2013	N/A
CSM-lig[51]	GP	Cutoff scanning matrix (CSM)	PDBbind v2007, v2013, v2014	http://biosig.unimelb.edu.au/csm_lig
RF-Score-v4[2]	RF	Terms from RF-Score v1 and Vina	CASF-2007 and PDBbind v2013 blind benchmark	http://ballester.marseille.inserm.fr/rf-score-4.tgz
$\Delta_{vina}RF_{20}$[62]	RF	Ten terms from Vina and 10 terms related to buried solvent-accessible surface area (bSASA)	CASF-2013 and CASF-2007	https://www.nyu.edu/projects/yzhang/DeltaVina
FFT-BP[65]	GBDT	Microscopic features	PDBbind v2007, v2015	N/A
T-Bind[52]	GBDT	Topological fingerprints	PDBbind v2007, v2013, v2015	N/A
RI-Score[66]	RF	Element-specific rigidity index	CASF-2007, CASF-2013, CASF-2016	http://weilab.math.msu.edu/RI-Score
TNet-BP[37]	CNN	Multichannel topological invariants	CASF-2007 and PDBbind v2016	https://weilab.math.msu.edu/TDL/TDL-BP
BT-Score[67] **MT-Net**[67]	GBDT, multitask DNN	2,714 descriptors from http://www.descriptordb.com	PDBbind v2014	N/A
K_{DEEP}[38]	CNN	Voxelized 24 Å representation of the binding site considering eight pharmacophoric-like properties	PDBbind v2016, CSAR datasets, congeneric series sets	https://playmolecule.org/Kdeep/
TopBP[70]	GBDT, CNN	Topological fingerprints	PDBbind v2007, v2013, v2015, v2016	N/A
Afifi and Al-Sadek[59]	RF	Terms from X-Score, Vina, AutoDock, and RF-Score v1	PDBbind v2016	N/A
DLSCORE[33]	DNN	348 BINANA descriptors	PDBbind v2016	https://github.com/sirimullalab/dlscore
Pafnucy[41]	CNN	Atomic coordinates and 19 features associated with atom type, hybridization, bonds, pharmacophoric-like properties or partial charges	CASF-2013, CASF-2016, Astex diverse set	https://gitlab.com/cheminfIBB/pafnucy
PotentialNet[42]	GCN	Basic information about atoms, bonds, and distances	CASF-2007	N/A

(Continues)

TABLE 2 (Continued)

SF	Machine learning (ML) method	Features or descriptors	Benchmark	Availability
EIC-Score[71]	GBDT	Element interactive curvatures (EICs)	CASF-2007, CASF-2013, CASF-2016	https://weilab.math.msu.edu/DG-GL
XGB-Score[57]	XGBoost	Terms from RF-Score and Vina	CASF-2007	https://github.com/HongjianLi/MLSF
AGL-Score[69]	GBDT	Graph invariants	CASF-2007, CASF-2013, CASF-2016	https://weilab.math.msu.edu/AGL-Score
Δ_{vina}XGB[64]	XGBoost	58 terms from Vina, 30 terms related to bSASA, three features related to water effect, two features related to ligand stability, one feature related to ions	CASF-2016	http://www.nyu.edu/projects/yzhang/DeltaVina
Boyles et al.[72]	RF-Score v3	RF-Score v3 features with 183 RDKit molecular descriptors	PDBbind 2007, 2013 and 2016 core sets	http://opig.stats.ox.ac.uk/resources

Note: In case of unnamed scoring functions (SFs), these are identified by the names of their authors instead. Abbreviations: BRT, boosted regression trees; CNN, convolutional neural network; DNN, deep neural network; GBDT, gradient boosting decision tree; GP, Gaussian processes; kNN, k-nearest neighbors; NN, neural network; RF, random forest; SVM, support vector machine; XGBoost, eXtreme gradient boosting.

Publication date	Scoring function	Regression model	Test set R_p
2019-08-26	Boyles et al.[72]	RF	0.836
2019-06-18	AGL-Score[69]	GBDT	0.830
2018-01-08	TopBP[70]	GBDT & CNN	0.827
2017-07-27	TNet-BP[37]	CNN	0.826
2017-06-29	T-Bind[52]	GBDT	0.818
2015-02-23	BsN-Score[46]	NN	0.816
2014-08-25	RF::XR[44]	RF	0.806
2014-02-16	RF-Score v2[43]	RF	0.803
2011-12-01	CScore[77]	NN	0.801
2010-03-17	RF-Score v1[10]	RF	0.776
<= 2009-04-09	X-Score[23]	Linear	0.644
<= 2009-04-09	DrugScore^{CSD}[23]	Linear	0.569
<= 2009-04-09	SYBYL::ChemScore[23]	Linear	0.555
<= 2009-04-09	DS::PLP1[23]	Linear	0.545
<= 2009-04-09	GOLD::ASP[23]	Linear	0.534
<= 2009-04-09	SYBYL::G-Score[23]	Linear	0.492
<= 2009-04-09	DS::LUDI3[23]	Linear	0.487
<= 2009-04-09	DS::LigScore2[23]	Linear	0.464
<= 2009-04-09	GlideScore-XP[23]	Linear	0.457
<= 2009-04-09	DS::PMF[23]	Linear	0.445
<= 2009-04-09	GOLD::ChemScore[23]	Linear	0.441
<= 2009-04-09	by NHA[23]	Linear	0.431
<= 2009-04-09	SYBYL::D-Score[23]	Linear	0.392
<= 2009-04-09	DS::Jain[23]	Linear	0.316
<= 2009-04-09	GOLD::GoldScore[23]	Linear	0.295
<= 2009-04-09	SYBYL::PMF-Score[23]	Linear	0.268
<= 2009-04-09	SYBYL::F-Score[23]	Linear	0.216

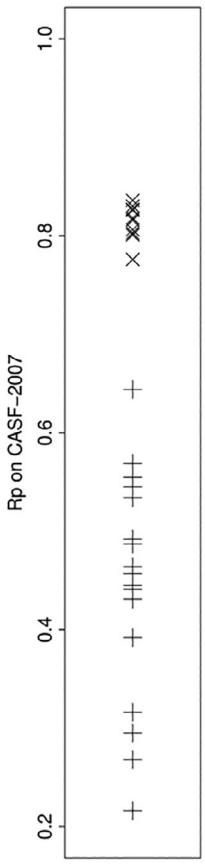

FIGURE 1 Performances of the classical scoring functions (SFs) tested on CASF-2007 ("+" signs) along with every SF that has surpassed the previous best R_p performance ("x" signs, all machine learning [ML]-based SFs). This shows that: (a) no competitive classical SF has been introduced since the advent of ML-based SFs, (b) Deep learning-based SFs are not necessarily the most predictive, and (c) as we can now predict the affinities of protein–ligand complexes with high accuracy on average across diverse targets, future efforts should focus instead on which SF is most predictive for each target (accuracy on some targets is still poor)

MolDock), with 173 generic crystallographic structures as the training set and 11 CDK2 crystallographic structures as the test set. Application of ML methods to build new SFs using the energy terms in MolDock and PLANTS was shown to generate a polynomial equation with improved prediction power (R_s of 0.845) when compared to AutoDock4,[55] Vina,[36] PLANTS, and MolDock whose R_s ranged from −0.773 to 0.682.

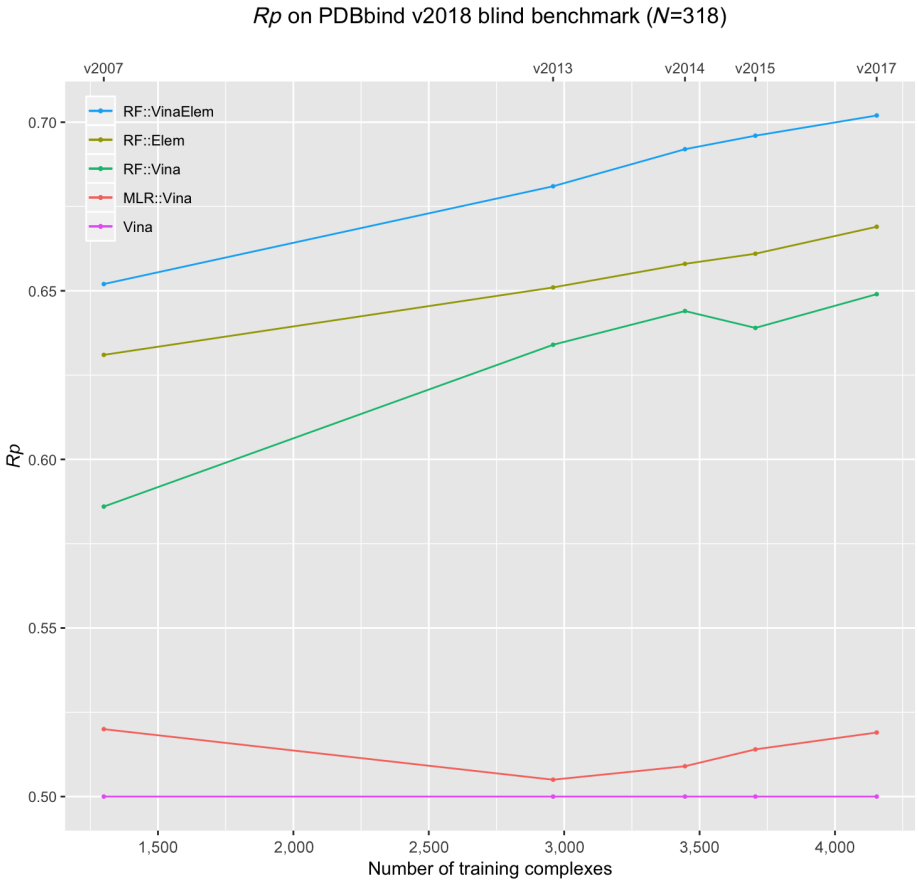

FIGURE 2 PDBbind v2018 blind test showing how test set performance (in terms of R_p) grows with more training data (from the PDBbind refined sets v2007 to v2017) when using random forest, but stagnates with multiple linear regression. AutoDock Vina acts as a baseline without retraining

In the study of HIV PR, Pintro and de Azevedo[82] built target-specific SFs to predict inhibition constants (K_i) for ligands against the HIV-1 protease. Seventy-one crystal structures of HIV protease were collected, of which 51 were for training and the other 20 were for testing. Energy terms from MolDock and PLANTS were used to build ML-based SFs, the best of which produced $R_s = 0.368$, outperforming PLANTS ($R_s = 0.\ 010$) and MolScore ($R_s = 0.086$).

In the study of CDK, de Azevedo and coworkers[83] focused on the development of CDK-targeted ML-based SFs. One hundred and seventy CDK structures were collected, of which 70% were for training. The overall performance of a ML-based SF ($R_s = 0.346$) were higher than MolDock ($R_s = -0.291$), AutoDock4 ($R_s = 0.213$), and Vina[36] ($R_s = 0.207$). Also with CDK as an use case, the authors implemented a program named Taba,[85] an acronym for tool to analyze the binding affinity. Taba represents protein–ligand interactions as a mass-spring system and considers the average intermolecular distances calculated from an ensemble of crystallographic structures of protein–ligand complexes. Seven ML techniques (ordinary linear regression, LASSO, LASSO with cross validation, Ridge, Ridge with cross validation, Elastic Net, and Elastic Net with cross validation) were utilized and a set of 31 structures of human CDKs was split in 70/30 for training/test partition. The best model obtained $R_p = 0.794$, better than those obtained by the classical SFs from MVD ($R_p = 0.001$ by PLANTS and $R_p = 0.010$ by MolDock), AutoDock4 ($R_p = 0.204$), and Vina ($R_p = 0.117$).

In the study of DHQD, de Ávila and de Azevedo[84] described the development of ML-based SFs to predict log(K_i) for the enzyme DHQD, which is the third step of the shikimate pathway and is responsible for the synthesis of chorismate. A dataset of 22 structures, each showing a different ligand with specific K_i value, was used in the analysis of ensemble docking and to develop DHQD-targeted SFs with different combinations of energy terms from MolDock and AutoDock4. Two resulting SFs have shown superior predictive performance ($R_s = 0.900$ and 0.943) when compared with classical SFs such as AutoDock4 ($R_s = 0.714$), MolDock ($R_s = -0.943$), and PLANTS ($R_s = 0.314$). Intermolecular electrostatic interactions between DHQD and competitive inhibitors were found to be of pivotal importance for the binding affinity against this enzyme. Previous studies showed that it is possible to discover novel DHQD inhibitors using ML-based SFs.[86]

Schneider et al.[73] focused on a well-known therapeutic target, the estrogen receptor ERα, a steroid binding receptor playing a key role in a range of diseases. When applied to this target using RF, 11 ligand-based features were found to

have lower predictive power ($R_p = 0.69$), compared to 19 structure-based features ($R_p = 0.78$). By contrast, Boyles et al.[72] showed that the scoring power of ML-based SFs can be consistently improved by the inclusion of ligand-based features on average across targets. This suggests that ERα is an exception to this general trend. Combining ligand-based and structure-based features maintained high accuracy ($R_p = 0.73$) in between the two reduced-variable models on the internal test set, still strongly outperforming PLANTS,[87] MedusaScore,[88] DSX,[68] and X-Score[40] ($R_p = 0.038$, 0.111, 0.118, and 0.076, respectively). A previous prospective study on this target[89] found that NNScore was able to discover a range of ERα ligands (some with nM potency).

Like Pintro and de Azevedo,[82] Schiffer and coworkers[90] also analyzed the application of ML-based SFs to rank actives against HIV proteases. Structure-based protein−ligand interaction fingerprints including vdW potential, hydrogen bonds, halogen bonds, salt bridges, π-interactions, and π-cation interactions were used as features and generated for 282 crystal structures of HIV protease in complex with competitive active site inhibitors. A GBDT-based SF yielded $R_p = 0.77$, outperforming three baseline ML-based SFs using Elastic Net, SVM, and RF. The recently proposed TreeSHAP method[91] was utilized to build an explanatory model with an additive feature attribution for the GBDT-based SF, attempting to quantify feature importance on the prediction performance and thus identify the most informative features.

In addition to basic research studies, the application of target-specific ML-based SFs to translational problems has been investigated. Nogueira and Koch[92] showed ML-based SFs can be used to predict the targets of a ligand. A NN model and an SVM model were built for each of the 20 considered protein targets. In this way, the targets of a ligand can be returned as the targets where the ligand is predicted to bind. This target-centric approach has also been investigated at large scale (5,454 crystal structures comprising 869 targets that can be predicted),[93] albeit using classical SFs. Note, however, that even this number of targets is low with respect to those offered by ligand-centric approaches.[94,95] Regarding how these SFs were built,[92] the protein−ligand interaction fingerprint PADIF (protein atom score contributions derived interaction fingerprint) calculated based on docking poses of active and inactive compounds was employed as a source of features. The authors reported that target-specific SFs were more predictive than off-the-shelf generic SFs for this application, as it has been seen as well in closely related applications.[96]

Another application of ML-based SFs for BAP is SBLO.[8] More concretely, ranking the chemical derivatives of a lead molecule by their predicted affinities against a target as a way to identify the most potent derivatives. Jiménez-Luna et al.,[97] the authors of the CNN-based SF K_{DEEP}[38] for generic BAP, recently presented DeltaDelta NN, a CNN specifically tailored for ranking congeneric series of molecules against the structure of a target. This CNN is innovative in that, instead of predicting the binding affinity of a single molecule as usual, predicts the relative affinities of two congeneric molecules. Therefore, each data instance consists of a pair of congeneric molecules docked, or bound, to the target of interest. The authors argue that a CNN predicting relative affinity is more accurate than subtracting the absolute affinity predictions of the pair of congeneric molecules because calculating relative affinities from an absolute predictor inevitably leads to the concatenation of errors from two separate predictions. Comprehensive and compelling numerical experiments are presented using public and private datasets from a range of targets, including BRD4, BACE, CDK2, JNK1, MCL1, P38, PTP1B, Thrombin, TYK2, PDE2, PDE3, PDE10, and ROS1. The authors declared that these were carried out blindly among them, while employing sequences of random and time-based data splits into training and test sets, the latter realistically mimicking the process of SBLO. The results show the power of using large target-specific datasets. For instance, in the Pfizer datasets with up to 362 ligands, an average test R_s of 0.64 across six targets was achieved. This correlation is much higher than those obtained by the baselines, including molecular weight, clogP, and MM-GBSA (molecular mechanics/generalized born surface area) based on molecular dynamics simulations and the generic CNN-based SF K_{DEEP} with average test R_s of 0.2, 0.18, 0.4, and 0.18, respectively. On several targets, these authors demonstrated that DeltaDelta NN was even more predictive than a far slower physics-based free energy perturbation (FEP) simulation method.[98]

Community blind tests have also been carried out to assess the SBLO ability of a range of in silico methods.[26–28] The D3R Grand Challenge 2[27] presented a blind evaluation of methods to predict the affinities of molecules against the nuclear receptor Farnesoid X receptor (FXR). Wei and coworkers obtained the top place in absolute free energy prediction for free energy set 1 in stage 2 with their mathematical DL models.[99] RF-Score v3[39] to rescore docked poses generated by idock[100] was among the top performing methods in some subchallenges. There were other top methods using ML, for example, those labeled "Trained," but their design principles were not stated (e.g., it was not clear if these were ML-based SFs or ligand-based 3D QSAR models). This study led to additional findings of particular interest. First, pose accuracy did not correlate well with ranking accuracy, indicating that much of the error in ranking ligands by predicted

affinity is due to SF limitations and not to a failure to identify the native poses. In a previous study, we also found that pose accuracy does not correlate with ranking accuracy,[2] which is the reason why we do not focus on pose generation error in this review. Second, explicit-solvent free energy simulation methods did not provide greater accuracy than much faster, less detailed methods (this has also been concluded in other studies[26,97]).

The D3R Grand Challenge 3[28] was a larger blind evaluation and more specific regarding whether methods employ ML or not. In comparison to the previous challenge,[26] only one participant employed free energy simulation methods, whereas an increase in the use of ML methods allied with structure-based modeling was observed. Also, this challenge yields the highest potency ranking accuracy, with values of Kendall's τ exceeding the highest prior GC2 value of 0.46, for ABL1 (0.52 ± 0.3), JAK2 SC2 (0.55 ± 0.08), JAK2 SC3 (0.71 ± 0.16), and TIE2 (0.57 ± 0.24). These authors also show that the top in silico methods outperformed high-throughput screening (HTS) single-concentration activity measurements for two of the six compared targets (ABL1, TIE2)[28] regarding their ability to correlate with low-throughput multiconcentration binding constant measurements. In contrast with these accurate affinity rankings, a prior large-scale evaluation concluded in 2006[101]: "For prediction of compound affinity, none of the docking programs or SFs made a useful prediction of ligand binding affinity." Comparing the results of both exercises does indeed show that SF development has substantially improved in recent years.

The D3R Grand Challenge 3 study[28] also stated that it is not clear whether structure-based ML methods perform better overall than those not using ML across the six different protein targets (Cathepsin S and the kinases VEGFR2, JAK2, p38-α, TIE2, and ABL1). However, eight subchallenges with top 3 methods each give 24 top submissions, of which 19 did use ML. This can be seen in Table 5 of that study,[28] where the three methods with the highest rank correlation to measured affinities per subchallenge are displayed (those with submission IDs in bold font indicate a method that used ML). To be more accurate, one can calculate per subchallenge how enriched are the top 3 submissions with ML methods. The numbers of ML submissions per subchallenge are 24, 51, 8, 17, 15, 18, 15, and 19 for CatS stage 1, CatS stage2, ABL1, JAK2 SC2, JAK2 C3, p38-α, TIE2, and VEGFR2, respectively. Taking into account the total numbers of submissions (see their Table 2[28]) and using the enrichment factor (EF) metric with ranked submissions instead of ranked molecules, EF_{top3} were 1.5 (2.3), 1.6 (1.6), 0.92 (1.4), 1.2 (1.8), 1.2 (1.2), 0.54 (1.6), 1.2 (1.2), and 1.8 (1.8) for CatS stage 1, CatS stage2, ABL1, JAK2 SC2, JAK2 C3, p38-α, TIE2, and VEGFR2, respectively (between brackets the highest possible EF_{top3} is provided for comparison). These results mean, for example, that ML methods were 1.8 times more likely to be among the top submissions for VEGFR2 (this is the maximum enrichment because all top 3 submissions employ ML). Overall, ML methods did only work worse on two of the eight subchallenges (those investigating targets ABL1 and p38-α). The ML-based SFs from Wei and colleagues achieved more ranked-first submissions than any other participant in this challenge.[99]

Community blind tests like D3R are important. They have the advantage that test set affinities are only made available to the participants after their models have been selected (e.g., FEP simulation settings, ligand preparation protocols, weights in classical SF terms, hyperparameters in ML methods), thus mimicking a realistic scenario. Furthermore, community blind tests permit a broader comparison across very diverse classes of methods, which is helpful to clarify the application niche of a given class. However, they have drawbacks too. Unconscious bias in selecting targets might be detrimental for a particular class of methods (e.g., targets with few known ligands tends to be harder for ML-based SFs), precluding the possibility of generalizing conclusions to unselected targets. Also, challenges are presented to have the same practical importance, but docking pose generation is not as important as BAP. Last, the results might be analyzed by experts from one class of methods only, which might result in unconsciously selecting types of analysis that favor them.

Another important consideration is that a community blind test is not the only valid way to evaluate SFs. A SF can be fully evaluated by the community if its code and training data is made freely available along with proper documentation (this is only the case for a few methods participating in D3R challenges). In this way, anyone can easily verify what is claimed about the SF, evaluate it on any target with any type of analysis, employ it to construct criticisms advancing our understanding of this problem and/or help others to build upon that work to generate more accurate SFs. Another way to evaluate SFs is prospectively (there are already impressive prospective applications of target-specific SVM-based SFs to SBLO[102]). Such applications should increase, as the number of open-source ML-based SFs, or at least freely available as executables or webservers, grows. Table 3 compiles studies in this section, including the DeltaDeltaNN webserver[97] that can build target-specific DNNs with user-supplied training and test sets and hence can be used for prospective SBLO applications. While expected to be generally less predictive due to their generic nature, the codes of ML-based SFs in Table 2 can also be used for this objective without any further training.

TABLE 3 Family-specific machine-learning scoring functions for binding affinity prediction (BAP). Targets of the test set used as benchmark is stated

Study	ML method	Features or descriptors	Test sets	Availability
De Azevedo and colleagues[81–84]	LASSO, Ridge, Elastic Net, Ordinary Linear Regression	Energy terms from AutoDock4, Vina, PLANTS and MolDock	CDK2, HIV PR, CDK and DHQD	N/A
Da Silva et al.[85]	LASSO, Ridge, Elastic Net, Ordinary Linear Regression	A mass-spring system with the average intermolecular distances	CDK	https://github.com/azevedolab/taba
Schneider et al.[73]	RF	19 terms from MedusaScore, DSX, X-Score, PLANTS, etc., and 11 QSAR topological, geometrical, constitutional and charge-based descriptors	ERα	N/A
Leidner et al.[90]	GBDT	Interaction fingerprints including vdW potential, hydrogen bonds, halogen bonds, salt bridges, π-interactions and π-cation interactions	HIV protease	N/A
Jiménez-Luna et al.[97]	CNN	Voxelized binding site with 10 atom types and 8 pharmacophoric-like properties	Public Schrödinger dataset (BACE, CDK2, JNK1, MCL1, p38, PTP1B, Thrombin, TYK2) and bromodomain dataset (BRD4), and private congeneric datasets from Janssen (PDE2, PDE3, PDE10, ROS1, BACE), Pfizer (kinase, enzyme, PDE, activator of transcription) and Biogen (tyrosine-protein kinase and receptor-associated kinase)	https://www.playmolecule.org/DeltaDelta/
D3R GC 2015[26]	Various	Various	HSP90	N/A
D3R GC 2015[26]	Various	Various	MAP4K4	N/A
D3R GC 2[27]	Various	Various	FXR	N/A
D3R GC 3[28]	Various	Various	Cathepsin S	N/A
D3R GC 3[28]	Various	Various	ABL1	N/A
D3R GC 3[28]	Various	Various	JAK	N/A
D3R GC 3[28]	Various	Various	P38-α	N/A
D3R GC 3[28]	Various	Various	TIE2	N/A
D3R GC 3[28]	Various	Various	VEGRF2	N/A

Abbreviations: CDK2, Cyclin-dependent kinase 2; CNN, convolutional neural network; DHQD, 3-dehydroquinate dehydratase; GBDT, gradient boosting decision tree; HIV PR, HIV protease; LASSO, least absolute shrinkage and selection operator; RF, random forest.

TABLE 4 Software tools and webservers for feature generation

Software name	Generating features	Availability
Descriptor Data Bank[103]	Over 2,700 proteins, ligands, and protein−ligand features	http://www.descriptordb.com
Cang et al.[70]	Topological fingerprints	https://doi.org/10.1371/journal.pcbi.1005929.s002
ODDT[104]	RF-Score, NNScore and PLEC fingerprints	https://github.com/oddt/oddt
BINANA[34]	Intermolecular descriptors	http://www.nbcr.net/binana
RF-Score v1[10]	36 intermolecular atom type pair occurrence count	http://ballester.marseille.inserm.fr/RF-Score-v1.zip
RF-Score v3[39]	36 RF-Score v1 features and 11 Vina features	https://github.com/HongjianLi/RF-Score

Last, in addition to curated structural and binding data (Table 1), another prerequisite to build ML-based SFs is to calculate some features or descriptors for every considered protein–ligand complex. Table 4 presents freely available codes for this task. Among them, we highlight Ashtawy and Mahapatra's Descriptor Data Bank (DDB)[103] for its comprehensiveness. DDB is an open-access hub for depositing, hosting, executing, and sharing descriptor extraction tools and data for a large number of interaction modeling hypotheses. The platform also implements a ML toolbox for automatic descriptor filtering and analysis and SF fitting and prediction. The descriptor filtering module is used to filter out irrelevant and/or noisy descriptors and to produce a compact subset from all available features. The authors seed DDB with 16 diverse descriptor extraction tools developed in-house and collected from the literature. The tools altogether generate over 2,700 descriptors that characterize (a) proteins, (b) ligands, and (c) protein−ligand complexes. The authors found that SFs built with multiperspective descriptor were 15% more predictive on average than their single-perspective counterparts. Furthermore, their proposed protein-specific descriptors also improved the accuracy of SFs.

5 | CONCLUSIONS

The number of studies presenting and/or evaluating ML-based SFs for BAP has boomed in the reviewed period (2015 to 2019). These SFs fall into two broad categories: generic (trained on complexes from a range of targets and hence of broad applicability) and target-specific (trained on complexes from a specific target or family of targets to be applied to this target or family).

The performance gap between generic classical and ML-based SFs was large[9] and has now broadened owing to methodological improvements. For example, on CASF-2007, RF-Score v3 obtained $R_p = 0.803$ 5 years ago,[39] whereas RF-Score v3 supplemented with ligand features now reaches $R_p = 0.836$[72] and a GBDT-based SF achieves $R_p = 0.830$.[69] On the same benchmark, DL SFs based on CNN such as TNet-BP[37] and PotentialNet[42] are closely behind with R_p values of 0.826 and 0.822, respectively. By contrast, 16 classical SFs tested on the same test set obtained a lower R_p ranging from 0.216 to 0.644 (e.g., GlideScore-XP obtained an R_p of 0.457). By comparing these SFs on the same 195 protein–ligand complexes of this pre-existing benchmark (CASF-2007), a broad comparison of ML-based and classical SFs can be made. It also has the advantage of ensuring that previously tested SFs were provided with optimal settings by their authors. Several of the classical SFs tested on CASF-2007 by Cheng et al.[23] have different versions or multiple options. However, for the sake of practicality, only the best-performing version/option of each SF was reported resulting in the set of 16 classical SFs.[23] It is important to note that the best classical SF on this test set is still X-Score for 10 years already since 2009[23] ($R_p = 0.644$). A similar performance gap is observed using other benchmarks, with X-Score also being the gold standard of classical SFs. On CASF-2013, AGL-Score achieves an R_p of 0.792, whereas the R_p of 21 classical SFs on the same test set range from 0.221 to 0.614.[69] On CASF-2016, a RF-based SF obtains an R_p of 0.824,[59] whereas the R_p of 32 classical SFs on the same test set range from 0.212 to 0.631.[64] This gap is also broadening thanks to the ever increasing training sets, as shown for CASF-2007[57] and time-stamped splits in Figure 2 and other articles.[39,61]

Generic ML-based SFs have not been systematically compared to target-specific ML-based SFs for BAP. However, there is some evidence that the latter is more predictive than the former.[97] Here too, ML-based SFs outperform classical SFs across targets, with very few exceptions like ABL1.[28] This is fully consistent with the observed performance gap between ML-based and linear regression methods in QSAR (e.g., see Table 2 of that study[105]). We are not aware of any prospective comparison between ML-based and classical SFs for SBLO. At most, SFs are compared retrospectively to

select that expected to generalize best to the test set. For example, a SVM-based SF was employed prospectively because it outperformed the five classical SFs that were considered (DrugScore, PMF, LigScore2, Jain, PLP1) on a retrospective validation.[102] Since ML-based SFs are consistently found to outperform classical SFs at retrospective BAP and SBLO validations, there is no reason to think that this trend would be any different in prospective comparisons. With that said, such in vitro confirmatory studies remain to be carried out.

Against the expectations of many experts, SFs employing DL techniques were not always more predictive than those based on more established ML techniques. For example, K_{DEEP} obtained an average R_p of 0.59 on four CSAR datasets on which RF-Score v3 achieved an average R_p of 0.70 (incidentally, the classical SF X-Score also obtained a higher average R_p of 0.64[38]). In fact, RF-Score v3 outperformed K_{DEEP} in all four CSAR datasets.[38] Another example was Pafnucy, which obtained an R_p of 0.70 at CASF-2013, but was also surpassed on this benchmark by RF-Score v3 (R_p of 0.74).[41] A further example is PotentialNet,[42] which obtained an R_p of 0.70 at sequence-based cross validations of the 2007 PDBbind refined set where RF-Score v1 obtained an R_p of 0.73 (the performance of the more advanced RF-Score v3 was not reported). Yet another example is an RF-based SF achieving lower root mean square error than MoleculeNet on several PDBbind datasets.[32] Beyond predictive accuracies, many non-DL methods possess the advantages of much shorter training times along with easier interpretation and faster rescoring of test set complexes. This trend has also been observed in other disciplines. For example, when compared to DNNs, other types of ML methods have been found to be trained faster and have overall better performance on some clinical problems.[106]

Data-driven identification of the most synergistic combination of ML regressor and featurization scheme has emerged as the most successful strategy.[70,72] When compared,[70] this strategy has been found to be more predictive than CNN's automatic feature extraction (feature learning), although this is likely to be target-dependent. In the future, we expect systematic studies that will shed light into which featurization schemes work best for each target. Such studies will be enabled by already available resources (Table 4).

Another promising avenue for future work is elucidating which datasets from other targets improve the performance of ML-based SFs on a given target. There is some work with multitarget test sets showing the benefit of training with complexes containing similar targets and similar ligands.[75] It is however unclear how much this helps depending on the considered target. Furthermore, there could as well be more effective ways to select training data instances from other targets improving performance on the considered target. For instance, the impact of similarities between complexes in terms of their features remains unexplored. While training on docked poses of known binders instead of their cocrystallized structures is now common,[2] it is still unknown how well this strategy works depending on the studied target.

The literature shows that using ML-based SFs for SBLO is a particularly promising opportunity. There are now many more of these SFs for others to use in prospective applications (Tables 2 and 3), compared to 5 years ago.[9] This should facilitate collaborations with the experimental groups that are required to validate predictions in vitro. For example, DeltaDeltaNN, which has demonstrated an outstanding level of performance on retrospective blind tests,[97] is freely available for prospective SBLO. This DNN-based target-specific SF achieved test set R_p correlations averaging 0.64 across the Pfizer targets with training sets ranging from 28 to 109 congeneric ligands docked to the target.[97] By contrast, molecular simulation-based MM-GBSA obtained an average R_p of 0.40 using the same training and test sets. In a previous prospective study, a SVM-based target-specific SF trained on a chemical series of 47 inhibitors docked to an Akt1 structure was able to discover a high proportion of nM inhibitors of this target.[102] As more targets with sufficient ligands and more user-friendly code to implement ML-based SFs become available, we expect more compelling prospective studies to be presented. Note that practically all reviewed SFs have been specifically designed for BAP and hence should not be applied to SBVS. We will dedicate a separate review of ML-based SFs designed for this related application.

CONFLICT OF INTEREST

The authors have declared no conflicts of interest for this article.

AUTHOR CONTRIBUTIONS

Hongjian Li: Investigation; software; writing-original draft, review, and editing. **Kam-Heung Sze**: Investigation; software; visualization. **Gang Lu**: Resources; writing-review and editing. **Pedro Ballester**: Conceptualization; investigation; resources; writing-original draft, review, and editing.

ORCID

Hongjian Li https://orcid.org/0000-0001-8467-638X
Pedro J. Ballester https://orcid.org/0000-0002-4078-743X

RELATED WIRES ARTICLES

Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening
From machine learning to deep learning: Advances in scoring functions for protein-ligand docking

REFERENCES

1. Coleman RG, Carchia M, Sterling T, Irwin JJ, Shoichet BK. Ligand pose and orientational sampling in molecular docking. *PLoS One*. 2013;*8*(10):e75992.
2. Li H, Leung K-S, Wong M-H, Ballester PJ. Correcting the impact of docking pose generation error on binding affinity prediction. *BMC Bioinformatics*. 2016;*17*(11):308.
3. Lee S, Blundell TL. BIPA: A database for protein-nucleic acid interaction in 3D structures. *Bioinformatics*. 2009;*25*(12):1559–1560.
4. Fontaine F, Overman J, Moustaqil M, et al. Small-molecule inhibitors of the SOX18 transcription factor. *Cell Chem Biol*. 2017;*24*(3): 346–359.
5. Makley LN, Gestwicki JE. Expanding the number of "druggable" targets: Non-enzymes and protein-protein interactions. *Chem Biol Drug Des*. 2013;*81*(1):22–32.
6. Wójcikowski M, Siedlecki P, Ballester PJ. Building machine-learning scoring functions for structure-based prediction of intermolecular binding affinity. *Methods Mol Biol*. 2019;*2053*:1–12.
7. Cheng T, Li Q, Zhou Z, Wang Y, Bryant SH. Structure-based virtual screening for drug discovery: A problem-centric review. *AAPS J*. 2012;*14*(1):133–141.
8. Joseph-McCarthy D, Baber JC, Feyfant E, Thompson DC, Humblet C. Lead optimization via high-throughput molecular docking. *Curr Opin Drug Discov Devel*. 2007;*10*(3):264–274.
9. Ain Qurrat U, Aleksandrova A, Roessler FD, Ballester PJ. Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening. *WIERs Comput Mol Sci*. 2015;*5*(6):405–424.
10. Ballester PJ, Mitchell JBO. A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking. *Bioinformatics*. 2010;*26*(9):1169–1175.
11. Schmidhuber J. Deep learning in neural networks: An overview. *Neural Netw*. 2015;*61*:85–117.
12. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Proceedings of the 25th International Conference on Neural Information Processing Systems—Volume 1; NIPS'12; Red Hook, NY: Curran Associates Inc.; 2012; p. 1097–1105.
13. Nassif AB, Shahin I, Attili I, Azzeh M, Shaalan K. Speech recognition using deep neural networks: A systematic review. *IEEE Access*. 2019;*7*:19143–19165.
14. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining—KDD '16; New York, NY: ACM Press; 2016; p. 785–794.
15. Yang X, Wang Y, Byrne R, Schneider G, Yang S. Concepts of artificial intelligence for computer-assisted drug discovery. *Chem Rev*. 2019;*119*(18):10520–10594.
16. Batool M, Ahmad B, Choi S. A structure-based drug discovery paradigm. *Int J Mol Sci*. 2019;*20*:2783.
17. Li J, Fu A, Zhang L. An overview of scoring functions used for protein–ligand interactions in molecular docking. *Interdiscip Sci Comput Life Sci*. 2019;*11*(2):320–328.
18. Jensen KF, Coley CW, Eyke NS. Autonomous discovery in the chemical sciences part I: Progress. *Angew Chem Int Ed*. 2020. https://doi.org/10.1002/anie.201909987
19. Shen C, Ding J, Wang Z, Cao D, Ding X, Hou T. From machine learning to deep learning: Advances in scoring functions for protein–ligand docking. *WIREs Comput Mol Sci*. 2019;*10*:e1429.
20. Pinzi L, Rastelli G. Molecular docking: Shifting paradigms in drug discovery. *Int J Mol Sci*. 2019;*20*(18):4331.
21. Zhang L, Tan J, Han D, Zhu H. From machine learning to deep learning: Progress in machine intelligence for rational drug discovery. *Drug Discov Today*. 2017;*22*:1680–1685.
22. Qiu T, Qiu J, Feng J, et al. The recent progress in proteochemometric modelling: Focusing on target descriptors, cross-term descriptors and application scope. *Brief Bioinform*. 2017;*18*(1):125–136.
23. Cheng T, Li X, Li Y, Liu Z, Wang R. Comparative assessment of scoring functions on a diverse test set. *J Chem Inf Model*. 2009;*49*(4): 1079–1093.
24. Li Y, Liu Z, Li J, et al. Comparative assessment of scoring functions on an updated benchmark: 1. Compilation of the test set. *J Chem Inf Model*. 2014;*54*(6):1700–1716.
25. Su M, Du Y, Yang Q, Wang R, Liu Z, Feng G, Li Y. Comparative assessment of scoring functions: The CASF-2016 update. *J Chem Inf Model*. 2018;*59*(2):895–913.
26. Gathiaka S, Liu S, Chiu M, et al. D3R grand challenge 2015: Evaluation of protein-ligand pose and affinity predictions. *J Comput Aided Mol Des*. 2016;*30*(9):651–668.

27. Gaieb Z, Liu S, Gathiaka S, et al. D3R grand challenge 2: Blind prediction of protein-ligand poses, affinity rankings, and relative binding free energies. *J Comput Aided Mol Des*. 2018;*32*(1):1–20.
28. Gaieb Z, Parks CD, Chiu M, et al. D3R grand challenge 3: Blind prediction of protein–ligand poses and affinity rankings. *J Comput Aided Mol Des*. 2019;*33*(1):1–18.
29. Gilson MK, Liu T, Baitaluk M, Nicola G, Hwang L, Chong J. BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res*. 2016;*44*(D1):D1045–D1053.
30. Carlson HA, Smith RD, Damm-Ganamet KL, et al. CSAR 2014: A benchmark exercise using unpublished data from Pharma. *J Chem Inf Model*. 2016;*56*(6):1063–1077.
31. Smith RD, Clark JJ, Ahmed A, Orban ZJ, Dunbar JB, Carlson HA. Updates to binding MOAD (mother of all databases): Polypharmacology tools and their utility in drug repurposing. *J Mol Biol*. 2019;*431*(13):2423–2433.
32. Wu Z, Ramsundar B, Feinberg EN, et al. MoleculeNet: A benchmark for molecular machine learning. *Chem Sci*. 2018;*9*(2):513–530.
33. Hassan MM, Mogollon DC, Fuentes O, Sirimulla S. DLSCORE: A deep learning model for predicting protein-ligand binding affinities. *ChemRxiv*. 2018.
34. Durrant JD, McCammon JA. BINANA: A novel algorithm for ligand-binding characterization. *J Mol Graph Model*. 2011;*29*(6):888–893.
35. Durrant JD, McCammon JA. NNScore 2.0: A neural-network receptor–ligand scoring function. *J Chem Inf Model*. 2011;*51*(11):2897–2903.
36. Trott O, Olson AJ. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem*. 2010;*31*(2):455–461.
37. Cang Z, Wei G-W. TopologyNet: Topology based deep convolutional and multi-task neural networks for biomolecular property predictions. *PLoS Comput Biol*. 2017;*13*(7):e1005690.
38. Jiménez J, Škalič M, Martínez-Rosell G, De Fabritiis G. K_{DEEP}: Protein-ligand absolute binding affinity prediction via 3D-convolutional neural networks. *J Chem Inf Model*. 2018;*58*(2):287–296.
39. Li H, Leung K-S, Wong M-H, Ballester PJ. Improving AutoDock Vina using random forest: The growing accuracy of binding affinity prediction by the effective exploitation of larger data sets. *Mol Inform*. 2015;*34*(2–3):115–126.
40. Wang R, Lai L, Wang S. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J Comput Aided Mol Des*. 2002;*16*(1):11–26.
41. Stepniewska-Dziubinska MM, Zielenkiewicz P, Siedlecki P. Development and evaluation of a deep learning model for protein-ligand binding affinity prediction. *Bioinformatics*. 2018;*34*(21):3666–3674.
42. Feinberg EN, Sur D, Wu Z, et al. PotentialNet for molecular property prediction. *ACS Cent Sci*. 2018;*4*(11):1520–1530.
43. Ballester PJ, Schreyer A, Blundell TL. Does a more precise chemical description of protein-ligand complexes lead to more accurate prediction of binding affinity? *J Chem Inf Model*. 2014;*54*(3):944–955.
44. Ashtawy HM, Mahapatra NR. A comparative assessment of predictive accuracies of conventional and machine learning scoring functions for protein-ligand binding affinity prediction. *IEEE/ACM Trans Comput Biol Bioinforma*. 2015;*12*(2):335–347.
45. Schnecke V, Kuhn LA. Virtual screening with solvation and ligand-induced complementarity. *Perspect Drug Discov Des*. 2000;*20*:171–190.
46. Ashtawy HM, Mahapatra NR. BgN-score and BsN-score: Bagging and boosting based ensemble neural networks scoring functions for accurate binding affinity prediction of protein-ligand complexes. *BMC Bioinformatics*. 2015;*16*(4):S8.
47. Jones G, Willett P, Glen RC, Leach AR, Taylor R. Development and validation of a genetic algorithm for flexible docking. *J Mol Biol*. 1997;*267*(3):727–748.
48. Khamis MA, Gomaa W. Comparative assessment of machine-learning scoring functions on PDBbind 2013. *Eng Appl Artif Dermatol Int*. 2015;*45*:136–151.
49. Hildebrandt A, Dehof AK, Rurainski A, et al. BALL—Biochemical algorithms library 1.3. *BMC Bioinformatics*. 2010;*11*(1):531.
50. Zavodsky MI, Sanschagrin PC, Korde RS, Kuhn LA. Distilling the essential features of a protein surface for improving protein-ligand docking, scoring, and virtual screening. *J Comput Aided Mol Des*. 2002;*16*(12):883–902.
51. Pires DEV, Ascher DB. CSM-Lig: A web server for assessing and comparing protein-small molecule affinities. *Nucleic Acids Res*. 2016;*44*(W1):W557–W561.
52. Cang Z, Wei G-W. Integration of element specific persistent homology and machine learning for protein-ligand binding affinity prediction. *Int J Numer Method Biomed Eng*. 2018;*34*(2):e2914.
53. Xavier MM, Heck GS, de Avila MB, et al. SAnDReS a computational tool for statistical analysis of docking results and development of scoring functions. *Comb Chem High Throughput Screen*. 2016;*19*(10):801–812.
54. Bitencourt-Ferreira G, de Azevedo WF. Development of a machine-learning model to predict Gibbs free energy of binding for protein-ligand complexes. *Biophys Chem*. 2018;*240*:63–69.
55. Morris GM, Huey R, Lindstrom W, et al. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J Comput Chem*. 2009;*30*(16):2785–2791.
56. Thomsen R, Christensen MH. MolDock: A new technique for high-accuracy molecular docking. *J Med Chem*. 2006;*49*(11):3315–3321.
57. Li H, Peng J, Sidorov P, et al. Classical scoring functions for docking are unable to exploit large volumes of structural and interaction data. *Bioinformatics*. 2019;*35*(20):3989–3995.
58. Cao Y, Li L. Improved protein-ligand binding affinity prediction by using a curvature-dependent surface-area model. *Bioinformatics*. 2014;*30*(12):1674–1680.

59. Afifi K, Al-Sadek AF. Improving classical scoring functions using random forest: The non-additivity of free energy terms' contributions in binding. *Chem Biol Drug Des*. 2018;*92*(2):1429–1434.
60. Li H, Leung KS, Wong MH, Ballester PJ. Substituting random forest for multiple linear regression improves binding affinity prediction of scoring functions: Cyscore as a case study. *BMC Bioinformatics*. 2014;*15*(1):291.
61. Li H, Leung KS, Wong MH, Ballester PJ. Low-quality structural and interaction data improves binding affinity prediction via random forest. *Molecules*. 2015;*20*(6):10947–10962.
62. Wang C, Zhang Y. Improving scoring-docking-screening powers of protein–ligand scoring functions using random forest. *J Comput Chem*. 2017;*38*(3):169–177.
63. Lu Y, Wang R, Yang CY, Wang S. Analysis of ligand-bound water molecules in high-resolution crystal structures of protein-ligand complexes. *J Chem Inf Model*. 2007;*47*(2):668–675.
64. Lu J, Hou X, Wang C, Zhang Y. Incorporating explicit water molecules and ligand conformation stability in machine-learning scoring functions. *J Chem Inf Model*. 2019;*59*(11):4540–4549.
65. Wang B, Zhao Z, Nguyen DD, Wei GW. Feature functional theory–binding predictor (FFT–BP) for the blind prediction of binding free energies. *Theor Chem Accounts*. 2017;*136*(4):55.
66. Nguyen DD, Xiao T, Wang M, Wei GW. Rigidity strengthening: A mechanism for protein-ligand binding. *J Chem Inf Model*. 2017;*57*(7):1715–1721.
67. Ashtawy HM, Mahapatra NR. Task-specific scoring functions for predicting ligand binding poses and affinity and for screening enrichment. *J Chem Inf Model*. 2018;*58*(1):119–133.
68. Neudert G, Klebe G. DSX: A knowledge-based scoring function for the assessment of protein-ligand complexes. *J Chem Inf Model*. 2011;*51*(10):2731–2745.
69. Nguyen DD, Wei G-W. AGL-score: Algebraic graph learning score for protein–ligand binding scoring, ranking, docking, and screening. *J Chem Inf Model*. 2019;*59*(7):3291–3304.
70. Cang Z, Mu L, Wei G-W. Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening. *PLoS Comput Biol*. 2018;*14*(1):e1005929.
71. Nguyen DD, Wei GW. DG-GL: Differential geometry-based geometric learning of molecular datasets. *Int J Numer Method Biomed Eng*. 2019;*35*(3):e3179.
72. Boyles F, Deane CM, Morris GM. Learning from the ligand: Using ligand-based features to improve binding affinity prediction. *Bioinformatics*. 2019;btz665.
73. Schneider M, Pons J-L, Bourguet W, Labesse G. Towards accurate high-throughput ligand affinity prediction by exploiting structural ensembles, docking metrics and ligand similarity. *Bioinformatics*. 2019;*36*(1):160–168.
74. Li Y, Yang J. Structural and sequence similarity makes a significant impact on machine-learning-based scoring functions for protein-ligand interactions. *J Chem Inf Model*. 2017;*57*(4):1007–1012.
75. Li H, Peng J, Leung Y, et al. The impact of protein structure and sequence similarity on the accuracy of machine-learning scoring functions for binding affinity prediction. *Biomol Ther*. 2018;*8*(1):12.
76. Li Y, Su M, Liu Z, et al. Assessing protein–ligand interaction scoring functions with the CASF-2013 benchmark. *Nat Protoc*. 2018;*13*(4):666–680.
77. Ouyang X, Handoko SD, Kwoh CK. CScore: A simple yet effective scoring function for protein–ligand binding affinity prediction using modified Cmac learning architecture. *J Bioinforma Comput Biol*. 2011;*9*(supp01):1–14.
78. Li X, Zhu M, Li X, Wang H-Q, Wang S. Protein-protein binding affinity prediction based on an SVR ensemble. In: Huang D-S, Jiang C, Bevilacqua V, Figueroa J, editors. *Intelligent computing technology*. Volume 7389. Berlin/Heidelberg, Germany: Springer, 2012; p. 145–151.
79. Geng C, Xue LC, Roel-Touris J, Bonvin AMJJ. Finding the ΔΔ *G* spot: Are predictors of binding affinity changes upon mutations in protein-protein interactions ready for it? *WIREs Comput Mol Sci*. 2019;*9*(5):e1410.
80. Han L, Yang Q, Liu Z, Li Y, Wang R. Development of a new benchmark for assessing the scoring functions applicable to protein–protein interactions. *Future Med Chem*. 2018;*10*(13):1555–1574.
81. de Ávila MB, Xavier MM, Pintro VO, de Azevedo WF. Supervised machine learning techniques to predict binding affinity. A study for Cyclin-dependent kinase 2. *Biochem Biophys Res Commun*. 2017;*494*(1–2):305–310.
82. Pintro VO, de Azevedo WF. Optimized virtual screening workflow. Towards target-based polynomial scoring functions for HIV-1 protease. *Comb Chem High Throughput Screen*. 2017;*20*(9):820–827.
83. Levin NMB, Pintro VO, Bitencourt-Ferreira G, de Mattos BB, de Castro Silvério A, de Azevedo WF. Development of CDK-targeted scoring functions for prediction of binding affinity. *Biophys Chem*. 2018;*235*:1–8.
84. de Ávila MB, de Azevedo WF. Development of machine learning models to predict inhibition of 3-dehydroquinate dehydratase. *Chem Biol Drug Des*. 2018;*92*(2):1468–1474.
85. da Silva AD, Bitencourt-Ferreira G, de Azevedo WF. Taba: A tool to analyze the binding affinity. *J Comput Chem*. 2019;*41*(1):69–73.
86. Ballester PJ, Mangold M, Howard NI, et al. Hierarchical virtual screening for the discovery of new molecular scaffolds in antibacterial hit identification. *J R Soc Interface*. 2012;*9*(77):3196–3207.
87. Korb O, Stützle T, Exner TE. Empirical scoring functions for advanced protein-ligand docking with PLANTS. *J Chem Inf Model*. 2009;*49*(1):84–96.

88. Yin S, Biedermannova L, Vondrasek J, Dokholyan NV. MedusaScore: An accurate force field-based scoring function for virtual drug screening. *J Chem Inf Model*. 2008;*48*(8):1656–1662.
89. Durrant JD, Carlson KE, Martin TA, et al. Neural-network scoring functions identify structurally novel estrogen-receptor ligands. *J Chem Inf Model*. 2015;*55*(9):1953–1961.
90. Leidner F, Kurt Yilmaz N, Schiffer CA. Target-specific prediction of ligand affinity with structure-based interaction fingerprints. *J Chem Inf Model*. 2019;*59*(9):3679–3691.
91. Rodríguez-Pérez R, Bajorath J. Interpretation of compound activity predictions from complex machine learning models using local approximations and shapley values. *J Med Chem*. 2019.
92. Nogueira MS, Koch O. The development of target-specific machine learning models as scoring functions for docking-based target prediction. *J Chem Inf Model*. 2019;*59*(3):1238–1252.
93. Lee A, Kim D. CRDS: Consensus reverse docking system for target fishing. *Bioinformatics*. 2019;btz656.
94. Peón A, Dang CC, Ballester PJ. How reliable are ligand-centric methods for target fishing? *Front Chem*. 2016;*4*:15.
95. Peón A, Li H, Ghislat G, et al. MolTarPred: A web tool for comprehensive target prediction with reliability estimation. *Chem Biol Drug Des*. 2019;*94*(1):1390–1401.
96. Wójcikowski M, Ballester PJ, Siedlecki P. Performance of machine-learning scoring functions in structure-based virtual screening. *Sci Rep*. 2017;*7*:46710.
97. Jiménez-Luna J, Pérez-Benito L, Martinez-Rosell G, et al. DeltaDelta neural networks for lead optimization of small molecule potency. *Chem Sci*. 2019;*10*(47):10911–10918.
98. Wang L, Wu Y, Deng Y, et al. Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. *J Am Chem Soc*. 2015;*137*(7):2695–2703.
99. Nguyen DD, Cang Z, Wu K, Wang M, Cao Y, Wei GW. Mathematical deep learning for pose and binding affinity prediction and ranking in D3R grand challenges. *J Comput Aided Mol Des*. 2019;*33*(1):71–82.
100. Li H, Leung K-S, Ballester PJ, Wong M-H. Istar: A web platform for large-scale protein-ligand docking. *PLoS One*. 2014;*9*(1):e85678.
101. Warren GL, Andrews CW, Capelli A-M, et al. A critical assessment of docking programs and scoring functions. *J Med Chem*. 2006;*49*(20):5912–5931.
102. Zhan W, Li D, Che J, et al. Integrating docking scores, interaction profiles and molecular descriptors to improve the accuracy of molecular docking: Toward the discovery of novel Akt1 inhibitors. *Eur J Med Chem*. 2014;*75*:11–20.
103. Ashtawy HM, Mahapatra NR. Descriptor data Bank (DDB): A cloud platform for multiperspective modeling of protein-ligand interactions. *J Chem Inf Model*. 2018;*58*(1):134–147.
104. Wójcikowski M, Zielenkiewicz P, Siedlecki P. Open drug discovery toolkit (ODDT): A new open-source player in the drug discovery field. *J Chem*. 2015;*7*(1):26.
105. Olier I, Sadawi N, Bickerton GR, et al. Meta-QSAR: A large-scale application of meta-learning to drug design and discovery. *Mach Learn*. 2018;*107*(1):285–311.
106. Chen D, Liu S, Kingsbury P, et al. Deep learning and alternative learning strategies for retrospective real-world clinical data. *npj Digit Med*. 2019;*2*(1):43.

How to cite this article: Li H, Sze K-H, Lu G, Ballester PJ. Machine-learning scoring functions for structure-based drug lead optimization. *WIREs Comput Mol Sci*. 2020;10:e1465. https://doi.org/10.1002/wcms.1465