JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

ACS AUTHORCHOICE

pubs.acs.org/jcim Article

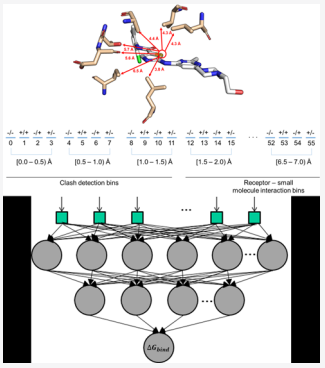
General Purpose Structure-Based Drug Discovery Neural Network Score Functions with Human-Interpretable Pharmacophore Maps

Benjamin P. Brown,* Jeffrey Mendenhall, Alexander R. Geanes, and Jens Meiler*

Cite This: J. Chem. Inf. Model. 2021, 61, 603−620 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The BioChemical Library (BCL) is an academic open-source cheminformatics toolkit comprising ligand-based virtual high-throughput screening (vHTS) tools such as quantitative structure−activity/property relationship (QSAR/QSPR) modeling, small molecule flexible alignment, small molecule conformer generation, and more. Here, we expand the capabilities of the BCL to include structure-based virtual screening. We introduce two new score functions, BCL-AffinityNet and BCL-DockANNScore, based on novel distance-dependent signed protein−ligand atomic property correlations. Both metrics are conventional feed-forward dropout neural networks trained on the new descriptors. We demonstrate that BCL-AffinityNet is one of the top performing score functions on the comparative assessment of score functions 2016 affinity prediction and affinity ranking tasks. We also demonstrate that BCL-AffinityNet performs well on the CSAR-NRC HiQ I and II test sets. Furthermore, we demonstrate that BCL-DockANNScore is competitive with multiple state-of-the-art methods on the docking power and screening power tasks. Finally, we show how our models can be decomposed into human-interpretable pharmacophore maps to aid in hit/lead optimization. Altogether, our results expand the utility of the BCL for structure-based scoring to aid small molecule discovery and design. BCL-AffinityNet, BCL-DockANNScore, and the pharmacophore mapping application, as well as the remainder of the BCL cheminformatics toolkit, are freely available with an academic license at the BCL Commons site hosted on http://meilerlab.org/.



INTRODUCTION

Computer-aided drug discovery (CADD) is a broad category of methods that can be employed to increase the efficiency of the drug discovery process. Broadly, CADD methods can be subdivided into two categories: ligand-based (LB) and structure-based (SB). LB methods predominantly employ similarity metrics to compare ligands with known biological activity or chemical attributes to a library of prospective small molecules. Among the most widely used LB methods are quantitative structure−activity relationship (QSAR) models, which relate quantitative chemical descriptors of molecules to known biological activities.[1,2] QSAR models lend themselves to supervised machine learning methods, such as artificial neural networks (ANNs) and random forest (RF).[3−8] Indeed, over the last two decades, we have demonstrated the efficacy of ANNs in LB classification tasks compared to other methods, such as support vector machines, and employed them to identify multiple G-protein-coupled receptor (GPCR) allosteric modulators.[3,9−12] At that time, we have contributed to multiple aspects of QSAR method development, including early efforts to expedite model training with graphics processing unit (GPU) programming,[13] chemical descriptor, and toolkit development,[14−16] improving QSAR ANN architectures with dropout,[5] and dataset assembly for community benchmarking.[17,18] We have accomplished this largely with the development of the BioChemical Library (BCL), a primarily ligand-based academic open-source cheminformatics toolkit. LB methods can often rank compounds many orders of magnitude faster than SB methods. Despite being very rapid and easily deployed on large databases for virtual high-throughput screening (vHTS), ligand-based methods have inherent limitations. Most notably, LB methods make predictions in the absence of binding pocket information. As a result, predictions made from LB methods must be target-specific, and generating LB models for a given target, especially QSAR models, may require a large amount of model training data.[1−8] Thus, there is considerable interest in developing target agnostic, rapid SB methods for vHTS.

SB methods provide information about small molecule interactions with the binding pocket. Critically, this should allow SB methods to be target agnostic and provide chemically meaningful insight with which to guide hit optimization. Unfortunately, the most accurate SB methods come with a computational cost prohibitive for vHTS. Accurate prediction of small molecule binding affinities to target proteins is a key challenge in SB CADD. Structure-based alchemical free energy

Received: August 26, 2020
Published: January 26, 2021

https://dx.doi.org/10.1021/acs.jcim.0c01001
J. Chem. Inf. Model. 2021, 61, 603−620

approaches, such as free energy perturbation (FEP) and thermodynamic integration (TI), are widely considered to be the most accurate.[19−21] Other approaches, such as molecular mechanics Poisson−Boltzmann or generalized-Born surface area (MM/PB(GB)SA), or protein−ligand docking semi-empirical scoring functions, can also provide reliable relative binding free energies, but with overall performance seemingly being more system dependent.[22−25] Faster but less accurate docking score functions are being increasingly scaled to medium- and high-throughput virtual screening.[26,27]

In the last decade, many machine learning approaches have been developed to increase the speed and accuracy of SB virtual screening approaches. As early as 2010, random forest (RF) rescoring of docked poses demonstrated that machine learning algorithms could provide rapid and competitive prediction of protein−ligand binding affinities (RF-Score).[28] A variation on RF as a modeling tool for protein−ligand binding affinity prediction is $\Delta_{Vina}RF_{20}$, which uses random forest (RF) to predict an error correction term for the AutoDock Vina docking score function.[29] More recently, deep learning with convolutional neural networks (CNNs) has been widely investigated to predict binding affinities. For example, DeepVS is a CNN that attempts to generalize binding mode information by encoding local atomic neighborhoods around each selected ligand atom using simple descriptors (i.e., atom types, charges, distances, and interacting amino acid identity).[30] Multiple grid-based CNNs have also been developed, such as K_{DEEP} and a CNN, by Ragoza et al., which treat protein−ligand complexes as three-dimensional (3D) images colored by specific atom type and pharmacophore properties.[31,32] AtomNet is another grid-based CNN that also includes features derived from protein−ligand interaction fingerprints.[33]

It is well known that cheminformatics machine learning algorithms can be strongly limited in their domain of applicability by the chosen training set and descriptors.[34−40] There is concern that some newer CNN techniques demonstrating exceptional performance may suffer from the lack of generalizability owing to the dataset and training biases.[31,41] Even in cases where machine learning models make accurate predictions, the chemical basis of these predictions is not easily interpreted without substantial input sensitivity and feature analysis. This infamously gives rise to the "black box" problem of machine learning algorithms, especially deep neural networks (DNNs).

Finally, a major motivation for the current project is to incorporate a modular and customizable SB score function into the BCL for use in the ongoing development of SB design algorithms. Currently, the BCL is only able to support LB design algorithms. Ultimately, we anticipate that increasing the capabilities of the BCL to perform both LB and SB design tasks will make it a valuable companion to other academic molecular modeling software projects, such as the Rosetta[42] macromolecular modeling and design software suite.

To address these issues, we have designed a novel SB protein−ligand binding affinity and pose prediction model based on distance-dependent signed atom property protein−ligand correlations (PLCs). Instead of encoding specific protein and ligand properties, our method encodes the protein−ligand interaction feature space. This is analogous to the formation of statistical pair potentials, except that here we do not formally provide any constraints on the function to be approximated. We demonstrate that fully connected feed-forward neural networks trained with our new descriptors are competitive with the existing state-of-the-art machine learning methods and docking methods at protein−ligand binding affinity prediction, pose prediction, and virtual screening power. Moreover, we explicitly demonstrate that the performance of our models is not dependent on exploiting dataset bias. Finally, we show how our models can be rapidly decomposed into human-interpretable pharmacophore maps. These pharmacophore maps allow users to visualize the atoms/substructures of their molecules that drive the activity prediction, as well as map predicted or known relative binding free energy changes across molecule ensembles to specific substructures. This will be the first SB scoring tool available in the BCL, and the pharmacophore mapping tool is fully compatible with the LB QSAR methods currently implemented. Together, we believe that these tools improve the utility of the BCL for SB hit identification and lead to optimization in drug discovery.

The new descriptors, models, and pharmacophore mapping application will be available in the upcoming BCL version 4.1 release, an academic open-source software package for cheminformatics written in the C++ programming language. It is our hope that our new method will be used in conjunction with other advancements in machine learning-based QSAR/QSPR to continue to improve the efficiency of drug discovery.

RESULTS

Development of a Pose-Dependent Protein−Ligand Property Correlation Descriptor. Currently, the top performing deep learning scoring algorithms that predict binding affinities from protein−ligand complexes are CNNs that encode neighboring ligands and receptor atoms spatially and/or chemically (e.g., hydrogen bond donor/acceptor heuristics).[31,32] One critique of these CNNs is that test-set performance can be attributed to learning ligand-specific features and not the protein−ligand interface features.[41] In other words, the neural network can perform well on the tests simply by learning the biases in the ligand datasets. To avoid any such potential limitations here, we developed a pose-dependent protein−ligand interaction descriptor based on sign-aware 3D autocorrelations (3DAs). This descriptor can be likened to a potential of mean force profile in which the collective variables are the pairwise interatomic distances between the protein and ligand atoms for specific chemical properties/heuristics.

Small Molecule Chemical Property Autocorrelations. Consider a property-weighted 3D autocorrelation (3DA) function for a single small molecule. An atom-based property allows the 3DA to represent the spatial distribution of properties of interest

$$\mathrm{3DA}(r_a, r_b) = \sum_{j}^{N}\sum_{i}^{N} \delta(r_a \le r_{i,j} < r_b) P_i P_j \, e^{-\beta r_{i,j}^2} \quad (1)$$

where r_a and r_b are the boundaries of the current distance interval, N is the total number of atoms in the molecule, $r_{i,j}$ is the distance between the two atoms being considered, δ is the Kronecker delta, β is a smoothing parameter referred to as "temperature",[15,43] and P is the property computed for each atom. 3DAs computed for signed properties (e.g., partial charge) contain, for each distance interval, three values corresponding to product sums of each of the three possible sign pairings (−/−, +/+, −/+).[14]

https://dx.doi.org/10.1021/acs.jcim.0c01001
J. Chem. Inf. Model. 2021, 61, 603−620

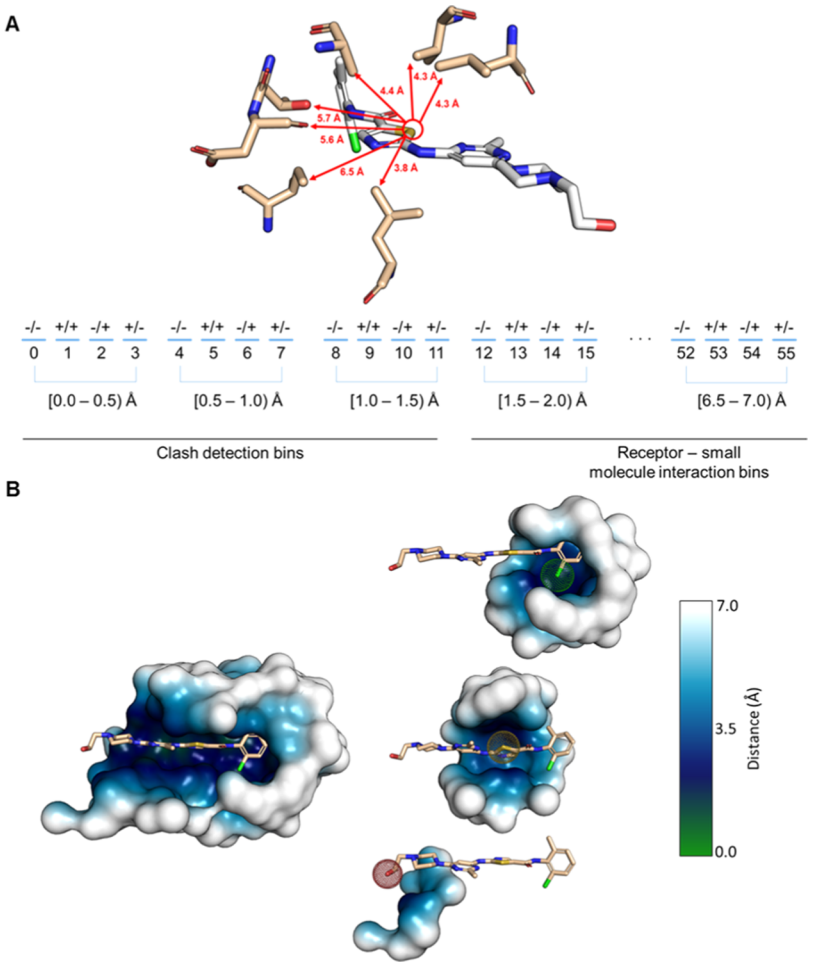


Figure 1. Schematic of a pose-dependent protein–ligand descriptor. (A) Schematic representation of pose-dependent protein–ligand interaction feature space. (B) Surface representation of discoidin domain receptor 1 (DDR1) kinase binding pocket heavy atoms within 7.0 Å of select atoms within dasatinib. The surface representation is colored by the distance to the selected atom. Dasatinib shown in stick configuration colored by element type with the selected atom is indicated by a dot sphere.

Recasting Property Space into Protein–Ligand Interaction Distance Bins. Instead of corresponding to intramolecular atomic distances, the distance bins now correspond to intermolecular protein–ligand interatomic distances. The property correlation is between each atom in the ligand and all atoms in the receptor within a specified radius (Figure 1)

$$\mathrm{PLC}(r_a, r_b) = \sum_{l}^{N_{lig}} \sum_{p}^{N_{prot}} \delta(r_a \le r_{l,p} < r_b) P_l P_p \, e^{-\beta r_{l,p}^2} \quad (2)$$

where r_a and r_b are the boundaries of the current protein–ligand interatomic distance interval, N_{lig} and N_{prot} are the total number of atoms in the ligand and receptor, respectively, $r_{l,p}$ is the distance between the current protein–ligand atom pair, δ is the Kronecker delta, β is the temperature, and P_l and P_p are the properties computed for ligand and receptor atoms l and p, respectively. As with 3DA in eq 1, protein–ligand correlation (PLC) descriptors distinguish signed pairs but can also optionally include an additional bin (−−/++/−+/+−) to account for opposite sign pairings between the protein and the ligand (Figure 1A). This can be useful if the properties between which the correlations are being taken are not identical or if the model being built is leveraging pre-existing knowledge about the chemical makeup of the system in the study.

For example, consider the descriptor "HBondDonorTernary". This descriptor returns a 1 if an atom is a hydrogen bond donor, −1 if it is a hydrogen bond acceptor, and 0 otherwise (Table S1). One could choose to differentiate hydrogen bond donor/acceptor pairs between the protein and the ligand (e.g., asymmetric: −+/+−) or to group all opposite sign pairs together (symmetric −/+). Sign pair discrimination is illustrated in Figure 1A for a property that tracks the protein–ligand directionality of opposite sign pairings. We empirically chose a total distance of 7.0 Å discretized at 0.50 Å intervals, resulting in either 42 (symmetric) or 56 (asymmetric) values per property (see the subsection on feature parameterization in Methods and the Supporting Information).

Representing Protein–Ligand Interactions with Property Correlation Descriptors. PLC descriptors (eq 2) encode interactions between protein and ligand atomic atoms as represented by a variety of atomic properties: partial charge, electronegativity, polarizability, hydrophobicity, hydrogen bond donors and acceptors, aromatic and generic ring membership, heavy and light atoms (Table S1). These atomic features are a superset of those we used previously for QSAR,[5,14] and are identical to those we used previously for the superimposition of similar molecules.[44]

https://dx.doi.org/10.1021/acs.jcim.0c01001
J. Chem. Inf. Model. 2021, 61, 603–620

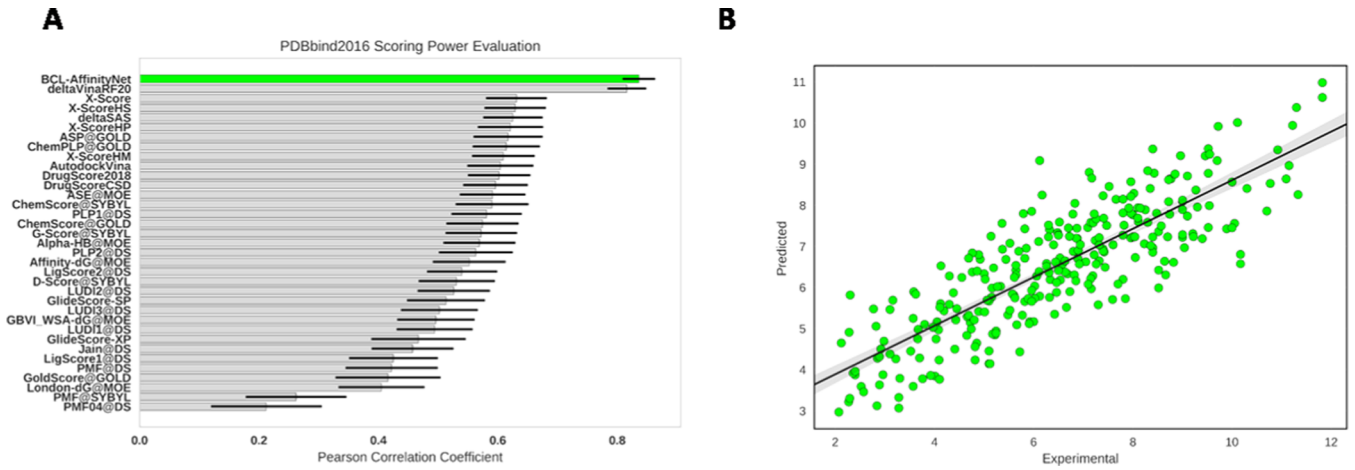


Figure 2. Scoring power evaluation of BCL-AffinityNet. (A) Comparison of BCL-AffinityNet scoring power to other methods from the CASF2016 benchmark by Su et al.[22] Error bars indicate the 90% confidence interval. (B) Linear regression of experimental vs predicted pK_d values in the CASF2016 coreset.

To mitigate feature redundancy, we summed feature interactions that were nominally equivalent. For example, consider the PLC descriptor that represents the signed correlation between atomic partial charges in receptor and ligand atoms: 3DAPairRS050(Atom_SigmaCharge) (Tables S2 and S3). In this descriptor, we summed −/+ (ligand negative charge, protein positive charge) with +/− (ligand positive charge, protein negative charge) interactions under the notion that these are equivalently favorable pairings. We took a similar approach for hydrogen bond donation, hydrophobic interactions, and heavy atom/hydrogen atom discrimination. Some descriptors, such as polarizability and electronegativity, are strictly positive valued, and therefore do not require binning by sign pairs (Tables S2 and S3).

While each of the previously mentioned descriptors can be considered symmetric in that we are correlating the same property for both the receptor and the ligand (e.g., partial charge), interactions can also be described by complementary interactions between dissimilar chemical properties. For example, interactions between aromatic ring systems and polar vs hydrophobic atoms. To create a property that can describe this interaction, we need to utilize Atom_HydrophobicTernary, which is an atom property that encodes hydrophobic atoms as +1, and polar atoms as −1. To better distinguish highly polar from less polar atoms, we multiply Atom_HydrophobicTernary by polarizability. We then encode aromatic−polar, aromatic−hydrophobic interactions with the PLC descriptor, "3DAPairRSAsym050(Multiply-(Atom_HydrophobicTernary, Atom_Polarizability), Atom_-IsInAromaticRingTernary)". In this descriptor, each distance bin is further discretized into −/− (ligand polar atom polarizability with a nonaromatic receptor atom), +/+ (ligand hydrophobic atom polarizability with an aromatic receptor atom), −/+ (ligand polar atom polarizability with an aromatic receptor atom), and +/− (ligand hydrophobic atom polarizability with a nonaromatic receptor atom) (Tables S2 and S3). An inverted version of this descriptor, in which hydrophobicity is with respect to the receptor and aromaticity to the ligand, is also employed here.

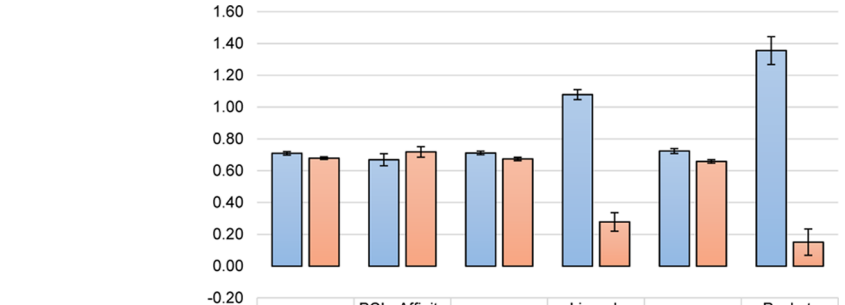
With these features, we trained two neural networks. BCL-AffinityNet is a "deep" single-task neural network (2 hidden layers, 512 neurons in the first hidden layer, and 32 neurons in the second layer) to directly predict log-scaled protein−ligand binding affinity values. BCL-DockANNScore is a multitasking shallow neural network (1 hidden layer with 32 neurons) that classifies binding poses as less-or-equal to 1.0, 2.0, 3.0, 5.0, or 8.0 Å from the native (cocrystallized) binding mode. Both of these models utilize only PLC descriptors (eq 2), with BCL-DockANNScore, including an additional PLC descriptor that discretizes hydrogen bond donor/receiver pair angles (Table S3; see the Supporting Information for details).

Finally, we note that we did not perform a deep exploration of possible base chemical descriptors and there are likely many additional features that could be effective (e.g., explicit consideration of π-interactions, σ-hole interactions, transition metal properties, solvation energies, etc.). Additionally, we did not perform feature selection to optimize the performance of our model on the benchmark training sets to avoid potentially over-optimizing the models for the training data. For a detailed evaluation of the importance of each feature in BCL-AffinityNet and BCL-DockANNScore, please see the top 20 features by model input sensitivity (Tables S4 and S5) and decomposition of each descriptor into the average input sensitivity per sign pair (Figures S1−S8) in the Supporting Information.

Scoring Power Evaluation of BCL-AffinityNet. We trained BCL-AffinityNet on protein−ligand complexes from the PDBbind v.2016 refined set and all general set protein−ligand (small molecule) complexes for which binding constants were available. Protein−ligand pairs comprising the coreset (285 unique test-set complexes) were entirely excluded from training. BCL-AffinityNet was trained with descriptors of the form eq 2. See the Supporting Information for a sample feature code object file and command lines to generate the model.

We first tested the performance of BCL-AffinityNet on the scoring power task described in the comparative assessment of score functions 2016 update (CASF2016). This task evaluates affinity prediction across the PDBbind v.2016 coreset comprised of 285 protein−ligand pairs on 57 targets (5 small molecules per target) by measuring the Pearson correlation coefficient (R) between the predicted and experimental values. It has previously been noted that binding affinities in this task correlate strongly with both the fraction of buried solvent accessible surface area (ΔSAS, R = 0.63) (Figure 2A)[22] and several scalar ligand descriptors, including molecular weight (MW, R = 0.50), topological polar surface

https://dx.doi.org/10.1021/acs.jcim.0c01001
J. Chem. Inf. Model. 2021, 61, 603−620



	BCL::Affinity Net Training (5xCV)	BCL::Affinity Net Combined AD Test (n=1377)	Ligand-based DNN Training (5xCV)	Ligand-based DNN Combined AD Test (n=1377)	Pocket-based DNN Training (5xCV)	Pocket-based DNN Combined AD Test (n=1377)
Normalized Mean Absolute Error	0.71	0.67	0.71	1.08	0.72	1.36
Correlation Coefficient	0.68	0.72	0.67	0.28	0.66	0.15

Figure 3. Performance evaluation on the combined AD test set. A total of 1377 training samples were excluded from the initial training set of 7568 samples (see Methods for details). The remaining 6191 training samples were used to train BCL-AffinityNet (i.e., a single-task regression DNN with PLC features), a signed 3DA LB QSAR model, or a signed 3DA pocket-based QSAR model. The training was completed with fivefold random-split cross-validation. Columns and error bars represent the mean and standard deviation of normalized mean absolute error (NMAE) (blue) or Pearson correlation coefficient (red) across either the fivefold random-split cross-validations (training) or fivefold random splits of the combined AD test set (testing).

area (TPSA, R = 0.20), log P (R = 0.32), and polarizability (R = 0.52) (Table S1). An important measure of success is whether or not the affinity prediction method is capable of performing better than these simple metrics that are unaware of specific protein−ligand interactions.

BCL-AffinityNet is among the best algorithms on the scoring power task (R = 0.84) (Figure 2A,B). $\Delta_{Vina}RF_{20}$, which is a protein−ligand interaction score function that uses a random forest (RF) algorithm to predict an error correction term on the AutoDock Vina score, performed similarly on the original CASF2016 report (Figure 2A).[22] However, as reported previously,[22] the training set for $\Delta_{Vina}RF_{20}$ includes 140 of the coreset test complexes. Lu and colleagues re-evaluated the scoring power of $\Delta_{Vina}RF_{20}$ after retraining it without any of the coreset complexes and found that it is still performed better than ΔSAS but with worse scoring power than originally reported (R = 0.73).[45]

BCL-AffinityNet performs competitively with other machine learning models, such as the grid-based CNN K_{DEEP} (R = 0.82) and RF-Score (R = 0.80) (Table S6). We note that K_{DEEP} was evaluated on the 290 molecule version of the PDBbind coreset, not the canonical 285 molecule set. Moreover, in the absence of the underlying distributions, it is unclear if these results are statistically different; however, the effect sizes are similar.

Explicit Assessment of Dataset Bias on BCL-Affinity-Net Scoring Power Performance. It is increasingly well documented that strong machine learning model performance on QSAR tasks can be the result of dataset bias.[41,46,47] Indeed, Yang et al. found that atomic CNNs (ACNNs) trained solely on ligand or receptor pocket features performed just as well as ACNNs trained on protein−ligand complexes,[46] suggesting that the model was unable to leverage features relating to the protein−ligand interactions in a meaningful way. Therefore, we sought to determine the extent to which dataset biases may be inflating BCL-AffinityNet performance.

First, we trained a BCL-AffinityNet Y-scramble model, in which the result labels were shuffled between training examples. The Y-scramble model is a negative control, and as expected, we find virtually no correlation between predicted and experimental results on the coreset with this model (Figure S9).

Next, we generated LB and pocket-based QSAR models with the same architecture as BCL-AffinityNet. These models were trained with the 3DA descriptor equivalent of the PLC features. In an ideal dataset, ligand and protein pocket controls would have near-zero correlation to experimental results; however, consistent with the findings of Yang et al.,[46] the LB and pocket-based QSAR models each had correlation coefficients greater than 0.50 at 0.72 and 0.61, respectively (Figure S9).

To assess the impact of dataset bias on our PLC model performance for out-of-class predictions, we generated three new leave-class-out test-set splits based on the ligand, protein pocket, or combined ligand and protein pocket similarity to the PDBbind v.2016 coreset. Specifically, we generated a k-means (k = 75) applicability domain (AD) model from the 3DAs of the ligands, protein pockets, or combination of ligands and protein pockets of the PDBbind v.2016 coreset. Using each of these AD models, we removed training samples that were further from their nearest Kohonen map node than the furthest point of the PDBbind v.2016 coreset was from the AD model. Intuitively, the new test sets thus include only points that are outside the nominal descriptor space given by the PDBbind v.2016 coreset for ligands, protein pockets, or combination ligand−protein pockets. This has the effect of making the training set feature space more representative of the PDBbind v.2016 coreset feature space while simultaneously creating new test sets that are outside PDBbind v.2016 coreset feature space.

This resulted in the creation of a LB AD test set (n = 995), pocket AD test set (n = 379), and combined AD test set (n = 1377) (see Methods for additional details). We hypothesized that the LB QSAR model would perform poorly on the LB AD test set, that the pocket-based QSAR model would perform poorly on the pocket AD test, and that both models would perform poorly on the combined AD test set. We further hypothesized that if models trained on PLC descriptors are truly generalizable SB score functions, then their performance on all three test splits ought not to be significantly worse than their training random-split cross-validation metrics.

We found that the LB QSAR models performed worse on the LB AD test set (R = 0.28) than on the random-split training cross-validation sets (R = 0.67) (Figure S10).

https://dx.doi.org/10.1021/acs.jcim.0c01001
J. Chem. Inf. Model. 2021, 61, 603−620

Table 1. Performance Evaluation of Models Trained on PDBbind Refined Version 2016 Dataset on Unique Complexes in the CSAR-NRC HiQ Test Sets. Gray shading indicates features/methods developed in this manuscript. Green shading indicates scalar properties employed a controls. Blue shading indicates comparisons to previously reported results with other software.[c]

		Benchmark Test Set					
		CSAR NRC-HiQ set 1 (n=55)			CSAR NRC-HiQ set 2 (n=49)		
Descriptor Set	**Model Type**	Pearson R	Spearman ρ	RMSE	Pearson R	Spearman ρ	RMSE
BCL-AffinityNet	DNN 2x512-32	0.72	0.77	2.02	0.85	0.82	1.37
Molecular Weight	Scalar Property	0.29	0.32	N/A	0.45	0.32	N/A
TPSA	Scalar Property	0.03	0.12	N/A	-0.10	-0.10	N/A
LogP	Scalar Property	0.06	0.10	N/A	0.08	0.10	N/A
Polarizability	Scalar Property	0.41	0.44	N/A	0.60	0.50	N/A
KDEEP [a]	Grid-based CNN	0.72	-- [b]	2.09	0.65	-- [b]	1.92
RF-Score [a]	RF Docking Score	0.78	-- [b]	1.99	0.75	-- [b]	1.66

[a]As reported in Jiménez et al.[32] [b]Not reported. [c]Results reported as Pearson correlation coefficient (R), Spearman rank correlation coefficient (ρ), and root mean square error (RMSE). Note that the Spearman rank correlation here is across all targets in the coreset, while the "ranking power" metric is based on within-target ranking of molecule affinities.

Table 2. Performance Evaluation of Models Trained on PDBbind Refined Version 2016 Dataset Sans CSAR-NRC HiQ Complexes on All Complexes in the CSAR-NRC HiQ Test Sets. Gray shading indicates features/methods developed in this manuscript. Green shading indicates scalar properties employed a controls.[a]

		Benchmark Test Set					
		CSAR NRC-HiQ set 1 (n=176)			CSAR NRC-HiQ set 2 (n=167)		
Descriptor Set	**Model Type**	Pearson R	Spearman ρ	RMSE	Pearson R	Spearman ρ	RMSE
BCL-AffinityNet	DNN 2x512-32	0.75	0.75	1.32	0.74	0.73	1.36
Molecular Weight	Scalar Property	0.50	0.51	N/A	0.66	0.67	N/A
TPSA	Scalar Property	-0.03	0.08	N/A	0.28	0.27	N/A
LogP	Scalar Property	0.27	0.38	N/A	0.21	0.36	N/A
Polarizability	Scalar Property	0.56	0.59	N/A	0.68	0.69	N/A

[a]Results reported as Pearson correlation coefficient (R), Spearman rank correlation coefficient (ρ), and root mean square error (RMSE). Note that the Spearman rank correlation here is across all targets in the coreset, while the "ranking power" metric is based on within-target ranking of molecule affinities.

Similarly, the pocket-based QSAR model performed worse on the pocket AD test set (R = 0.33) than on the training splits (R = 0.63) (Figure S11). We also note a reduction in the performance of the pocket-based QSAR model on the LB AD test set relative to training (R = 0.51 vs 0.64, respectively), as well as a reduction in the performance of the LB QSAR model on the pocket AD test set relative to training (R = 0.54 vs 0.65, respectively) (Figures S10 and S11). On the combined AD test set, we observe the worst performance of the LB (R = 0.28) and pocket-based (R = 0.15) QSAR models (Figure 3).

In contrast, we observed that BCL-AffinityNet, when retrained to exclude each AD test set, consistently performs

https://dx.doi.org/10.1021/acs.jcim.0c01001
J. Chem. Inf. Model. 2021, 61, 603−620

well (R = 0.72, 0.75, and 0.72 for the LB, pocket-based, and combined AD test sets, respectively) despite the reduced training set size and coverage (Figures 3, S10, and S11).

To evaluate whether PLC descriptors are effective with other machine learning model types, we have utilized WEKA[48] to train a random forest version of BCL-AffinityNet (termed AffinityRF for ease) for evaluation on the PDBbind v.2016 coreset and the combined AD test split. AffinityRF achieves a good correlation (R = 0.79 and 0.70, respectively) on both tasks, suggesting that PLC descriptors may be suitable in multiple machine learning paradigms (Figure S12). Altogether, these results suggest that the PLC descriptors encode generalized representations of protein−ligand interactions.

Performance Evaluation on Subsets of the CSAR-NRC HiQ Test Sets. As additional independent tests, we evaluated the performance of our models on the CSAR-NRC HiQ test sets. For the purposes of a head-to-head comparison with two of the leading machine learning methods in the field, K_{DEEP} and RF-Score, we first compared our model to the 55 and 49 compounds of the CSAR-NRC HiQ test sets 1 and 2, respectively, which were previously evaluated for K_{DEEP} and RF-Score in Jiménez et al.[32] For this evaluation, we retrained our models with the PDBbind set as described previously, but we also excluded any of the 55 or 49 compounds found in the CSAR test set from training.

RF-Score performed the best on set 1 (R = 0.78, root mean square error (RMSE) = 1.99) with K_{DEEP} (R = 0.72, RMSE = 2.09) and BCL-AffinityNet (R = 0.72, ρ = 0.77, RMSE = 2.02) performing similarly to one another (Table 1). In contrast, BCL-AffinityNet is the top performing model (R = 0.85, ρ = 0.82, RMSE = 1.37) on set 2, followed by the RF-Score (R = 0.78, RMSE = 1.66) and K_{DEEP} (R = 0.65, RMSE = 1.92) (Table 1).

Next, in the interest of obtaining a more complete benchmark and facilitating future comparisons, we extended our evaluation of the CSAR-NRC HiQ test sets to the full molecule lists, which included 176 and 167 molecules in sets 1 and 2, respectively. Again, we retrained our models on the PDBbind set, excluding now either the 176 or 167 compounds in test set 1 or 2 in addition to the remaining molecules in the 285 compounds from the coreset. Performance of BCL-AffinityNet on set 1 (R = 0.75, ρ = 0.75, RMSE = 1.32) is very similar to performance on set 2 (R = 0.74, ρ = 0.73, RMSE = 1.36) (Table 2).

A summary of BCL-AffinityNet's scoring power performance on the PDBbind v.2016 coreset, the CSAR-NRC HiQ I and II subsets from Jiménez et al.,[32] and the full CSAR-NRC HiQ I and II sets are shown in Figure S13.

Ranking Power Performance Evaluation. The CASF2016 ranking power evaluation analyzes the ability of score functions to rank ligands targeting the same receptor. Among the methods originally compared in Su et al.,[22] BCL-AffinityNet (ρ = 0.69) places just after $\Delta_{Vina}RF_{20}$ (ρ = 0.75) (Figure 4). Again taking into consideration Lu et al. retraining $\Delta_{Vina}RF_{20}$ to exclude the 140 overlapped test-set compounds, $\Delta_{Vina}RF_{20}$ achieves a ranking power ρ = 0.63 compared to Δ_{Vina}XGB, which achieves a ranking power of ρ = 0.65.[45]

Altogether results on the scoring power and ranking power tests suggest that BCL-AffinityNet is competitive with state-of-the-art SB virtual screening methods for binding affinity prediction and affinity ranking.

Docking Power Performance Evaluation. Despite its success in the scoring and ranking power evaluations, BCL-AffinityNet is not ideally suited for decoy discrimination. This is because the training set for BCL-AffinityNet is composed entirely of native protein−ligand complexes. Thus, while BCL-AffinityNet could likely be used with an AD model generated in the same feature space to exclude clashed structures (by virtue of the lack of occupancy in the shortest distance bins,

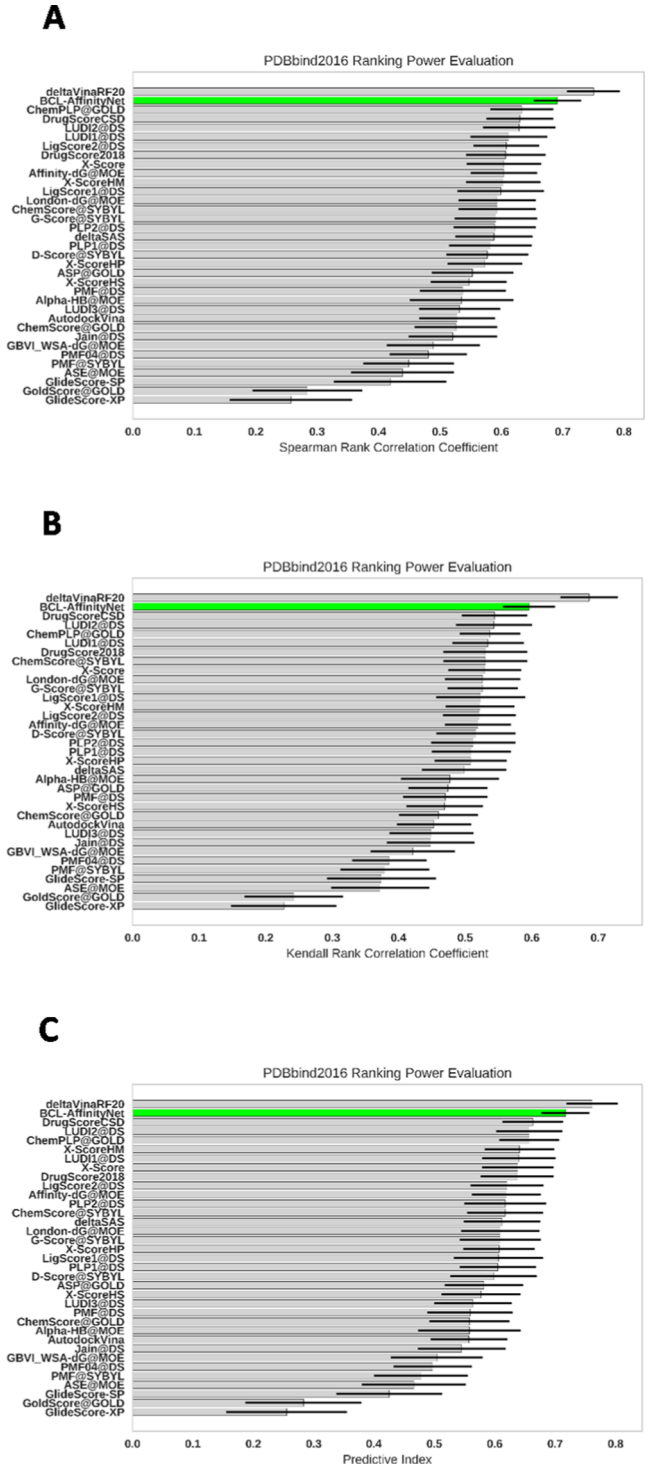


Figure 4. Ranking power evaluation of BCL-AffinityNet. Comparison of BCL-AffinityNet ranking power to other methods from the CASF2016 benchmark by Su et al.[22] with (A) Spearman rank correlation coefficient, (B) Kendall rank correlation coefficient, and (C) predictive index. Error bars indicate the 90% confidence interval. Green bars indicate BCL-AffinityNet.

https://dx.doi.org/10.1021/acs.jcim.0c01001
J. Chem. Inf. Model. 2021, 61, 603−620

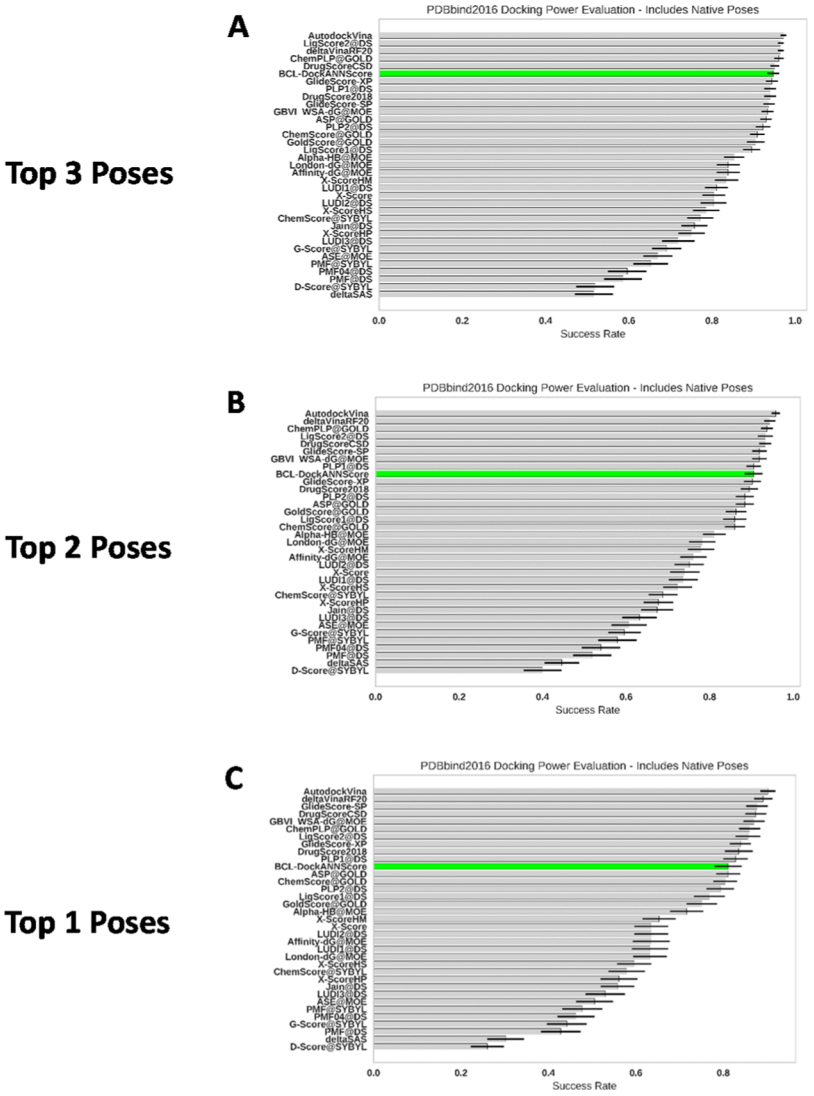


Figure 5. Docking power evaluation of BCL-DockANNScore. Comparison of BCL-DockANNScore docking power to other methods from the CASF2016 benchmark by Su et al.[22] when recovering the native pose under 2.0 Å RMSD (A) within the top 3 poses, (B) within the top 2 poses, and (C) within the top 1 poses. Error bars indicate the 90% confidence interval. Green indicates BCL-DockANNScore.

Figure 1A), it is unlikely to be able to discriminate plausible docking poses.

To address this limitation, we built a shallow multitasking ANN trained with the same PLC descriptors (eq 2) as BCL-AffinityNet, with the addition of the hydrogen bond angle descriptor described above (see the Supporting Information for a sample code object file). We reasoned that in differentiating properly docked poses, it would be insufficient to consider only hydrogen bond donor/acceptor distances. In our experience, we have also found that separating a categorical prediction task (i.e., is this pose most likely to be 1.0 Å from the native pose, 2.0, or 5.0 Å) into separate classification tasks for each category generally does not worsen model performance but may improve it. Thus, we nominally organized the output layer as five correlated classification tasks: determining whether a pose was less than 1.0, 2.0, 3.0, 5.0, and 8.0 Å from the native pose.

We trained this ANN on the PDBbind v.2016 refined set, excluding all coreset protein−ligand complexes. For each complex in the training set, 250 additional decoys were generated with RosettaLigand (see Methods for details). The final model score, which we refer to as BCL-DockANNScore, is the product of the classification probability of a pose being less than 2.0 Å from the native pose (referred to elsewhere as a probability calibration curve[49,50]) and the BCL-AffinityNet affinity prediction score for that pose.

BCL-DockANNScore performs reasonably well on the docking power benchmark with success rates of 0.81, 0.91, and 0.95 for native pose recovery at a 2.0 Å threshold for poses within the best scoring 1, 2, and 3 poses, respectively (Figure 5). When native poses are excluded, BCL-DockANNScore success rates reduce by ∼5%, consistent with performance reductions in multiple other methods (Figure S15). Binding funnel analysis of BCL-DockANNScore demonstrates good Spearman rank correlation coefficients at wide RMSD ranges but performs less well in the 0−2.0 Å range (Figure S16). This suggests that one possible route to improve BCL-DockANN-Score further is to provide additional training decoys within the 0−2.0 Å range or additional high-resolution descriptors.

Overall, these results are especially encouraging because the decoy poses for the docking power benchmark are generated with a combination of GOLD[51] version 5.2, SYBYL's Surflex,[52]

https://dx.doi.org/10.1021/acs.jcim.0c01001
J. Chem. Inf. Model. 2021, 61, 603−620

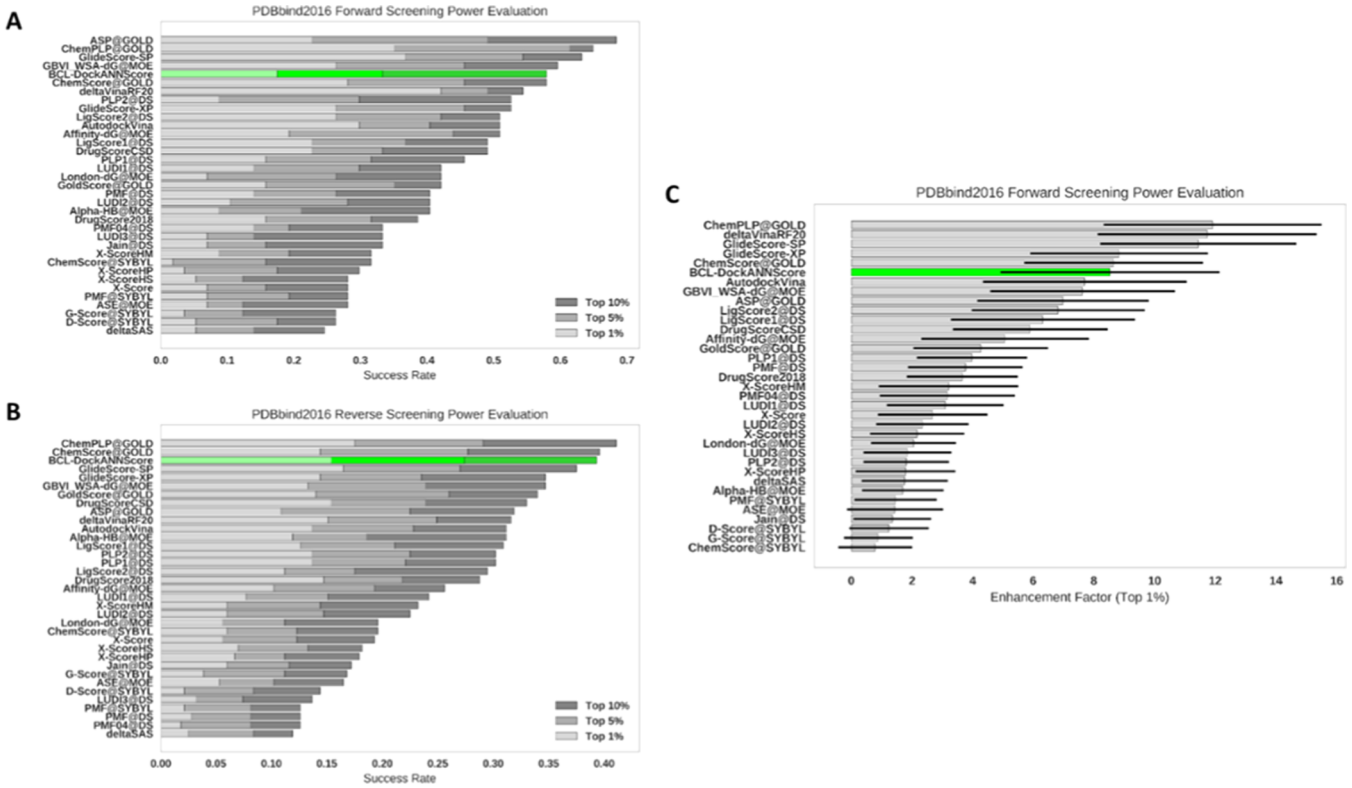


Figure 6. Screening power evaluation of BCL-DockANNScore. Comparison of BCL-DockANNScore screening power to other methods from the CASF2016 benchmark by Su et al.[22] (A) Forward screening power evaluation success rates, (B) reverse screening power evaluation success rates, (C) forward screening power evaluation enhancement factor (top 1%). Error bars indicate the 90% confidence interval. Green indicates BCL-DockANNScore.

and Chemical Computing Group's Molecular Operating Environment (MOE) docking algorithm,[53] while our training decoys were generated with RosettaLigand. It suggests that the BCL-DockANNScore feature space (i.e., PLC descriptor space) is not overly dependent on RosettaLigand poses and can be used to build modular score functions.

Screening Power Performance Evaluation. We evaluated BCL-DockANNScore on the forward and reverse screening tests. The forward screening power task evaluates the ability of a score function to identify small molecule ligands that bind to a target protein. The reverse screening power task evaluates the ability of a score function to identify the protein that most effectively binds a small molecule ligand (i.e., cross-docking).[22]

Similar to the docking power evaluation, we find that BCL-DockANNScore performs reasonably well, but not always among the very best docking scores. On the forward screening task, BCL-DockANNScore has a success rate of 0.18, 0.33, and 0.58 when identifying the ligand amongst the top 1, 5, and 10% of candidates, respectfully (Figure 6A). This is competitive with the best score functions at the 10% level; however, performance at the 1% level is more mid-tier (ranking alongside several of the MOE score functions, while the top performers are from GOLD, Glide, and the AutoDock Vina and derived methods). The overall enhancement factor at the 1% level is 8.5 (Figure 6C). In contrast, we find that the performance on the reverse screening task is competitive even with the top performers when identifying the top 1, 5, and 10% of candidates, with success rates of 0.15, 0.24, and 0.39, respectively (Figure 6B).

Generating Absolute Pharmacophore Maps. Finally, one important consideration in the development of a SB score function for the BCL was model interpretability. One of the strengths of SB CADD is that the predicted changes in activity can be attributed to specific interactions with the target. Neural networks are, however, often negatively characterized as "black boxes" because usually the function learned in the model cannot be decomposed into human-interpretable parts. Traditional docking scoring functions, such as RosettaLigand, have the advantage that they can be decomposed into target per-residue contributions to the overall predicted affinity. This is important in drug discovery, where predictions need to be actionable. Here, we demonstrate that BCL-AffinityNet predictions can be decomposed into a map of atom contributions to the predicted bioactivity.

We take two general approaches for constructing a pharmacophore map: (1) absolute feature contributions (Figure 7) and (2) relative feature contributions (Figure 8). The first case generates a map on any individual molecule by evaluating the contributions of specific atoms to the overall predicted activity. This can be likened to evaluating model input sensitivity, except in this case, the molecule of interest is being perturbed instead of the weights connecting individual neurons in the model.

To generate an absolute pharmacophore map of a given molecule, we perturb the chemical structure by sequentially removing individual atoms and closing the newly opened valence(s) with hydrogen atoms. Afterward, we compute the predicted affinity for each perturbed molecule with BCL-AffinityNet. The predicted binding affinity of the perturbed molecules is compared to that of the original molecule. The

https://dx.doi.org/10.1021/acs.jcim.0c01001
J. Chem. Inf. Model. 2021, 61, 603−620

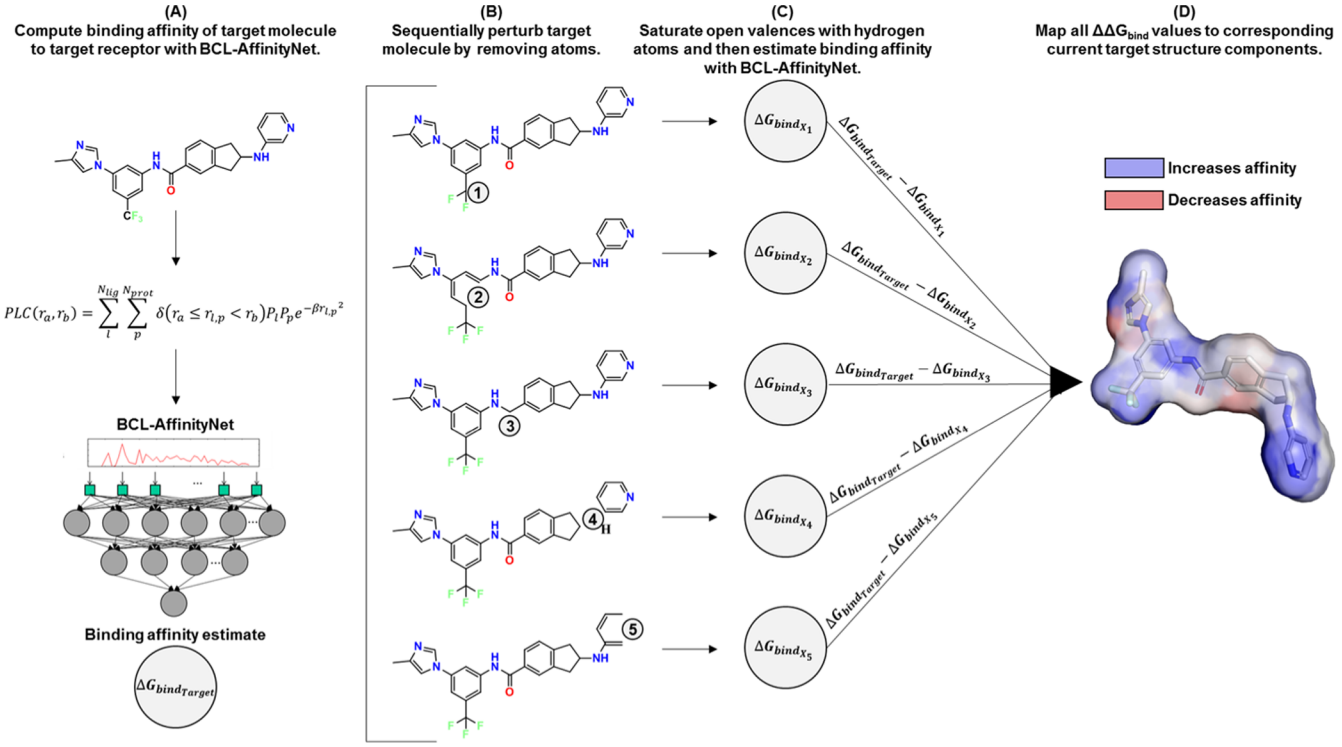


Figure 7. Construction of absolute pharmacophore maps. (A) The target molecule, in this case, compound 7c from Zhu et al.,[54] is first modeled in complex with its target receptor using PLC descriptors and scored with BCL-AffinityNet. (B) Then, we iterate over each atom in the target molecule and sequentially remove it from the molecule to create a perturbed molecule, X. (C) Perturbed molecules are saturated with hydrogen atoms to close any open valences resulting from the perturbation, and then they are scored with BCL-AffinityNet. (D) Differences in predicted binding affinity between the starting molecule and each perturbed molecule are mapped to the corresponding atoms of the starting structure. Here, predictions are in units of kcal/mol at 300 K. The surface representation of atoms that contribute beneficially to BCL-AffinityNet's binding affinity prediction is blue, while atoms that worsen the prediction are in red. Atoms that contribute neutrally/negligibly are white.

differences in predicted activity between the perturbed and original molecules are assigned to the corresponding atoms (Figure 7).

Generating Relative Pharmacophore Maps. Relative pharmacophore maps leverage structural similarity within a congeneric ligand series to attribute predicted affinity differences to specific substructures. It has been shown that highly accurate binding affinity estimates can be obtained with alchemical free energy methods when reference structures with experimentally determined binding affinities within a congeneric ligand series are available.[19−21]

To generate a relative pharmacophore map between two molecules, we first identify a common substructure (MCS) via one of two methods: (1) identify the largest subgraph isomorphism between the two molecules or (2) assign spatially mutually matched atoms to be common to one another (the first approach is more accurate and is the default approach). Component substructures that graphically correspond to the same common atoms are then iteratively removed, newly opened valences are closed with hydrogen atoms, and the perturbed molecules are scored with BCL-AffinityNet (Figure 8).

Thus, for each non-MCS substructure in the reference and target molecules, there is a $\Delta\Delta G_{bind}$ between the nonperturbed and perturbed molecules. A final $\Delta\Delta\Delta G_{bind}$ is computed for each non-MCS substructure as the difference between the reference and target perturbation $\Delta\Delta G_{bind}$ values (Figure 8). The $\Delta\Delta\Delta G_{bind}$ values are mapped to the target molecule for visualization.

Consider a series of type II tyrosine kinase inhibitors (TKIs) of DDR1 kinase developed recently by Zhu et al.[54] We generated relative pharmacophore maps of compounds 7c, 7f, and 7j to compound 7i[54] (Figure 9A−D). We also modified the compound 7 scaffold to include N → C mutations in the hinge-binding region analogous to prior substitutions done by Wang et al.[55] in a previous DDR1 TKI series (Figure 9A,E−G).

From the pharmacophore maps, we also compute relative binding affinities of each molecule to compound 7i by summing the $\Delta\Delta\Delta G_{bind}$ values for each non-MCS component in the target molecule: $\Delta\Delta G_{bind} = \sum\Delta\Delta\Delta G_{bind}$. In all comparisons, the trifluoromethyl group is preferable to the methyl. Relative binding affinity estimates of compounds 7c and 7f from 7i are within 0.50 kcal/mol of experimental values (−2.62 vs −2.82 kcal/mol and −2.32 vs −2.25 kcal/mol, respectively) (Figure 9A,C,D). The ethyl in 7j is also correctly estimated to improve binding affinity relative to methyl in 7i; however, BCL-AffinityNet underestimates the extent of the affinity improvement (−0.69 vs −2.25 kcal/mol) (Figure 9A,B). Conversion of both hinge-binding nitrogen atoms to carbon atoms is strongly unfavorable even in the presence of the trifluoromethyl group, consistent with prior SAR[55] (Figure 9A,G). Thus, the relative pharmacophore maps provide meaningful QSAR insights that can be readily visualized.

Relative pharmacophore maps can be generated with respect to one or more reference input molecules (e.g., hit compounds or scaffolds) or in a pairwise manner across a series of input molecules. If more than one molecule is used as a reference,

https://dx.doi.org/10.1021/acs.jcim.0c01001
J. Chem. Inf. Model. 2021, 61, 603−620

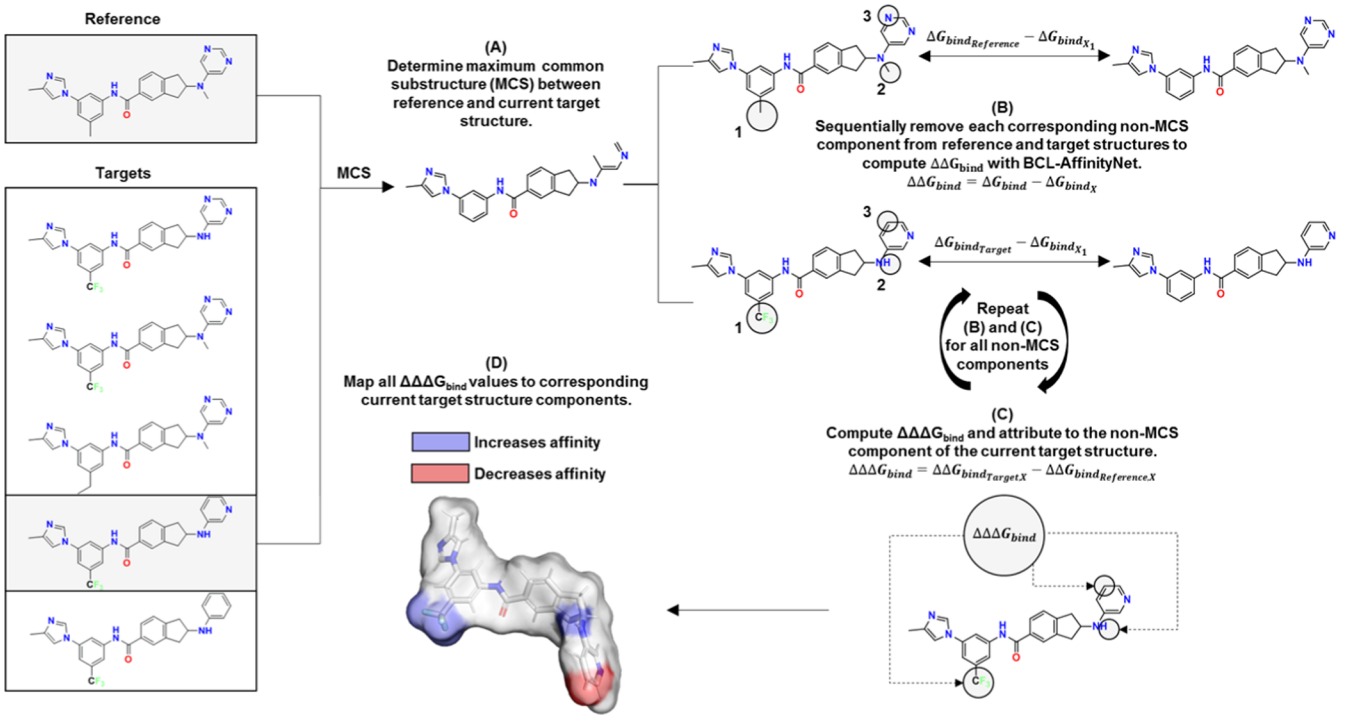


Figure 8. Construction of relative pharmacophore maps. Relative pharmacophore maps are generated from a target molecule and a reference molecule. (A) Determine the MCS between the reference and target structure. (B) Identify the MCS atoms that connect to the corresponding non-MCS substructures in both the reference and target molecules. Non-MCS atoms are circled in gray, and the corresponding substructures between the reference and target share numerical labels (e.g., the reference molecule methyl circled in gray and the target molecule trifluoromethyl circled in gray correspond structurally and are labeled "1"). For both the reference and target molecules, non-MCS substructures are independently removed. The binding affinities of the reference, target, and perturbed molecules are estimated with BCL-AffinityNet. The $\Delta\Delta G_{bind}$ between starting and perturbed molecules is determined for both the reference and target. (C) For each corresponding non-MCS substructure, compute $\Delta\Delta\Delta G_{bind}$ as $\Delta\Delta\Delta G_{bind} = \Delta\Delta G_{bind_{Target,X}} - \Delta\Delta G_{bind_{Reference,X}}$, where X indicates the perturbed target or the reference molecule. (D) Map the $\Delta\Delta\Delta G_{bind}$ values back to the target molecule non-MCS substructures. The surface representation of atoms that contribute beneficially to BCL-AffinityNet's binding affinity prediction is blue, while atoms that worsen the prediction are in red. Atoms that contribute neutrally/negligibly are white.

the final map for each target molecule indicates the favorability of each molecule's substitutions in comparison to the whole ensemble. For example, command-line to generate a relative pharmacophore map, see the Supporting Information.

Case Study on Guiding Chemical Modifications with Pharmacophore Maps. To further illustrate this approach, consider three congeneric dysiherbaine analogues in complex with ionotropic glutamate receptor 5 (iGluR5). These molecules differ from one another by small substitutions at carbon atoms (1) and (2) (Figure 10A−C, first row). Each of the analogues was scored with BCL-AffinityNet and ranked correctly. For each of these three compounds, we generated absolute and relative pharmacophore maps (see Methods for command-line details).

First, we generated relative pharmacophore maps of the dysiherbaine analogues in the pairwise manner described above (Figure 8). The pharmacophore maps of dysiherbaine and neodysiherbaine suggest that the methylamine and hydroxyl substitutions, respectively, at position (2) provide a net increase in affinity relative to the proton in 8,9-dideoxyneodysiherbaine (Figure 10A−C, third row). Furthermore, the pharmacophore maps predict that the methylamine modification increases binding affinity more than the hydroxyl substitution, in agreement with experimental observation (Figure 10B,C, third row).

Interestingly, the relative pharmacophore map of neodysiherbaine also predicts that the hydroxyl substitution at position (2) is more important for binding affinity than the hydroxyl substitution at position (1) (Figure 10B, third row). Similarly, the methylamine at position (2) of dysiherbaine is predicted to contribute more to the binding affinity than the hydroxyl at position (1) (Figure 10C, third row). Finally, we see from the absolute pharmacophore maps of all three analogues that the two carboxylic acid groups contribute favorably to the binding. Indeed, we see that their contributions are predicted to be more important than the substitutions at (1) and (2), supporting the notion that these substituents are an important component of the conserved scaffold (Figure 10A−C, fourth row).

Together with the DDR1 TKI congeneric series, these comparisons illustrate how BCL-AffinityNet can yield structure−activity insight. To our knowledge, this is the first modern machine learning-based SB score function that is readily accompanied by an interpretable decomposition scheme. In principle, our pharmacophore mapping procedure is compatible with any LB or SB machine learning score function in the BCL. Thus, these results demonstrate a fast and simple approach to generate interpretable pharmacophore maps from BCL machine learning model predictions.

DISCUSSION

Here, we develop a novel machine learning-based score function for vHTS SB scoring. Our approach centers around the development of novel protein−ligand signed property

https://dx.doi.org/10.1021/acs.jcim.0c01001
J. Chem. Inf. Model. 2021, 61, 603−620

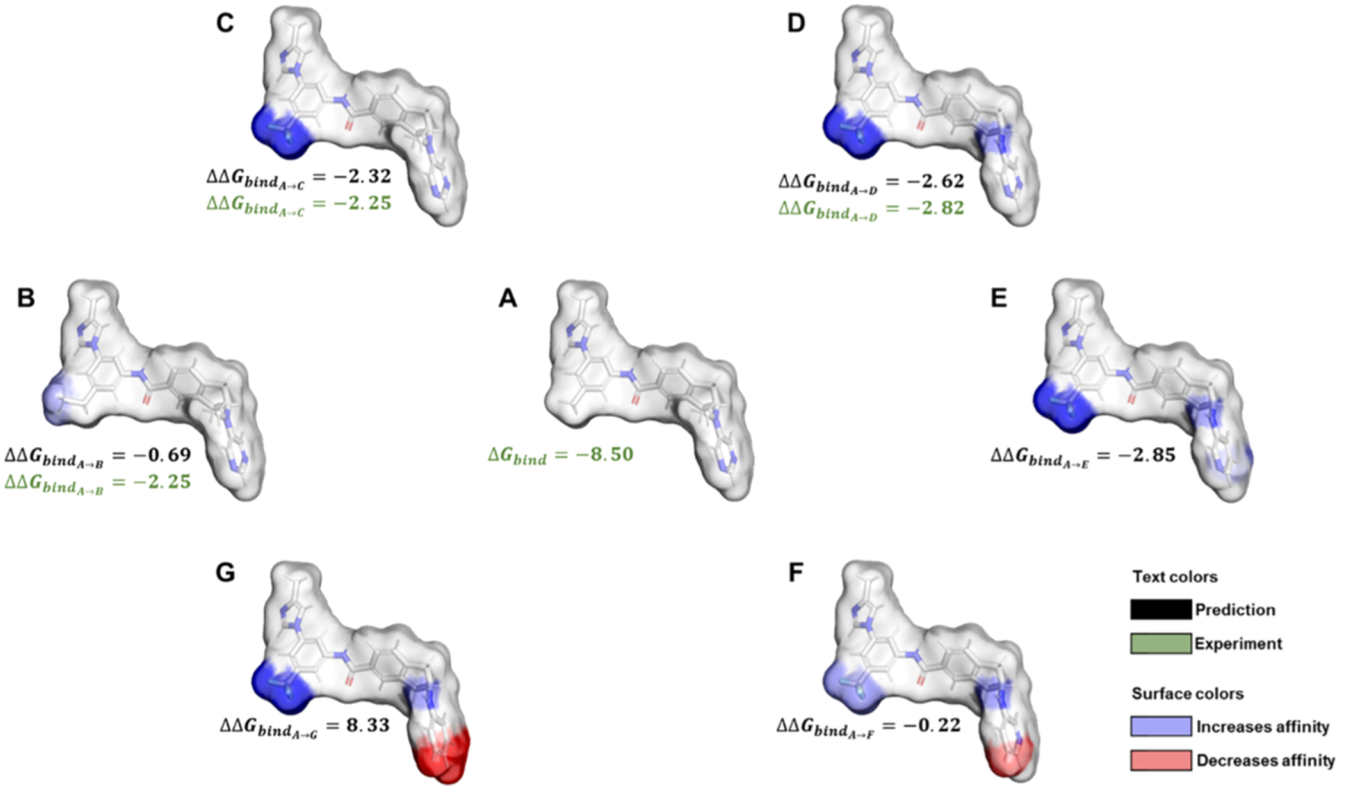


Figure 9. Relative pharmacophore maps of a congeneric DDR1 inhibitor series. (A) Compound 7i is the reference molecule for the creation of the pharmacophore maps. Compounds (B) 7j, (C) 7f, and (D) 7c from Zhu et al.[54] Compounds with the N → C alteration at (F) the hinge-binding nitrogen atom, (E) the symmetrically placed hinge-binding nitrogen rotated away from the hydrogen bond donor partner, and (G) both nitrogen atom positions at the hinge-binding ring. Binding affinities in black text are predicted by BCL-AffinityNet, while green values are from Zhu et al.[54] The surface representation of atoms that contribute beneficially to BCL-AffinityNet's binding affinity prediction is blue, while atoms that worsen the prediction are in red. Atoms that contribute neutrally/negligibly are white.

correlation descriptors. In addition to the new descriptors, our models avoid the use of ligand-specific features to reduce the risk of training dataset bias. The new models, BCL-AffinityNet and BCL-DockANNScore, have been evaluated on current best practice benchmarks and compared to other standard and leading methods.

BCL-AffinityNet generally performs on par with or better than currently available SB virtual screening scores in affinity prediction and affinity ranking. BCL-DockANNScore, while generally not as good as GOLD, Glide, or the AutoDock Vina and derived methods at pose recovery or screening, performs competitively with respect to all of the evaluated methods. We therefore suggest that it may be a generally useful SB scoring algorithm with especially strong affinity prediction. Indeed, some of the best methods for docking and screening failed to provide estimates for power scoring (e.g., statistics for GlideScore-XP are based on 258/285 protein−ligand pairs, GlideScore-SP 252/285, GoldScore@GOLD 244/285).[22] Thus, when considering all of the tasks together (scoring power, ranking power, docking power, and screening power), the new SB scoring models in the BCL demonstrate the utility of our novel signed property protein−ligand correlation descriptors for SB CADD. Moreover, BCL-AffinityNet and BCL-DockANNScore represent the first instantiation of SB scoring in the BCL.

While a number of algorithms consider multiple ligand-specific descriptors in their feature space alongside the protein−ligand interaction features (e.g., AutoDock Vina incorporates, e.g., the ligand length, number of hydrophobic atoms, etc.; both $\Delta_{Vina}RF_{20}$ and $\Delta_{Vina}XGB$ include ligand-specific pharmacophore features; $\Delta_{Vina}XGB$ includes an estimate of ligand conformational stability; K_{DEEP} contains ligand-specific voxels colored by pharmacophore features),[22,29,32,45] we made a conscious decision to avoid inclusion of such features in BCL-AffinityNet and BCL-DockANNScore. This was done to reduce the ligand bias of the models and hopefully yield a more generalizable score function. Nevertheless, efforts are underway to incorporate other aspects of protein−ligand binding affinity other than just interaction score terms into the BCL-AffinityNet and BCL-DockANNScore in an unbiased manner. These include improvements to both the neural network architectures employed here as well as the incorporation of efficient metrics for solvation energy, ligand conformational preference, and entropy changes.

An important limitation of our work is that all models were trained in the absence of explicit water molecules, metal ions, and/or other cofactors. Others have recently demonstrated that the incorporation of explicit water molecules can improve model performance,[45] and future improvements to our model will incorporate these elements. As these updates are introduced, we will also continue to retrain the models leveraging the increasing availability of high-quality protein−ligand co-crystal structures with K_i/K_d data.

Another limitation is the under-optimized protein−ligand interaction feature space of the current models. The generalizability of the PLC descriptors used to build BCL-AffinityNet and BCL-DockANNScore should not be conflated with

https://dx.doi.org/10.1021/acs.jcim.0c01001
J. Chem. Inf. Model. 2021, 61, 603−620

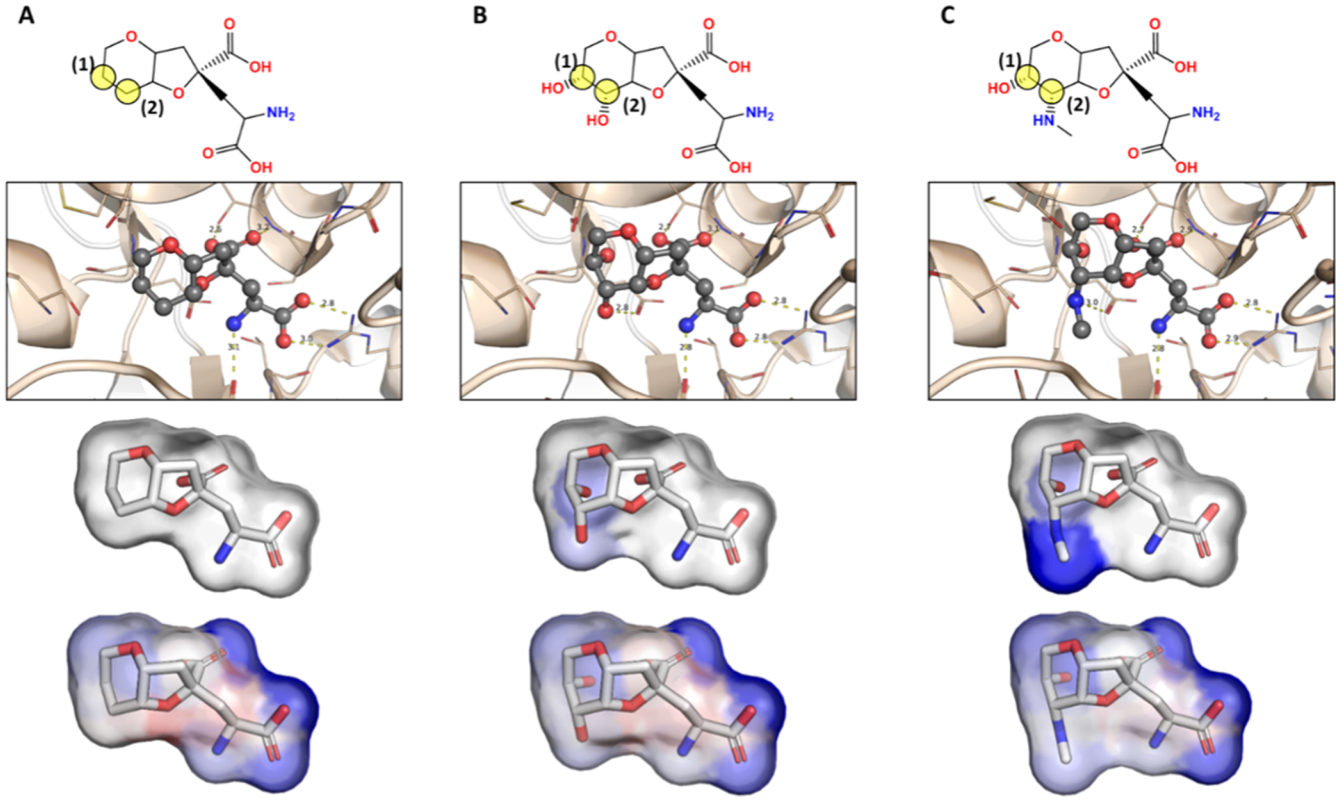


Figure 10. Pharmacophore maps of dysiherbaine analogues in complex with iGluR5 generated from BCL-AffinityNet. Pharmacophore maps were generated for iGluR5 complexed with (A) 8,9-dideoxyneodysiherbaine (PDB ID 3GBB; pK_d = 6.9, ΔG = −9.79 kcal/mol at 310 K), (B) neodysiherbaine (PDB ID 3FV2; pK_d = 8.1, ΔG = −11.49 kcal/mol at 310 K), and (C) dysiherbaine (PDB ID 3FV1; pK_d = 9.3, ΔG = −13.19 kcal/mol at 310 K) and mapped onto the native bound pose. Labeled yellow transparent circles in the top panel are used to reference the substituted carbon atoms of interest. Per atom pharmacophore map scores are output to a PyMol script for visualization as a molecular surface colored on a per atom basis by spectrum from blue (negative) to white (zero) to red (positive). In this example, negative values indicate atoms whose removal results in a loss in predicted binding affinity. The second row illustrates each ligand in complex with iGluR5. The third row illustrates the common substructure pharmacophore map (i.e., pairwise per-substructure relative binding free energy changes). The fourth row illustrates the raw pharmacophore map for each ligand upon sequentially removing individual atoms and saturating open valences.

completeness of the score function. By analogy, RosettaLigand with the Rosetta Talaris2014 score function[56] does not model halogen σ-hole interactions with aromatic ring systems and is thus unlikely to accurately determine the protein−ligand binding affinities of systems with these interactions. In the same way, BCL-AffinityNet and BCL-DockANNScore are incomplete representations of protein−ligand interactions. Further score function development will focus on expanding the availability of training data as well as describing additional salient chemical features.

Ongoing work in the Meiler Lab is focused on the development of both LB and SB small molecule de novo design and focused library design algorithms. A critical motivator for the present work was the need for the BCL to have a rapid and flexible SB score function that can be deployed for design tasks where there is insufficient data to build a reliable LB QSAR model. BCL-AffinityNet and BCL-DockANNScore are fully integrated into the BCL descriptor framework, allowing them to be called and mathematically combined with a multitude of other features, including AD scores, ligand descriptors, and more.

Another fundamental hurdle that we wanted to overcome was the so-called "black box" problem. This problem arises whenever the underlying feature space of the score function cannot be decomposed into human-interpretable parts, and it presents a major challenge when relying on complex score functions for rational drug design. In this manuscript, we have demonstrated a simple approach that can be employed with any score function in the BCL (machine learning or not) to convert predictions into all-atom pharmacophore maps. These pharmacophore maps can be generated with respect to underlying substructures or spatially matched atoms between different molecules, or they can be generated for individual molecules without a reference structure. We demonstrate how this can be accomplished with the BCL-AffinityNet score function for a series of congeneric DDR1 TKIs and dysiherbaine analogues. The relative pharmacophore maps provide an interpretable decomposition of affinity with respect to scaffold modifications that can be used to guide further molecule optimization. The absolute pharmacophore map procedure can tell the user which atoms are most salient to BCL-AffinityNet's predictions. In addition to being a useful tool for interpreting machine learning score functions in the BCL, we anticipate that such pharmacophore maps will be valuable in automated drug design tasks.

All of our models and applications for generating new models are freely available with an academic license for the BCL at http://meilerlab.org/. We hope that our descriptors and models may be integrated with future machine learning architecture development and descriptor optimization for the continued advancement of drug discovery.

https://dx.doi.org/10.1021/acs.jcim.0c01001
J. Chem. Inf. Model. 2021, 61, 603−620

METHODS

Training Dataset Preparation. BCL-AffinityNet was trained using the refined set plus protein−ligand complexes from the general set of the PDBbind v.2016 dataset that satisfied the following criteria: (1) the ligand was a small molecule; (2) the binding affinity was measured as either K_i or K_d; and (3) all-atom types had defined Gasteiger atom types. The PDBbind v.2016 coreset was not included in the training set for any of the models for any of the performance evaluations. BCL-DockANNScore was trained using the refined set protein−ligand complexes from the PDBbind v.2016 dataset, excluding the 285 coreset compounds. For each protein−ligand complex in the PDBbind v.2016 refined set, 250 additional pose decoys were generated with RosettaLigand flexible docking with the Talaris2014 score function.[57−59]

Model Validation. Metrics for scoring power, ranking power, docking power, screening power, and confidence interval bootstrapping were performed with the scripts made available with download of PDBbind v.2016.[22] All models were trained with fivefold random-split cross-validation. The final model prediction value is the average prediction value obtained across all five splits (i.e., as opposed to selecting a single best model from the five splits). PDBbind v.2016 coreset complexes were always excluded from training. For other external test-set evaluations, the models were always retrained, excluding all test-set complexes explicitly. Thus, the final training set sizes for testing on the PDBbind 2016 coreset (n = 285), CSAR-NRC HiQ 1 Jiménez et al. subset (n = 55), CSAR-NRC HiQ 2 Jiménez et al. subset (n = 49), CSAR-NRC HiQ 1 full set (n = 176), and CSAR-NRC HiQ 2 full set (n = 167) were 7568, 7551, 7537, 7442, and 7440 (not every complex in the CSAR sets is in the PDBbind v.2016 set, hence the differences are not equivalent to 7568 − n). For comparisons to the CSAR-NRC HiQ benchmarks in Jiménez et al.,[32] complexes present in both the CSAR test sets and the PDBbind v.2016 refined subset were removed from the CSAR test sets. This resulted in two CSAR test sets of sizes 55 and 49, respectively, with the exact same PDB IDs as reported in the Supporting Information of Jiménez et al.[32]

For our baseline assessment of ligand and receptor pocket bias on the PDBbind v.2016 coreset, we trained two DNNs identical in architecture to BCL-AffinityNet. For descriptors, we utilized the same chemical features, distance bins, and sign pairings as in the PLC descriptors, except we instead generated signed 3D autocorrelations of the ligand and/or receptor itself.[14] As inputs, we used the structures provided in the PDBbind v.2016 dataset such that the ligand-based DNNs were trained on the native poses of the ligands and the pocket-based DNNs were trained on the receptor binding pockets as extracted for inclusion in PDBbind v.2016.[22,60]

For validation splits that explicitly address ligand and pocket bias of the training datasets, we generated k-means (k = 75) AD models of the PDBbind v.2016 coreset (n = 285) based on ligand 3DAs, pocket 3DAs, or column-combined ligand and pocket 3DAs (using the same descriptors that were used to create ligand- and pocket-based QSAR models; see the Supporting Information). We then scored all 7568 training set samples with each of these AD models. Previous studies on appropriate cutoffs for distance-based AD models have suggested that test-set samples further away from their closest node than 95−100% of the training samples can reliably be considered outside of the domain of applicability.[34,61] We therefore made three test-set splits (one for each AD model) containing all training samples that had AD scores greater than 1.0. The resulting test sets are those samples whose ligands, proteins, or ligands and proteins can be considered within the same AD as the PDBbind v.2016 coreset. Put another way, this creates larger PDBbind v.2016 coreset-like leave-class-out test-set splits based on the properties of the ligands, protein pockets, or combined ligands and protein pockets. We refer to these test sets, respectively, as LB AD test (n = 995), pocket AD test (n = 379), and combined AD test (n = 1377). For these evaluations, the total model training sample size is 7568 − n. For details on command-line syntax, see the Supporting Information.

Training Neural Networks for Affinity Prediction and Pose Discrimination. All neural networks were trained with the BCL. Our binding affinity prediction model, which we call BCL-AffinityNet, is a single-task, feed-forward regression neural network trained to predict p$K_{i/d}$. While technically a "deep" neural network in that we utilize two hidden layers (512 and 32 neurons, respectively) instead of just one, BCL-AffinityNet is quite small compared to neural networks recently published for similar tasks.[31−33] Our pose prediction model, which we call BCL-DockANNScore, is a shallow (single hidden layer, 32 neurons) multitasking feed-forward classification neural network that predicts whether a protein−ligand pose is less than 1.0, 2.0, 3.0, 5.0, and 8.0 Å from the correct pose. Both networks can thus be formalized as follows.

For a network with L hidden layers indexed $l \in (1, ..., L)$, forward propagation for $l \in (0, ..., L - 1)$ can be described as

$$z^{(l+1)} = w^{(l+1)} y^{l} + b^{(l+1)} \tag{3}$$

$$y^{(l+1)} = f(z^{(l+1)}) \tag{4}$$

where $\boldsymbol{y}^l$ is the output vector at layer l connected to the input vector $z^{(l+1)}$ at layer $l + 1$ by weights $\boldsymbol{w}$ and biases $\boldsymbol{b}$ and f is the transfer function applied to each set of inputs into the $l + 1$ layer. Correspondingly, the activation of a single neuron i in hidden layer $l + 1$ can be represented as

$$z_i^{(l+1)} = w_i^{(l+1)} y^{l} + b_i^{(l+1)} \tag{5}$$

$$y_i^{(l+1)} = f(z_i^{(l+1)}) \tag{6}$$

to yield the output $y_i^{(l+1)}$ from layer $l + 1$. A mean-squared error (MSE) cost function was employed in all studies. Overtraining is prevented through the use of dropout in the input and hidden layers. During forward propagation, each output value y_i^l of each i neuron in the layer l of the ANN is randomly multiplied either by a value of 0 (corresponding to a "dropped" neuron) or 1.

$$z_i^{(l+1)} = w^{(l+1)}(r^{l} \times y^{l}) + b_i^{(l+1)} \tag{7}$$

Here, $\boldsymbol{r}^l$ is a vector with the same dimensions as $\boldsymbol{y}^l$ whose values are either 0 (at fraction p) or 1 (at fraction $1 - p$). At the end of every training batch, $\boldsymbol{r}^l$ is shuffled. At test time, the corresponding weights are scaled down by the factor $1 - p$.

The BCL-AffinityNet DNN contains two hidden layers with 512 and 32 neurons, respectively. It was trained with a 5% dropout in the input layer, a 25% dropout in the first hidden layer, and a 5% dropout in the second hidden layer.[5] All neurons utilized a leaky rectifier transfer function

https://dx.doi.org/10.1021/acs.jcim.0c01001
J. Chem. Inf. Model. 2021, 61, 603−620

$$f(x) = \begin{cases} x & x > 0 \\ 0.05x & x \leq 0 \end{cases} \quad (8)$$

where x is the total input to a neuron. We utilized normalized mean absolute error (NMAE; defined as the quotient of mean absolute error and mean absolute deviation) as our objective function during training.

The BCL-DockANNScore ANN contained a single hidden layer with 32 neurons. It was trained with 5% dropout in the input layer and 25% dropout in the hidden layer.[5] All neurons utilized a sigmoid transfer function

$$f(x) = \frac{e^x}{e^x + 1} \quad (9)$$

where x is the total input to a neuron. We utilized the area under the curve (AUC) as our objective function during training.

The AffinityRF random forest model was trained with WEKA v.3.8.4 utilizing default settings.[48]

Feature Parameter and Neural Network Hyperparameter Tuning. Our adoption of 5% input layer dropout and 25% dropout in the first hidden layer (for both models) as well as the selection of a 32 neuron hidden layer prior to the output layer is based on extensive prior evaluation in Mendenhall et al.[5] For classification models, it has been shown that shallow networks often perform equivalently and sometimes better than deep networks at a substantially reduced training cost.[4] This, coupled with our own experience with QSAR classification tasks,[5,17,18] led us to use our previously utilized single hidden layer architecture for BCL-DockANNScore.[5]

With respect to BCL-AffinityNet, we nominally selected the nearest power of 2 ($2^9 = 512$) to our input feature size as an upper limit for our first hidden layer size. We investigated two PLC descriptor feature parameters using fivefold random-split cross-validation with the DNN of this size: (1) the interaction bin distance and (2) the smoothing parameter β (eq 2). We selected an initial smoothing parameter value of 5.0 based on prior 3DA QSAR investigations in which values greater than one were effective.[16−18] Subsequently, we varied the interaction bin distances at 1.0 Å intervals between 4.0 and 9.0 Å and compared NMAE and Pearson correlation across the cross-validation splits. Our results suggested that distances greater than 5.0 Å were best (Figure S17). In the interest of keeping our feature set relatively small, we selected 7.0 Å for our final models. Similarly, we varied the smoothing parameter between 0.1 and 10.0 at a fixed bin distance of 7.0 Å. We found that β values between 3.0 and 10.0 produced similar results (Figure S18); therefore, we retained a value of 5.0 for all additional studies.

With the PLC parameters selected, we then performed additional fivefold random-split cross-validation studies to determine an appropriate first hidden layer size. We decreased the number of neurons from 512 by powers of 2 down to the size of the second hidden layer (32 neurons). For completeness, we also evaluated a shallow ANN ranging in size from 32 to 128 neurons using either a leaky rectifier (eq 8) or sigmoid (eq 9) transfer function. Generally, we observed that shallow and deep networks with smaller (32−64 neurons) first hidden layers performed the worst independent of transfer function. We also noted that two hidden layers seemed better than one, with little improvement in cross-validation performance between 256 and 512 neurons (Figure S19).

We note that all cross-validation studies for PLC feature parameter and model hyperparameter tuning were done with the BCL-AffinityNet training set of size 7568 protein−ligand complexes (PDBbind v.2016 refined set, excluding the coreset and including select general set complexes; see the Methods subsection Model Validation for details). Model performance on the external test sets was not evaluated during feature parameter or model hyperparameter tuning.

Resolving Hydrogen Bond Angles in Feature Space. BCL-DockANNScore contains an additional feature type not present in BCL-AffinityNet. Specifically, we binned hydrogen bonding pairs by both distance and angle. We considered that the strength of hydrogen bonding interactions is often approximated not only with distances between donor and acceptor atoms but also with the orientation angle. Therefore, we also developed a complementary feature to (eq 2) to assist with the description of well-formed hydrogen bonds. While (eq 2) is generalizable to any atom-based descriptor (or pair of descriptors if performing an asymmetric correlation) returning a scalar value, this descriptor is exclusively for hydrogen bond donor/acceptor pairs. Essentially, each distance interval specified by the boundaries r_a and r_b in eq 2 is equally partitioned into a user-specified number of bins (for this manuscript, nominally 45 bins of 8° each). Thus, for each distance bin there is also an angular component. See the Supporting Information for sample BCL code object files containing all properties employed in this study.

Input Sensitivity Analysis. The predictions for BCL-AffinityNet (and separately, BCL-DockANNScore) are the average predictions of the five cross-validated models. We can readily calculate feature importance for a single ANN by computing the magnitude of the input sensitivity across a dataset with respect to a given feature, after appropriate rescaling of the inputs. For model ensembles, the magnitude cannot be used or meaningfully averaged because feature input sensitivity may differ in the sign for various feature-instance pairings. While we could look at the raw average of input sensitivity of models across a given instance-feature pairing and then average the absolute value of that over the dataset, we suffer from an issue with relative scaling of the input sensitivities due to the nonlinearity of the ANN's transfer function. Rather than deriving an optimal weighted feature importance metric for ANN ensembles by some criteria, we chose to simply evaluate how often the models in the ensemble agreed on the sign of the derivative for each feature, averaged across the dataset.

This is a form of input sensitivity analysis we refer to as "consistency". Here, we specifically evaluate the consistency of feature column perturbations on result labels across cross-validation models. Features for which models in the ensemble agree on the derivative sign most routinely are interpreted as those that are of most importance to the ensemble's performance. Consistency is thus insensitive to the magnitude of feature's influence.

To calculate consistency, we iterate across all input feature columns of a training sample, perturb the feature value by a small amount (e.g., 0.01), propagate the perturbed inputs, and measure the result. For efficiency, we perform a forward propagation pass, followed by a backpropagation pass with a slightly modified result, which is readily transformed into the forward input sensitivities. This is done for each cross-validation model (in this manuscript, we performed fivefold cross-validation for all models). For each feature column, we

https://dx.doi.org/10.1021/acs.jcim.0c01001
J. Chem. Inf. Model. 2021, 61, 603−620

count the number of models that predict that the perturbation will improve the score vs the number of models that predict that the perturbation will worsen the score. This number is normalized such that when half of the models predict a negative change to the result and the other half predicts a positive change to the result, the net consistency is zero. The consistency result is averaged across all examples in the training set for each individual feature.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.0c01001.

Additional details, including command-line syntax for file preparation, model training, and pharmacophore map generation (PDF)

BCL code object files (ZIP)

■ AUTHOR INFORMATION

Corresponding Authors

Benjamin P. Brown − *Chemical and Physical Biology Program, Medical Scientist Training Program, Center for Structural Biology, Vanderbilt University, Nashville, Tennessee 37232, United States;* orcid.org/0000-0001-5296-087X; Email: benjamin.p.brown17@gmail.com

Jens Meiler − *Department of Chemistry, Center for Structural Biology and Departments of Pharmacology and Biomedical Informatics, Vanderbilt University, Nashville, Tennessee 37232, United States; Institute for Drug Discovery, Leipzig University Medical School, Leipzig SAC 04103, Germany;* Email: jens@meilerlab.org

Authors

Jeffrey Mendenhall − *Department of Chemistry, Center for Structural Biology, Vanderbilt University, Nashville, Tennessee 37232, United States*

Alexander R. Geanes − *Department of Chemistry, Center for Structural Biology, Vanderbilt University, Nashville, Tennessee 37232, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.0c01001

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

Work in the Meiler laboratory is supported through the NIH (R01GM080403 and R01DA046138). B.P.B. is supported through the NIH by a Ruth L. Kirschstein NRSA fellowship (F30DK118774). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

■ REFERENCES

(1) Leelananda, S. P.; Lindert, S. Computational Methods in Drug Discovery. *Beilstein J. Org. Chem.* **2016**, *12*, 2694−2718.

(2) Sliwoski, G.; Kothiwale, S.; Meiler, J.; Lowe, E. W., Jr. Computational Methods in Drug Discovery. *Pharmacol. Rev.* **2014**, *66*, 334−395.

(3) Butkiewicz, M.; Mueller, R.; Selic, D.; Dawson, E.; Meiler, J. In *Application of Machine Learning Approaches on Quantitative Structure Activity Relationships*, IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, 2009.

(4) Dahl, G. E. Multi-Task Neural Networks for Qsar Predictions. 2014, arXiv:1406.1231. arXiv.org e-Print archive. https://arxiv.org/abs/1406.1231.

(5) Mendenhall, J.; Meiler, J. Improving Quantitative Structure-Activity Relationship Models Using Artificial Neural Networks Trained with Dropout. *J. Comput.-Aided Mol. Des.* **2016**, *30*, 177−189.

(6) Hillebrecht, A.; Klebe, G. Use of 3d Qsar Models for Database Screening: A Feasibility Study. *J. Chem. Inf. Model* **2008**, *48*, 384−396.

(7) Manchester, J.; Czerminski, R. Samfa: Simplifying Molecular Description for 3d-Qsar. *J. Chem. Inf. Model.* **2008**, 1167.

(8) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and Qsar Modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947−1958.

(9) Geanes, A. R.; Cho, H. P.; Nance, K. D.; McGowan, K. M.; Conn, P. J.; Jones, C. K.; Meiler, J.; Lindsley, C. W. Ligand-Based Virtual Screen for the Discovery of Novel M5 Inhibitor Chemotypes. *Bioorg. Med. Chem. Lett.* **2016**, *26*, 4487−4491.

(10) Lowe, E. W., Jr.; Ferrebee, A.; Rodriguez, A. L.; Conn, P. J.; Meiler, J. 3d-Qsar Comfa Study of Benzoxazepine Derivatives as Mglur5 Positive Allosteric Modulators. *Bioorg. Med. Chem. Lett.* **2010**, *20*, 5922−5924.

(11) Mueller, R.; Rodriguez, A. L.; Dawson, E. S.; Butkiewicz, M.; Nguyen, T. T.; Oleszkiewicz, S.; Bleckmann, A.; Weaver, C. D.; Lindsley, C. W.; Conn, P. J.; Meiler, J. Identification of Metabotropic Glutamate Receptor Subtype 5 Potentiators Using Virtual High-Throughput Screening. *ACS Chem. Neurosci.* **2010**, *1*, 288−305.

(12) Bleckmann, A.; Meiler, J. Epothilones: Quantitative Structure Activity Relations Studied by Support Vector Machines and Artificial Neural Networks. *QSAR Comb. Sci.* **2003**, *22*, 722−728.

(13) Lowe, E. W.; Butkiewicz, M.; Woetzel, N.; Meiler, J. In *Gpu-Accelerated Machine Learning Techniques Enable Qsar Modeling of Large Hts Data*, 2012 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), San Diego, CA, 2012; pp 314−320.

(14) Sliwoski, G.; Mendenhall, J.; Meiler, J. Autocorrelation Descriptor Improvements for Qsar: 2da_Sign and 3da_Sign. *J. Comput.-Aided Mol. Des.* **2015**, 209.

(15) Sliwoski, G.; Lowe, E. W.; Butkiewicz, M.; Meiler, J. Bcl::Emas−Enantioselective Molecular Asymmetry Descriptor for 3d-Qsar. *Molecules* **2012**, *17*, 9971−9989.

(16) Mendenhall, J.; Brown, B. P.; Kothiwale, S.; Meiler, J. Bcl::Conf: Improved Open-Source Knowledge-Based Conformation Sampling Using the Crystallography Open Database. *J. Chem. Inf. Model.* **2020**, DOI: 10.1021/acs.jcim.0c01140.

(17) Butkiewicz, M.; Wang, Y.; Bryant, S. H.; Lowe, E. W.; Weaver, D. C.; Meiler, J. High-Throughput Screening Assay Datasets from the Pubchem Database. *Chem. Inf.* **2017**, *3*, 1−7.

(18) Butkiewicz, M.; Lowe, E. W., Jr.; Mueller, R.; Mendenhall, J. L.; Teixeira, P. L.; Weaver, C. D.; Meiler, J. Benchmarking Ligand-Based Virtual High-Throughput Screening with the Pubchem Database. *Molecules* **2013**, *18*, 735−756.

(19) Wang, L.; Chambers, J.; Abel, R. Protein−Ligand Binding Free Energy Calculations with Fep+. In Biomolecular Simulations In *Methods and Protocols*; Bonomi, M.; Camilloni, C., Eds.; Springer New York: New York, NY, 2019; pp 201−232.

(20) Wang, L.; Wu, Y.; Deng, Y.; Kim, B.; Pierce, L.; Krilov, G.; Lupyan, D.; Robinson, S.; Dahlgren, M. K.; Greenwood, J.; Romero, D. L.; Masse, C.; Knight, J. L.; Steinbrecher, T.; Beuming, T.; Damm, W.; Harder, E.; Sherman, W.; Brewer, M.; Wester, R.; Murcko, M.; Frye, L.; Farid, R.; Lin, T.; Mobley, D. L.; Jorgensen, W. L.; Berne, B. J.; Friesner, R. A.; Abel, R. 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*, 2695−2703.

(21) Zou, J.; Tian, C.; Simmerling, C. Blinded Prediction of Protein−Ligand Binding Affinity Using Amber Thermodynamic Integration for the 2018 D3r Grand Challenge 4. *J. Comput.-Aided Mol. Des.* **2019**, *33*, 1021−1029.

(22) Su, M.; Yang, Q.; Du, Y.; Feng, G.; Liu, Z.; Li, Y.; Wang, R. Comparative Assessment of Scoring Functions: The Casf-2016 Update. *J. Chem. Inf. Model.* **2019**, *59*, 895−913.
(23) Wang, E.; Sun, H.; Wang, J.; Wang, Z.; Liu, H.; Zhang, J. Z. H.; Hou, T. End-Point Binding Free Energy Calculation with Mm/Pbsa and Mm/Gbsa: Strategies and Applications in Drug Design. *Chem. Rev.* **2019**, *119*, 9478−9508.
(24) Sun, H.; Duan, L.; Chen, F.; Liu, H.; Wang, Z.; Pan, P.; Zhu, F.; Zhang, J. Z. H.; Hou, T. Assessing the Performance of Mm/Pbsa and Mm/Gbsa Methods. 7. Entropy Effects on the Performance of End-Point Binding Free Energy Calculation Approaches. *Phys. Chem. Chem. Phys.* **2018**, *20*, 14450−14460.
(25) Miller, B. R.; McGee, T. D.; Swails, J. M.; Homeyer, N.; Gohlke, H.; Roitberg, A. E. Mmpbsa.Py: An Efficient Program for End-State Free Energy Calculations. *J. Chem. Theory Comput.* **2012**, *8*, 3314−3321.
(26) Stein, R. M.; Kang, H. J.; McCorvy, J. D.; Glatfelter, G. C.; Jones, A. J.; Che, T.; Slocum, S.; Huang, X.-P.; Savych, O.; Moroz, Y. S.; Stauch, B.; Johansson, L. C.; Cherezov, V.; Kenakin, T.; Irwin, J. J.; Shoichet, B. K.; Roth, B. L.; Dubocovich, M. L. Virtual Discovery of Melatonin Receptor Ligands to Modulate Circadian Rhythms. *Nature* **2020**, *579*, 609−614.
(27) DeLuca, S.; Khar, K.; Meiler, J. Fully Flexible Docking of Medium Sized Ligand Libraries with Rosettaligand. *PLoS One* **2015**, *10*, No. e0132508.
(28) Ballester, P. J.; Mitchell, J. B. A Machine Learning Approach to Predicting Protein-Ligand Binding Affinity with Applications to Molecular Docking. *Bioinformatics* **2010**, *26*, 1169−1175.
(29) Wang, C.; Zhang, Y. Improving Scoring-Docking-Screening Powers of Protein-Ligand Scoring Functions Using Random Forest. *J. Comput. Chem.* **2017**, *38*, 169−177.
(30) Pereira, J. C.; Caffarena, E. R.; dos Santos, C. N. Boosting Docking-Based Virtual Screening with Deep Learning. *J. Chem. Inf. Model.* **2016**, *56*, 2495−2506.
(31) Ragoza, M.; Hochuli, J.; Idrobo, E.; Sunseri, J.; Koes, D. R. Protein-Ligand Scoring with Convolutional Neural Networks. *J. Chem. Inf. Model.* **2017**, *57*, 942−957.
(32) 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*, 287−296.
(33) Wallach, I.; Dzamba, M.; Heifets, A. Atomnet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-Based Drug Discovery. 2015, arXiv:1510.02855. arXiv.org e-Print archive. https://arxiv.org/abs/1510.02855.
(34) Minovski, N.; Zuperl, S.; Drgan, V.; Novic, M. Assessment of Applicability Domain for Multivariate Counter-Propagation Artificial Neural Network Predictive Models by Minimum Euclidean Distance Space Analysis: A Case Study. *Anal. Chim. Acta* **2013**, *759*, 28−42.
(35) Sheridan, R. P. Three Useful Dimensions for Domain Applicability in Qsar Models Using Random Forest. *J. Chem. Inf. Model.* **2012**, *52*, 814−823.
(36) Tetko, I. V.; Sushko, I.; Pandey, A. K.; Zhu, H.; Tropsha, A.; Papa, E.; Oberg, T.; Todeschini, R.; Fourches, D.; Varnek, A. Critical Assessment of Qsar Models of Environmental Toxicity against Tetrahymena Pyriformis: Focusing on Applicability Domain and Overfitting by Variable Selection. *J. Chem. Inf. Model.* **2008**, *48*, 1733−1746.
(37) Schroeter, T.; Schwaighofer, A.; Mika, S.; Ter Laak, A.; Suelzle, D.; Ganzer, U.; Heinrich, N.; Müller, K.-R. Estimating the Domain of Applicability for Machine Learning Qsar Models: A Study on Aqueous Solubility of Drug Discovery Molecules. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 651−664.
(38) Ruiz, I. L.; Gómez-Nieto, M. Á. Study of the Applicability Domain of the Qsar Classification Models by Means of the Rivality and Modelability Indexes. *Molecules* **2018**, *23*, No. 2756.
(39) Roy, K.; Kar, S.; Ambure, P. On a Simple Approach for Determining Applicability Domain of Qsar Models. *Chemom. Intell. Lab. Syst.* **2015**, *145*, 22−29.
(40) Carrió, P.; Pinto, M.; Ecker, G.; Sanz, F.; Pastor, M. Applicability Domain Analysis (Adan): A Robust Method for Assessing the Reliability of Drug Property Predictions. *J. Chem. Inf. Model.* **2014**, *54*, 1500−1511.
(41) Sieg, J.; Flachsenberg, F.; Rarey, M. In Need of Bias Control: Evaluating Chemical Data for Machine Learning in Structure-Based Virtual Screening. *J. Chem. Inf. Model.* **2019**, *59*, 947−961.
(42) Koehler Leman, J.; Weitzner, B. D.; Renfrew, P. D.; Lewis, S. M.; Moretti, R.; Watkins, A. M.; Mulligan, V. K.; Lyskov, S.; Adolf-Bryfogle, J.; Labonte, J. W.; Krys, J.; Consortium, R.; Bystroff, C.; Schief, W.; Gront, D.; Schueler-Furman, O.; Baker, D.; Bradley, P.; Dunbrack, R.; Kortemme, T.; Leaver-Fay, A.; Strauss, C. E. M.; Meiler, J.; Kuhlman, B.; Gray, J. J.; Bonneau, R. Better Together: Elements of Successful Scientific Software Development in a Distributed Collaborative Community. *PLoS Comput. Biol.* **2020**, *16*, No. e1007507.
(43) Hemmer, M. C.; Steinhauer, V.; Gasteiger, J. Deriving the 3d Structure of Organic Molecules from Their Infrared Spectra. *Vib. Spectrosc.* **1999**, *19*, 151−164.
(44) Brown, B. P.; Mendenhall, J.; Meiler, J. Bcl::Molalign: Three-Dimensional Small Molecule Alignment for Pharmacophore Mapping. *J. Chem. Inf. Model.* **2019**, *59*, 689−701.
(45) 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*, 4540−4549.
(46) Yang, J.; Shen, C.; Huang, N. Predicting or Pretending: Artificial Intelligence for Protein-Ligand Interactions Lack of Sufficiently Large and Unbiased Datasets. *Front. Pharmacol.* **2020**, *11*, No. 69.
(47) Chen, L.; Cruz, A.; Ramsey, S.; Dickson, C. J.; Duca, J. S.; Hornak, V.; Koes, D. R.; Kurtzman, T. Hidden Bias in the Dud-E Dataset Leads to Misleading Performance of Deep Learning in Structure-Based Virtual Screening. *PLoS One* **2019**, *14*, No. e0220113.
(48) Witten, I. H.; Frank, E.; Hall, M. A.; Pal, C. J. Data Mining. *Practical Machine Learning Tools and Techniques*, 4th ed.; Morgan Kaufmann Publishers Inc., 2016.
(49) Niculescu-Mizil, A.; Caruana, R. In *Predicting Good Probabilities with Supervised Learning*, ICML'05, 2005.
(50) Zadrozny, B.; Elkan, C. In *KDD '02*, Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Alberta, Canada; Association for Computing Machinery, 2002; pp 694−699.
(51) Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. Improved Protein-Ligand Docking Using Gold. *Proteins: Struct., Funct., Genet.* **2003**, *52*, 609−623.
(52) Jain, A. N. Surflex: Fully Automatic Flexible Molecular Docking Using a Molecular Similarity-Based Search Engine. *J. Med. Chem.* **2003**, *46*, 499−511.
(53) *CCG Molecular Operating Environment (MOE)*, version 2007.09; Chemical Computing Group (CCG): Montreal, Quebec, 2007.
(54) Zhu, D.; Huang, H.; Pinkas, D. M.; Luo, J.; Ganguly, D.; Fox, A. E.; Arner, E.; Xiang, Q.; Tu, Z.-C.; Bullock, A. N.; Brekken, R. A.; Ding, K.; Lu, X. 2-Amino-2,3-Dihydro-1h-Indene-5-Carboxamide-Based Discoidin Domain Receptor 1 (Ddr1) Inhibitors: Design, Synthesis, and in Vivo Antipancreatic Cancer Efficacy. *J. Med. Chem.* **2019**, *62*, 7431−7444.
(55) Wang, Z.; Bian, H.; Bartual, S. G.; Du, W.; Luo, J.; Zhao, H.; Zhang, S.; Mo, C.; Zhou, Y.; Xu, Y.; Tu, Z.; Ren, X.; Lu, X.; Brekken, R. A.; Yao, L.; Bullock, A. N.; Su, J.; Ding, K. Structure-Based Design of Tetrahydroisoquinoline-7-Carboxamides as Selective Discoidin Domain Receptor 1 (Ddr1) Inhibitors. *J. Med. Chem.* **2016**, 5911.
(56) O'Meara, M. J.; Leaver-Fay, A.; Tyka, M.; Stein, A.; Houlihan, K.; DiMaio, F.; Bradley, P.; Kortemme, T.; Baker, D.; Snoeyink, J.; Kuhlman, B. A Combined Covalent-Electrostatic Model of Hydrogen Bonding Improves Structure Prediction with Rosetta. *J. Chem. Theory Comput.* **2015**, *11*, 609−622.

https://dx.doi.org/10.1021/acs.jcim.0c01001
J. Chem. Inf. Model. 2021, 61, 603−620

(57) Fu, D. Y.; Meiler, J. Rosettaligandensemble: A Small-Molecule Ensemble-Driven Docking Approach. *ACS Omega* **2018**, *3*, 3655−3664.
(58) Meiler, J.; Baker, D. Rosettaligand: Protein-Small Molecule Docking with Full Side-Chain Flexibility. *Proteins* **2006**, *65*, 538−548.
(59) Smith, S. T.; Meiler, J. Assessing Multiple Score Functions in Rosetta for Drug Discovery. *PLoS One* **2020**, *15*, No. e0240450.
(60) Liu, Z.; Li, Y.; Han, L.; Li, J.; Liu, J.; Zhao, Z.; Nie, W.; Liu, Y.; Wang, R. Pdb-Wide Collection of Binding Data: Current Status of the Pdbbind Database. *Bioinformatics* **2015**, *31*, 405−412.
(61) Sahigara, F.; Mansouri, K.; Ballabio, D.; Mauri, A.; Consonni, V.; Todeschini, R. Comparison of Different Approaches to Define the Applicability Domain of Qsar Models. *Molecules* **2012**, *17*, 4791.

https://dx.doi.org/10.1021/acs.jcim.0c01001
J. Chem. Inf. Model. 2021, 61, 603−620