ACS OMEGA

http://pubs.acs.org/journal/acsodf Article

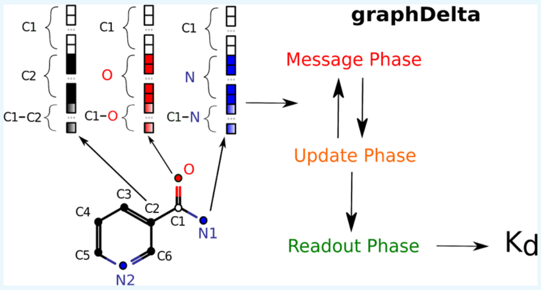
graphDelta: MPNN Scoring Function for the Affinity Prediction of Protein−Ligand Complexes

Dmitry S. Karlov, Sergey Sosnin, Maxim V. Fedorov,* and Petr Popov*

Cite This: *ACS Omega* 2020, 5, 5150−5159 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: In this work, we present graph-convolutional neural networks for the prediction of binding constants of protein−ligand complexes. We derived the model using multi task learning, where the target variables are the dissociation constant (K_d), inhibition constant (K_i), and half maximal inhibitory concentration (IC_{50}). Being rigorously trained on the PDBbind dataset, the model achieves the Pearson correlation coefficient of 0.87 and the RMSE value of 1.05 in pK units, outperforming recently developed 3D convolutional neural network model K_{deep}.



1. INTRODUCTION

The majority of marketed drugs act via non-covalent binding to a macromolecular target in a human organism, such as protein molecules or nucleic acids.[1] The binding affinity is one of the major determinants along with absorption, distribution, metabolism, and excretion properties of the dose necessary to achieve a biological response and, consequently, the additional off-target impact on the organism. Many efforts are made to develop robust and powerful binding affinity prediction models.[2] With the substantial growth of atomic structures in the Protein Data Bank (PDB[3]), now it is possible to derive reliable scoring functions.[4−11] Depending on the formulation of the optimization problem, the scoring functions aim (i) to identify correct (near-native) binding pose amongst conformations of modeled putative binding candidates or (ii) to rank a given set of chemical compounds (ligands) with respect to its binding affinity to a particular target.[12] These task-specific SFs should combine the high speed of computation with accuracy,[13] and usually, the parameters of the recent scoring functions were obtained without affinity information.[14] SFs can be divided into three categories based on the way parameterization: (1) force-field-based SFs were designed from physical principles based on the theoretical representation of interatomic potentials;[15] (2) empirical SFs, which utilize force-field based canvas with the parameter set, were tuned to reproduce experimental affinity measurements;[4,8,9] (3) knowledge-based SFs trained with experimental structural data to approximate interatomic potentials as arbitrary functions defined by piece-wise linear interpolation or in the other way.[7,11]

1.1. Machine Learning Approaches for Protein−Ligand Scoring Functions.

The majority of the mentioned approaches basically utilize a linear regression approach to account for different terms describing intermolecular interactions such as hydrogen bonding, $\pi-\pi$ stacking, $\pi-$cation interactions, entropic contribution, van der Waals interactions, and so forth[4] for the sake of efficiency and are easy to interpret because of the possibility of per atom decomposition of the score value. Ain et al. reported[2] the possible improvement in performance when the scoring function is not constrained to a predefined functional form. These machine learning approaches[16] were applied for both classification and regression tasks in the different areas of science and technology. One of the most known machine learning-based SF for the evaluation of protein−ligand interactions is the RF-score[10] that utilizes the ensemble of decision trees to approximate binding affinity using the receptor−ligand interatomic interactions counts as the descriptors. Some authors customized this methodology by training target-specific scoring functions using AutoDock Vina scoring terms as descriptors.[17,18]

1.2. Sampling Approaches to Binding Affinity Estimation.

It should be noted that because of the peculiarities of the common drug discovery pipelines, hit identification and hit to lead optimization are considered to be separate stages, and fast scoring is used for the former and the sophisticated molecular simulations are performed for the latter. The free energy perturbation (FEP) technique[19] is based on the alchemical transformations and allows us to achieve in some cases good results for the affinity ranking in a series of closely related compounds. The end-state free energy approaches (MM-PB(GB)SA)[20] being not so demanding for

Received: December 6, 2019
Accepted: February 21, 2020
Published: March 9, 2020

https://dx.doi.org/10.1021/acsomega.9b04162

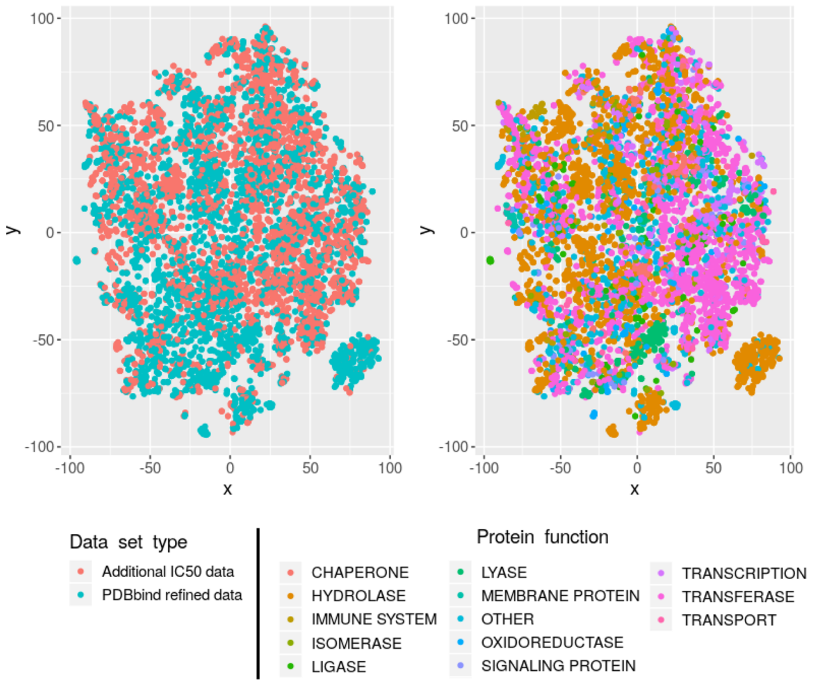


Figure 1. Results of *t*-SNE mapping of ligand protein interactions represented by SILIRID[46] fingerprints: (left) blue color mark complexes that consist the initial PDBbind refined set while the red one represents the additional data; (right) color scheme is based on protein functions.

computational resources can be considered as a cheaper but a less accurate alternative to FEP.[21] While the chemical space exploration[22,23] leads to creation of enormous virtual databases, there is a strong demand to assess for better techniques that are faster than ensemble methods (FEP, MM-PBSA) and more accurate than scoring functions developed for virtual screening.

1.3. Progress in Applications of Deep Learning for Chemical Problems. Deep neural networks (DNN) are powerful machine learning models with broad applicability to different regression and classification tasks. Progress in hardware and software development for large-scale training of convolutional neural networks (CNN)[24] and recurrent neural networks (RNN)[25] resulted in great achievements in computer vision, natural language and signal processing, and other related problems[26] The promising results obtained in computational chemistry[27,28] structure generation tasks[29,30] and QSPR/QSAR[31,32] by means of DNN, demonstrate that the application of DNN undoubtedly has a strong potential to growth in the area of the computer-aided molecular design.

1.4. 3D Convolutional Neural Networks for Protein–Ligand Scoring Functions. Recently it was shown that 3D convolutional neural networks (3D CNN) can be applied to derive scoring functions for binding affinity prediction.[33,34] Current approaches use voxelized representation of a molecular complex, where voxel channels encode physicochemical properties, similar to the RGB channels in images. Molecular representations for processing by 3D CNN can be constructed in several ways: each atom or a group of atoms can be represented either by a separate channel or a channel which can represent some kind of superposition of atoms. For example, one can calculate interactions with a probe atom to construct 3D molecular field or use some physicochemical or DFT (3D electron density) calculations as 3D filed representations.[31,35] Both of these approaches have limitations: atom-to-channel representation leads to dramatic increase in the number of input channels, which are crucial for memory consumption. It is also inefficient because many channels are empty or sparse. The use of molecular fields, in turn, results in losing information. The balance between the quality of representation and memory requirements is a fundamental problem with 3D CNNs. This fact motivates us to search for new ways and architectures for 3D deep learning in molecular science. K_{deep}[33] trained using PDBbind data and aimed at predicting absolute binding affinities takes a set of 3D grids representing map of certain structural features (hydrophobic, aromatic, h-bond acceptors, and so forth.) as the input data. Also, the model developed by Ragoza et al.[34] with the goal to improve virtual screening results act as a classification model which operates with 3D maps defined by smina (scoring and minimization with AutoDock Vina) atom types.[36] However, the performance of these models can still be improved. The major drawback of 3D CNN is the enormous number of parameters, which results in high-demand of computational resources and GPU memory; meanwhile, GPU memory is limited.[37] Interestingly, a special architecture which is applied simultaneously for two maps representing similar ligands was used to predict differences in affinity to input ligands and achieved better results compared to MM-GBSA and QSAR in blind predictions.[38]

1.5. Geometric Deep Learning and Tasks of This Work. The limitations of 3D CNN architectures motivated researchers to search more natural ways of processing chemical structures. Geometric deep learning is a bunch of approaches which aim to generalize neural network to non-Euclidian manifolds, in particular, to graphs.[39] Simultaneously, molecules can be represented as labeled and weighted graphs in chemoinformatics, and the idea of the applications of geometric deep learning seems to be natural and can lead to very promising results.[40,41] One of the graph convolutional architectures (PotentialNet[42]) was trained on PDBBind 2007[43] and applied to the affinity prediction problem. Authors use

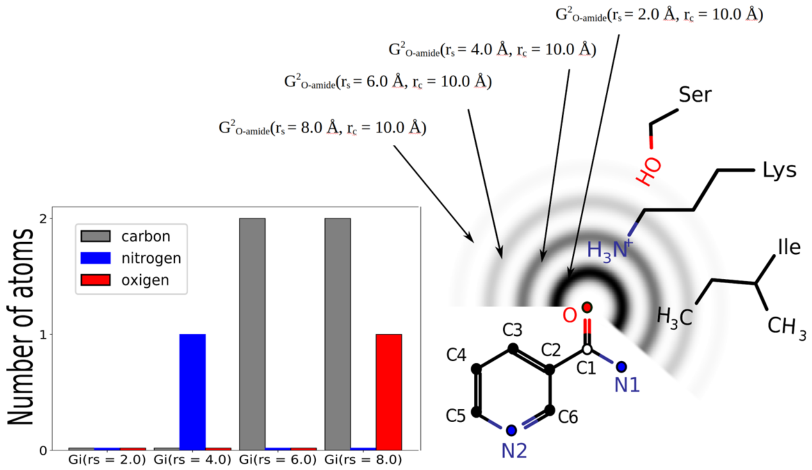


Figure 2. Illustration of the G_i^2 computation.

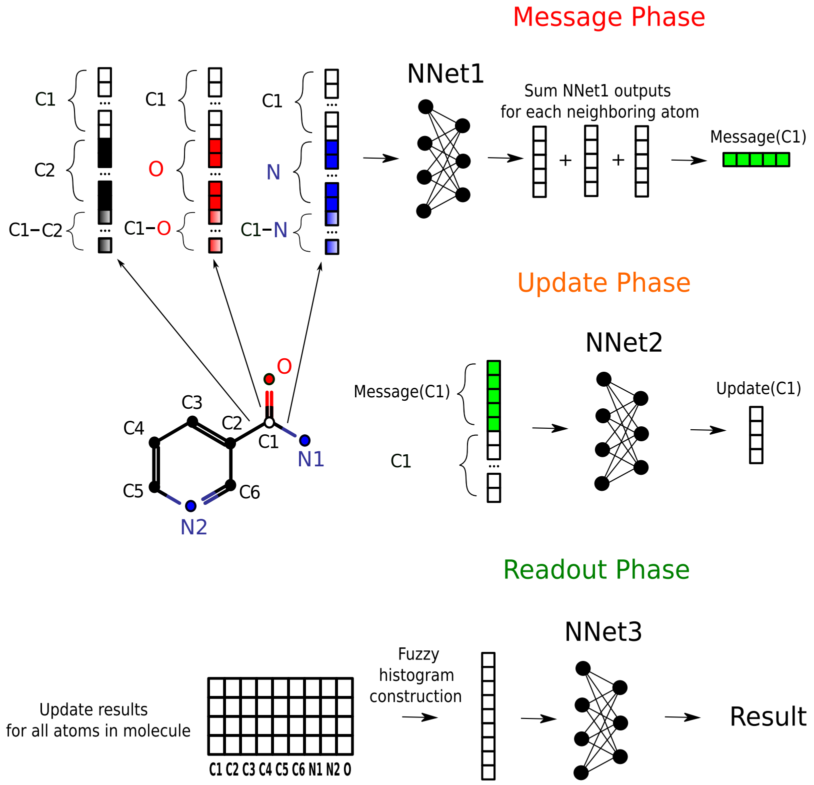


Figure 3. General description of the MPNN forward pass with interaction net architecture.

gated graph neural network architectures which utilizes RNN for the update stage. PotentialNet can perform graph convolution operations for both covalent and non-covalent interactions; in other words, authors include in an initial graph the nearby residues.

The goal of this work is the design of the scoring function with a possibility to predict binding free energy for a diverse set of chemical compounds and protein targets based on subclass of graph CNN—message passing neural networks which demonstrated very promising results in approximation of DFT electron energies[28] for the QM9 small molecule data set.[44] We assess its performance on different data sets and compare it with the existing tools. We showed that the message passing neural network (MPNN) can be a very efficient tool for modeling protein−ligand interactions.

2. RESULTS AND DISCUSSION

The thorough description of the training set properties is necessary for the applicability domain definition of a scoring function. The extended training set obtained by IC_{50} data addition contains more druglike molecules (Figure S1) than the initial data set. It may indicate that the optimal applicability domains of the model are structures which fall into line with Lipinski's rule of five. To analyze the distributions of protein targets, we performed a *t*-SNE[45] mapping (by *scikit-learn 0.19.1* package) of protein−ligand interaction descriptors[46] into a 2D space using an Open Drug Discovery Toolkit[47] (ODDT). The obtained distribution is shown in Figure 1. It should be noted that the usage of the additional IC_{50} data improves the representation of the known types of interactions rather than introduce completely undescribed binding modes. The addi-

Table 1. Network Architecture of Message, Update, and Readout Functions[a]

		message function			update function			readout function		
T	layer	in	out	BN	in	out	BN	in	out	BN
1	1	751	200	yes	473	200	yes			
	2	200	100	yes	200	100	yes			
	3	100	100	no	100	100	no			
2	1	205	200	yes	200	200	yes			
	2	200	100	yes	200	100	yes			
	3	100	100	no	100	100	no			
3	1	205	200	yes	200	200	yes	1100	300	yes
	2	200	100	yes	200	100	yes	300	200	yes
	3	100	100	no	100	100	no	200	100	yes
	4							100	2	no

[a] "In" and "Out" means the number of input and output neurons in the current layer, and "BN" denotes the application of the batch normalization layer.

Table 2. Results of graphDelta Evaluation on the CSAR Data Compared to K_{deep} and RF-Score[a]

			graphDelta		K_{deep}[33]		RF-score[33]	
dataset	epochs	MT/ST	*r*	RMSE	*r*	RMSE	*r*	RMSE
CASP2016	500	true	0.82	1.22	0.82	1.27	0.80	1.39
	500	false	0.84	1.17				
	1000	true	0.86	1.11				
	1000	false	0.84	1.16				
	2000	true	0.84	1.17				
	2000	false	**0.87**	**1.05**				
CSAR NRC HiQ set1	500	true	0.74	1.67	0.72	2.08	**0.77**	1.99
	500	false	0.64	1.81				
	1000	true	0.71	1.70				
	1000	false	0.71	1.66				
	2000	true	0.74	**1.59**				
	2000	false	0.74	**1.59**				
CSAR NRC HiQ set2	500	true	0.60	1.86	0.65	1.91	**0.75**	1.66
	500	false	0.59	1.72				
	1000	true	0.56	1.92				
	1000	false	0.71	**1.52**				
	2000	true	0.64	1.73				
	2000	false	0.71	1.53				
CSAR12	500	true	0.52	1.16	0.37	1.59	0.46	1.00
	500	false	0.41	1.37				
	1000	true	**0.59**	**0.94**				
	1000	false	0.54	1.11				
	2000	true	0.52	1.10				
	2000	false	0.48	1.14				
CSAR14	500	true	0.72	1.40	0.61	1.75	**0.80**	**0.87**
	500	false	0.66	1.51				
	1000	true	0.65	1.34				
	1000	false	0.59	1.67				
	2000	true	0.70	1.32				
	2000	false	0.74	1.22				
average	500	true	0.68	1.46	0.62	1.72	**0.72**	1.38
	500	false	0.63	1.52				
	1000	true	0.67	1.40				
	1000	false	0.68	1.42				
	2000	true	0.69	1.38				
	2000	false	0.71	**1.31**				

[a] Bold font is used to stress the best correlation coefficient and RMSE for the selected data set.

tional data, in general, describe the interaction with transferase and hydrolase enzymes, which appeared to be the most representative class of proteins in the current data set.

The trained MPNN scoring function demonstrates very good results for the CASF2016 test set composed of the X-ray structures significantly outperforming both K_{deep} and RF score in terms of Pearson R. Unfortunately, it is difficult to accurately

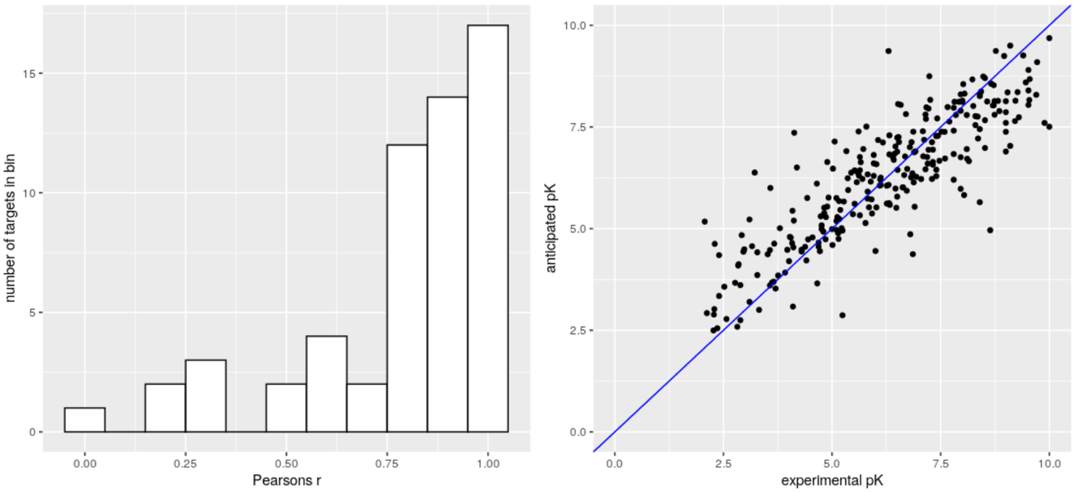


Figure 4. Results of prediction (graphDelta, 2000 epochs, single task) for the CASF2016 data set: (left) histogram of correlation coefficients computed for all targets from CASF2016, (right) the depiction of the prediction results with the trend line.

Table 3. Results of graphDelta Evaluation on the Data Set Used for FEP and MM-PBSA Evaluation (graphDelta, 2000 Epochs, Multi Task)[a]

	graphDelta		K_{deep}		RF-score		FEP or MM-PBSA	
subset	r	RMSE	r	RMSE	r	RMSE	r	RMSE
p38	0.64	1.56	0.36	1.57	0.48	**0.9**	**0.6**	1.03
PTP1B	0.46	1.22	0.58	0.93	0.26	**0.9**	**0.80**	1.22
thrombin	0.39	0.74	0.58	**0.44**	0.08	0.71	**0.71**	0.93
Tyk2	0.17	1.08	0.05	1.23	0.41	0.94	**0.89**	**0.93**
Bace	0.65	0.78	−0.06	0.84	−0.14	**0.65**	**0.78**	1.03
CDK2	0.19	1.94	**0.69**	1.26	−0.23	1.05	0.48	**0.91**
JNK1	0.33	1.53	0.69	1.18	**0.61**	0.5	**0.85**	1.00
MCL1	0.22	1.12	0.34	1.04	0.52	**0.99**	**0.77**	1.41
AMPA	0.39	1.37	0.74	**1.32**	0.38	1.71	**0.78**	0.62
average	0.35	1.31	0.41	1.07	0.26	**0.92**	**0.75**	1.00

[a] Bold font is used to stress the best correlation coefficient and RMSE for the selected data set. MM-PBSA data are provided for AMPA receptor ligands, while FEP data are provided for other targets.

compare RMSE because of the unavailability of the raw data from K_{deep} and RF-score. The authors of the K_{deep} mentioned that the inclusion of the low-quality data to the training set did not improve the results, probably because of the excessive noise contribution to the model. In this work, we showed that the inclusion of the IC_{50} subset with a similar structural quality can improve the model and speed up model training when multi task learning is used. The neural network can predict a set of properties simultaneously, and the prediction of IC_{50} and K_d, which are strongly correlated but still slightly differs, significantly improves the CASF2016 prediction results at 1000 epochs of training. At the same time, the multi task model trained for 2000 epochs appeared to be slightly overtrained compared to the single task model, which demonstrated the best performance for CASP2016 among all the used models. The results are presented in Table 2. The distribution of Pearson correlation coefficients computed for each target from CASF2016 demonstrates the improvement compared to K_{deep}: the number of Pearson $r < 0.75$ is 14 for graphDelta and is 32 (more than the half of targets) for K_{deep} and the number of Pearson $r < 0.0$ is zero for graphDelta and is six for K_{deep} (Figure 4). graphDelta (single task, 2000 epochs) demonstrates the best prediction rates for CASF2016 compared to the RF-score (one-tailed, $z = -2.78$, $P = 0.0027$) and K_{deep} (one-tailed, $z = -2.09$, $P = 0.018$) in terms of Pearson correlation coefficients. The single task graphDelta model trained for 2000 epochs outperforms K_{deep} or yields the similar results in terms of RMSE and Pearson correlation coefficients for CASF2016 and CSAR sets, except CSAR NRC HiQ set2, while RF-score yields better results than graphDelta in CSAR NRC HiQ set2 and CSAR14 (RMSE) and for CSAR NRC HiQ set1, CSAR NRC HiQ set2, CSAR14 (Pearson r). In average, graphDelta outperforms K_{deep} for these four data sets and yields practically similar results as the RF-score in terms of Pearson r and outperforms it in terms RMSE (Table 2).

To compare the obtained results with PotentialNet,[42] we trained our model on PDBbind v.2007 which is about eight times less in size compared to the initial training set. PotentialNet for the smaller training sets yields better results ($r = 0.82$) than GraphDelta ($r = 0.38$) possibly because of the small size of the training set. It should be noted that the learning procedure was accompanied by fast overtraining and leaps to the prediction of the mean value.

The graphDelta evaluation on the FEP and MM-PBSA data sets demonstrated worse results compared to the other SF where K_{deep} demonstrated the best results in terms of the Pearson correlation coefficient and RF-score shows the best RMSE (Table 3). The application of the MPNN scoring function (multi task, 2000 epochs of training) yields better results among ML scoring functions only for two subsets of the

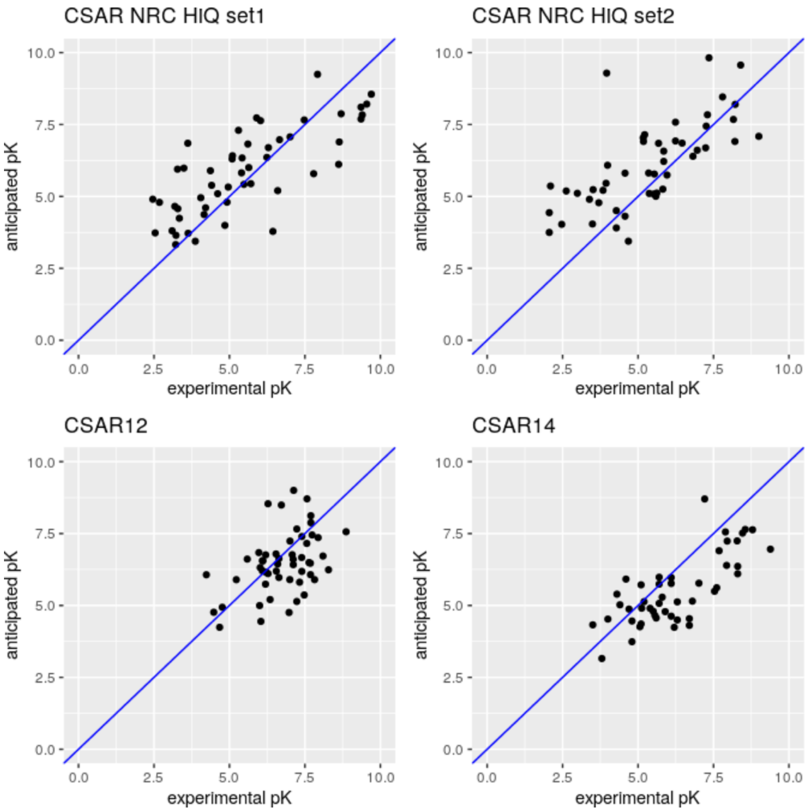


Figure 5. Results of prediction (graphDelta, 2000 epochs, single task) for CSAR data sets.

FEP data set: p38 and BACE. This dataset consists presumably of hydrolase (PTP1B, thrombin, Bace) and kinase (Tyk2, Jnk1, p38, CDK2) targets, and multi task demonstrated better results because of the extended training set by these types of proteins. The other graphDelta models yielded even worse results in terms of RMSE and Pearson r (see Table S1 and Figures S7—S11; Supporting Information). However, it should be noted that all examined SF gave poor results for this set of closely related structures.

In this work, our goal was to develop a novel tool for scoring of protein−ligand interactions based on graph-CNN. Despite the performance improvement in CASF2016 and some other X-ray data sets, both the RMSE and correlation coefficient decreases for some data sets obtained by docking. We believe that the performance improvement is caused by the addition of high-quality IC_{50} data to the training set which allowed the increase of the training set size in more than two times. This data set extension allowed us to train the model obtaining similar results faster (1000 epochs vs 2000 for Single task model). Noteworthy, the additional IC_{50} data may introduce a skew performing better on kinases and hydrolases which are the main content of the cleaned IC_{50} data set. This fact may be confirmed by the better performance of the multi task model on the docked data (Tables 3 and S1). At the same time, the performance of the examined models is still worse than trajectory-based approaches (FEP, MM-PBSA). Although trajectory-based approaches work well for some systems, sometimes, deep-learning-based models surpass their results.[38] It should be noted that trajectory-based affinity prediction methods works in the case of sufficient sampling which may be tricky for some types of proteins.[48] Modest results obtained for the complexes obtained by docking suggest the necessity of model improvement. We suggest two ways to accomplish this task. The first one is the application of this model to molecular dynamics trajectories or Monte-Carlo ensembles of structures which may be even slower than the trajectory-based approaches. The other possible approach is to apply proper data augmentation scheme which is not a straightforward task. Augmentation techniques which are often for 3D-CNN such as a shift of the box center and random rotations around are not suitable for graph CNN. We made available this scoring function at http://mpnn.syntelly.com/.

3. COMPUTATIONAL METHODS

3.1. Data Sets. The main source of data for the current work was PDBbind[49] (v.2018) containing 16151 protein−ligand complexes derived from Protein Data Bank (PDB) accompanied with their binding data in terms of dissociation (K_d) and inhibition (K_i) constants and half maximal inhibitory concentration (IC_{50}). A smaller "refined" set (4463 complexes) was compiled based on the following rules: the structure resolution less than 2.5 Å, an R-factor less than 0.25, ligand should be bound noncovalently and without steric clashes (any distance between pairs of ligand−protein atoms is more than 2.0 Å), pK is inside the range from 2 to 12, and complexes labeled only by IC_{50} are eliminated. The reason for the latter action is the substrate concentration dependence of IC_{50} (Cheng−Prusoff eq 1, where $[S]$ and K_m are the substrate concentration in the experiment and Michaelis constant, respectively) and cannot be in union with the K_i/K_d subset. However, in practice, the pIC_{50} values (logarithmically transformed IC_{50}) are usually less and within one logarithmic value compared to pK_i/pK_d. This bias can be easily learned using the model in the multi task learning mode[50] when the last layer simultaneously predicts both pK and pIC_{50}. Thus, we prepared a novel subset containing both pIC_{50} and pK in all

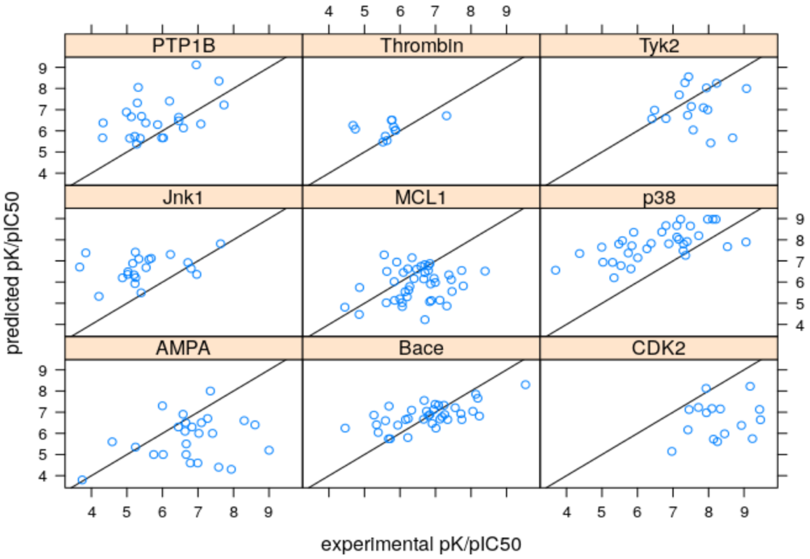


Figure 6. Results of prediction (graphDelta, 2000 epochs, multi task) for data sets used to assess FEP[52] and MM-PBSA[53] performance.

other quality criteria identical to the mentioned "refined" subset yielding 8766 complex structures. The idea to extend PDBbind refined set by low-quality data was also reported by Li et al.[51] A core set (285 items) used for critical assessment of scoring[12] (CASF) 2016 was not changed compared to the previous version of the database, facilitating the performance matching with the other scoring functions. Additional test sets were used for comparison with other available models: two subsets of CSAR NRC-HiQ containing after removing intersections with training data 53 and 49 complexes (csardock.org), and CSAR12 and CSAR14 sets downloaded from D3R (drugdesigndata.org) were prepared according to Jiménez et al.[33] Finally, we considered a bunch of data set serving as benchmarks[52,53] for the ensemble methods (FEP, MM-PB(GB)SA) developed for binding free energy estimation.

$$K_{\mathrm{i}} = \frac{\mathrm{IC}_{50}}{1 + \frac{[S]}{K_{\mathrm{m}}}} \tag{1}$$

3.2. Descriptors. The choice of the descriptor set which reflects the atomic environment in the relevant manner was influenced by a success of electron energy approximation by neural network potentials. The good representation of an atomic environment should be invariant to the permutation, rotational, reflection, and translation symmetries. It is worth mentioning that Behler–Parrinello symmetric functions (BPS)[54] made a basis for the first transferable NN potential and smooth overlap of atomic positions,[55] which defines a similarity metric for direct comparison of atomic environments. In this work, we employed BPS to describe the atomic environment in a binding site.

It is natural to prioritize local environment defining a cut-off function $f_c(r_{ij})$ (eq 2)[56] which smoothly decreases the weights of atoms outside the proximal environment and assigns the zero weight for atoms outside the cut-off distance. In the present work, a cutoff of 12 Å had been used. Table S2 (Supporting Information) lists the parameters defining BPS descriptors used in this study.

BPSF contains terms which depend only on the distance to the neighboring atoms and terms which are based on angles formed by all atom pairs in the environment and the central atom as an angle vertex. Functions with the radial symmetry are constructed as follows (eq 3) with the sum of Gaussian functions. Their role is to indicate the existence of certain atom approximately at distance r_s.

$$f_{\mathrm{c}}(r_{ij}) = \begin{cases} \tan \mathrm{h}^{3}\left(1 - \frac{r_{ij}}{r_{\mathrm{c}}}\right) & r_{ij} \le r_{\mathrm{c}} \\ 0 & r_{ij} > r_{\mathrm{c}} \end{cases} \tag{2}$$

$$G_i^{\,2} = \sum_{i \neq j}^{\mathrm{all}} \mathrm{e}^{-\eta (r_{ij} - r_s)^2} f_{\mathrm{c}}(r_{ij}) \tag{3}$$

It is to be noted that that radial symmetry functions take into account only pair-wise atom interactions. We illustrated the $G_i^{\,2}$ computation procedure as a simple example (Figures 2, 3, 5 and 6) which shows nicotinic acid amide schematically surrounded by three amino acid residues (Ser, Lys, and Ile), and the descriptors are computed for its oxygen atom. The BPS functions are denoted by concentric circles with decreasing intensity which is caused by the f_c function application. The bar plot on the left part of Figure 2 shows the number of atoms which contribute to the $G^{i}_{\mathrm{O-amide}}$ values for different r_s parameter values; taking into account triplewise interactions, we compute the angular dependent function (eq 4)

$$G_i^{\,3} = 2^{1-\zeta} \sum_{j,k \neq i}^{\mathrm{all}} (1 + \lambda \cos \theta_{jik})^{\zeta} \mathrm{e}^{-\eta (r_{ij}^{\;2} + r_{jk}^{\;2} + r_{ik}^{\;2})} f_{\mathrm{c}}(r_{ij}) f_{\mathrm{c}}(r_{ik}) \\ f_{\mathrm{c}}(r_{jk}) \tag{4}$$

where r_{ij} corresponds to the distance between i and j atoms. All angles θ_{jik} are defined with atom i as the central one and atoms j and k are the atoms from the environment. The η value control the guassian sharpness, while the role of parameter ζ is to provide angular resolution. High ζ values produce a narrower range of nonzero symmetry function values. The parameter λ can take values +1 or −1 allows to shift the maximum value of the cosine part from π (+1) to 0 (−1) and

https://dx.doi.org/10.1021/acsomega.9b04162

describe the atomic environment in a better way. It should be noted that G_i^2 and G_i^3 represent two-body and three-body interactions and can be expanded to high-order terms, but they are not used in the current work.

The parameter set used to compute descriptors can be found in Table S2 (Supporting Information), and all possible combinations of parameters yielded 52 descriptors of the atomic environment. We defined several atom types representing the most common elements which can be found in the protein structure: C, O, N, S, P, M1, M2, where M1 and M2 represent single charged metal ions and metal ions in the other charged state, respectively. BPS computed for each atom type of the protein environment leads to 364 environmental descriptors and combined with one-hot encoded ligand atom type gives 373 descriptors calculated for each atom of the ligand molecule. Ligand atom types (C, O, N, S, P, F, Cl, Br, I) were selected by their relative occurrence in the data set. Hydrogen atoms were ignored to reduce the memory requirements. It should be noted that boron atoms were considered to be included in this set, but the majority of boron-containing ligands contain carborane substructures difficult to describe using the standard valence model or boronic acids which usually form a covalent bond with certain protein atoms.

3.3. Neural Network Architecture.

Chemical structures are naturally represented as undirected graphs, where nodes and edges correspond to atoms and bonds, respectively. Recently Gilmer et al. designed the MPNN framework,[57] that operates with chemical graphs, and which is invariant to graph isomorphism.[28,41,58] In this study, we consider a chemical graph G with node features x_v and edge features e_{vw} where v and w are node indexes. According to Gilmer et al., the forward pass consists of two main stages: (i) the message passing phase and (ii) the readout phase. The message passing phase can be divided into T stages, which are performed sequentially, and at each time step, two functions are carried out on the graph elements: message function (M_t) and update function (U_t). The message and update functions are learned differentiable functions with fixed length input and output. To perform the message phase, first, for each node v of graph G, we select neighboring nodes of v and denote them $N(v)$. Then, for each pair v and w, where $w \in N(v)$, we concatenate two node descriptor vectors h_v^t and h_w^t with edge descriptor vector e_{vw}, and the obtained vector of fixed length becomes an input for a message function M_t. Then, we summarize all these outputs (eq 5), yielding the m_v^{t+1} vector of the fixed length finishing the message phase. The update phase is performed for each node v as a result of application of an update function (eq 6) to the concatenation of the hidden state vector h_v^t and the newly computed message vector m_v^{t+1}.

$$m_v^{t+1} = \sum_{w \in N(v)} M_t(h_v^t, h_w^t, e_{vw}) \tag{5}$$

$$h_v^{t+1} = U_t(h_v^t, m_v^{t+1}) \tag{6}$$

$$\hat{y} = R(\{h_v^t | v \in G\}) \tag{7}$$

One can imagine this iterative process as information flooding across the graph from node to node. It should be noted that in our implementation vector, h_v^1 is initialized as the BPS node descriptor vector calculated to represent the atomic environment. Vector e_{vw} is added for all time steps without changes. The readout phase consists of an application of the readout function (eq 7) to a set of hidden states h_v^t obtained at the final update step yielding the target variable. The readout function is constructed to be invariant to the node permutations which make the designed MPNN invariant to chemical graph isomorphisms. The simplest way to achieve this property is to summarize all hidden state vectors reducing N (number of nodes) vectors of length L (number of resulting features) to one vector of length L. Unfortunately, this approach leads to the significant information loss. That is why we followed Kearnes et el.[41] and applied a fuzzy histogram approach[59] to capture the distribution of each L features.

To construct histograms, we apply a set of membership functions of length equals to the number of predefined bins to the data. Each membership function returns one, if the data element is in the current bin and zero otherwise. For the fuzzy histogram approach, the normalized Gaussian membership function (eq 8) was used where i is a bin index and K is the number of bins, respectively. Each fuzzy membership function is defined by the bin center x_i where x denotes the current data element. In this work, 11 fuzzy membership functions centered at −2.75, −2.0, −1.35, −0.8, −0.35, 0, 0.35, 0.8, 1.35, 2.0, and 2.75 were used with all σ_i^2 equals to 0.5. Then, the summation performed over all nodes yields 11 × L permutationally invariant descriptors.

$$\mathrm{GM}_i = \frac{e^{-(x-x_i)^2/\sigma_i^2}}{\sum_{1 \le i \le K} e^{-(x-x_i)^2/\sigma_i^2}} \tag{8}$$

The choice of message, update, and readout functional forms were inspired by Battaglia et al.,[58] where each of the function is a multilayer fully connected perceptron with a specific architecture defined in Table 1. Batch normalization[60] was applied for each layer of all neural networks except the output layer. We found that the dropout technique significantly increased the training time; thus, we did not use it to obtain the final model.

3.4. Error Metrics and Training Details.

We used the *Pytorch 0.4* framework for DNN training and *networkx 2.1* and *rdkit 2018.03.1* for molecular graph processing and chemoinformatics routines. It should be noted that reading of some sdf files in database yields an error by rdkit. The most common issue is the lack of positive charge on tertiary amine nitrogen atoms. These structures were corrected manually using MarvinSketch 18.10 (http://www.chemaxon.com). The detailed description of the training procedure is given in the Supporting Information. Because scoring function training was reformulated as a multi task learning problem, we should describe in more detail the loss function and quality metrics. The loss function is the modified MSE loss (eq 9), where N and T are the number of complex in a batch multiplied by two (for both pIC_{50} and pK predictions) and the number of available activities for the batch, respectively. The root-mean-square error (RMSE) and the mean average error, as well as the Pearson correlation coefficient (r) and the Spearman rank correlation coefficient (ρ) were computed for the performance comparison.

$$L(\hat{y}, y) = \sum_{i=1}^{N} \begin{cases} \frac{1}{T}(\hat{y} - y)^2 & y \text{ is not } NaN \\ 0 & y \text{ is } NaN \end{cases} \tag{9}$$

It was stressed in the literature that the usage of the train and test split provided by the PDBbind tends to provide too optimistic results.[61,62] Thus, we performed fivefold cross-validation, selecting the best pool of models and subsequently assessing their performance on the selected test sets by averaging the prediction results from all of the five models. Our scoring model is available at http://mpnn.syntelly.com.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.9b04162.

> Training details, basic molecular properties of the training sets, results of prediction for data sets, learning curves for the refined set and refined set + IC_{50} data, results of graphDelta evaluation on the data set used for FEP and MM-PBSA evaluation; and the parameter set used to compute BPS functions (PDF)

AUTHOR INFORMATION

Corresponding Authors

Petr Popov − *Skolkovo Institute of Science and Technology, Moscow 143026, Russia; Moscow Institute of Physics and Technology, Dolgoprudny 141701, Russia*; Email: p.popov@skoltech.ru

Maxim V. Fedorov − *Skolkovo Institute of Science and Technology, Moscow 143026, Russia; Skolkovo Innovation Center, Syntelly LLC, Moscow 143026, Russia; University of Strathclyde, Glasgow UK G4 0NG, U.K.*; Email: m.fedorov@skoltech.ru

Authors

Dmitry S. Karlov − *Skolkovo Institute of Science and Technology, Moscow 143026, Russia*; orcid.org/0000-0002-7194-1081

Sergey Sosnin − *Skolkovo Institute of Science and Technology, Moscow 143026, Russia; Skolkovo Innovation Center, Syntelly LLC, Moscow 143026, Russia*; orcid.org/0000-0002-3042-7369

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.9b04162

Notes

The authors declare the following competing financial interest(s): Maxim Fedorov and Sergey Sosnin are co-founders of Syntelly LLC.

ACKNOWLEDGMENTS

P.P. acknowledges Russian Science Foundation (RSF) research grant 18-74-00117. The results of the work were obtained using computational resources of MCC NRC ”Kurchatov Institute, http://computing.nrcki.ru/.

ABBREVIATIONS

MPNN, message passing neural network; NN, neural networks; DNN, deep neural networks; FEP, free energy perturbation; BPS, Behler−Parrinello symmetric function; SOAP, smooth overlap of atomic positions

REFERENCES

(1) Rehman, S. U.; Sarwar, T.; Husain, M. A.; Ishqi, H. M.; Tabish, M. Studying non-covalent drug−DNA interactions. *Arch. Biochem. Biophys.* **2015**, *576*, 49−60.

(2) Ain, Q. U.; Aleksandrova, A.; Roessler, F. D.; Ballester, P. J. Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2015**, *5*, 405−424.

(3) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.

(4) Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening. *J. Med. Chem.* **2004**, *47*, 1750−1759.

(5) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267*, 727−748.

(6) Verkhivker, G.; Appelt, K.; Freer, S. T.; Villafranca, J. E. Empirical free energy calculations of ligand-protein crystallographic complexes. I. Knowledge-based ligand-protein interaction potentials applied to the prediction of human immunodeficiency virus 1 protease binding affinity. *Protein Eng., Des. Sel.* **1995**, *8*, 677−691.

(7) Muegge, I. PMF Scoring Revisited. *J. Med. Chem.* **2006**, *49*, 5895−5902.

(8) Krammer, A.; Kirchhoff, P. D.; Jiang, X.; Venkatachalam, C. M.; Waldman, M. LigScore: a novel scoring function for predicting binding affinities. *J. Mol. Graphics Modell.* **2005**, *23*, 395−407.

(9) Jain, A. N. Scoring noncovalent protein-ligand interactions: A continuous differentiable function tuned to compute binding affinities. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 427−440.

(10) Ballester, P. J.; Mitchell, J. B. O. A machine learning approach to predicting protein−ligand binding affinity with applications to molecular docking. *Bioinformatics* **2010**, *26*, 1169−1175.

(11) Kadukova, M.; Grudinin, S. Convex-PL: a novel knowledge-based potential for protein-ligand interactions deduced from structural databases using convex optimization. *J. Comput.-Aided Mol. Des.* **2017**, *31*, 943−958.

(12) 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.

(13) McGaughey, G. B.; Sheridan, R. P.; Bayly, C. I.; Culberson, J. C.; Kreatsoulas, C.; Lindsley, S.; Maiorov, V.; Truchon, J.-F.; Cornell, W. D. Comparison of Topological, Shape, and Docking Methods in Virtual Screening. *J. Chem. Inf. Model.* **2007**, *47*, 1504−1519.

(14) Dittrich, J.; Schmidt, D.; Pfleger, C.; Gohlke, H. Converging a Knowledge-Based Scoring Function: DrugScore2018. *J. Chem. Inf. Model.* **2019**, *59*, 509−521.

(15) Ewing, T. J. A.; Makino, S.; Skillman, A. G.; Kuntz, I. D. DOCK 4.0: Search strategies for automated molecular docking of flexible molecule databases. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 411−428.

(16) Bishop, C. M. *Pattern Recognition and Machine Learning*; Springer, 2006.

(17) Li, H.; Leung, K.-S.; Wong, M.-H.; Ballester, P. J. Improving AutoDock Vina Using Random Forest: The Growing Accuracy of Binding Affinity Prediction by the Effective Exploitation of Larger Data Sets. *Mol. Inf.* **2015**, *34*, 115−126.

(18) Berishvili, V. P.; Voronkov, A. E.; Radchenko, E. V.; Palyulin, V. A. Machine Learning Classification Models to Improve the Docking-based Screening: A Case of PI3K-Tankyrase Inhibitors. *Mol. Inf.* **2018**, *37*, 1800030.

(19) Abel, R.; Wang, L.; Harder, E. D.; Berne, B. J.; Friesner, R. A. Advancing Drug Discovery through Enhanced Free Energy Calculations. *Acc. Chem. Res.* **2017**, *50*, 1625−1632.

(20) Genheden, S.; Ryde, U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opin. Drug Discovery* **2015**, *10*, 449−461.

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

(22) Blaschke, T.; Olivecrona, M.; Engkvist, O.; Bajorath, J.; Chen, H. Application of Generative Autoencoder in De Novo Molecular Design. *Mol. Inf.* **2018**, *37*, 1700123.

(23) Karlov, D. S.; Sosnin, S.; Tetko, I. V.; Fedorov, M. V. Chemical space exploration guided by deep neural networks. *RSC Adv.* **2019**, *9*, 5151−5157.

(24) Rawat, W.; Wang, Z. Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review. *Neural Comput.* **2017**, *29*, 2352−2449.

(25) Liu, P.; Qiu, X.; Huang, X. Recurrent Neural Network for Text Classification with Multi-Task Learning. **2016**, arXiv:1605.05101.

(26) LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436.

(27) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci.* **2017**, *8*, 3192−3203.

(28) Schütt, K. T.; Arbabzadah, F.; Chmiela, S.; Müller, K. R.; Tkatchenko, A. Quantum-chemical insights from deep tensor neural networks. *Nat. Commun.* **2017**, *8*, 13890.

(29) Popova, M.; Isayev, O.; Tropsha, A. Deep reinforcement learning for de novo drug design. *Sci. Adv.* **2018**, *4*, No. eaap7885.

(30) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4*, 268−276.

(31) Sosnin, S.; Misin, M.; Palmer, D.; Fedorov, M. 3D matters! 3D-RISM and 3D convolutional neural network for accurate bioaccumulation prediction. *J. Phys.: Condens. Matter* **2018**, *30*, 32LT03.

(32) Sosnin, S.; Karlov, D.; Tetko, I. V.; Fedorov, M. V. Comparative Study of Multitask Toxicity Modeling on a Broad Chemical Space. *J. Chem. Inf. Model.* **2019**, *59*, 1062−1072.

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

(34) 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.

(35) Golkov, V.; Skwark, M. J.; Mirchev, A.; Dikov, G.; Geanes, A. R.; Mendenhall, J.; Meiler, J.; Cremers, D. 3D Deep Learning for Biological Function Prediction from Physical Fields. **2017**, arXiv:1704.04039.

(36) Koes, D. R.; Baumgartner, M. P.; Camacho, C. J. Lessons Learned in Empirical Scoring with smina from the CSAR 2011 Benchmarking Exercise. *J. Chem. Inf. Model.* **2013**, *53*, 1893−1904.

(37) Wallach, I.; Dzamba, M.; Heifets, A. AtomNet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-based Drug Discovery. **2015**, arXiv:1510.02855.

(38) Jiménez-Luna, J.; Pérez-Benito, L.; Martínez-Rosell, G.; Sciabola, S.; Torella, R.; Tresadern, G.; De Fabritiis, G. DeltaDelta neural networks for lead optimization of small molecule potency. *Chem. Sci.* **2019**, *10*, 10911−10918.

(39) Bronstein, M. M.; Bruna, J.; LeCun, Y.; Szlam, A.; Vandergheynst, P. Geometric Deep Learning: Going beyond Euclidean data. *IEEE Signal Process. Mag.* **2017**, *34*, 18−42.

(40) Duvenaud, D.; Maclaurin, D.; Aguilera-Iparraguirre, J.; Gómez-Bombarelli, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional Networks on Graphs for Learning Molecular Fingerprints. **2015**, arXiv:1509.09292.

(41) Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P. Molecular graph convolutions: moving beyond fingerprints. *J. Comput.-Aided Mol. Des.* **2016**, *30*, 595−608.

(42) Feinberg, E. N.; Sur, D.; Wu, Z.; Husic, B. E.; Mai, H.; Li, Y.; Sun, S.; Yang, J.; Ramsundar, B.; Pande, V. S. PotentialNet for Molecular Property Prediction. *ACS Cent. Sci.* **2018**, *4*, 1520−1530.

(43) 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*, 1079−1093.

(44) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **2014**, *1*, 140022.

(45) van der Maaten, L.; Hinton, G. Visualizing Data Using T-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579−2605.

(46) Chupakhin, V.; Marcou, G.; Gaspar, H.; Varnek, A. Simple Ligand−Receptor Interaction Descriptor (SILIRID) for alignment-free binding site comparison. *Comput. Struct. Biotechnol. J.* **2014**, *10*, 33−37.

(47) Wójcikowski, M.; Zielenkiewicz, P.; Siedlecki, P. Open Drug Discovery Toolkit (ODDT): a new open-source player in the drug discovery field. *J. Cheminf.* **2015**, *7*, 26.

(48) Mobley, D. L. Let's get honest about sampling. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 93−95.

(49) Liu, Z.; Su, M.; Han, L.; Liu, J.; Yang, Q.; Li, Y.; Wang, R. Forging the Basis for Developing Protein−Ligand Interaction Scoring Functions. *Acc. Chem. Res.* **2017**, *50*, 302−309.

(50) Sosnin, S.; Vashurina, M.; Withnall, M.; Karpov, P.; Fedorov, M.; Tetko, I. A Survey of Multi-Task Learning Methods in Chemoinformatics. *Mol. Inf.* **2018**, *38*, 1800108.

(51) Li, H.; Leung, K.-S.; Wong, M.-H.; Ballester, P. Low-Quality Structural and Interaction Data Improves Binding Affinity Prediction via Random Forest. *Molecules* **2015**, *20*, 10947−10962.

(52) 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.

(53) Karlov, D. S.; Lavrov, M. I.; Palyulin, V. A.; Zefirov, N. S. MM-GBSA and MM-PBSA performance in activity evaluation of AMPA receptor positive allosteric modulators. *J. Biomol. Struct. Dyn.* **2018**, *36*, 2508−2516.

(54) Behler, J.; Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401.

(55) Bartók, A. P.; Kondor, R.; Csányi, G. On representing chemical environments. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2013**, *87*, 184115.

(56) Imbalzano, G.; Anelli, A.; Giofré, D.; Klees, S.; Behler, J.; Ceriotti, M. Automatic selection of atomic fingerprints and reference configurations for machine-learning potentials. *J. Chem. Phys.* **2018**, *148*, 241730.

(57) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural Message Passing for Quantum Chemistry. **2017**, arXiv:1704.01212.

(58) Battaglia, P.; Pascanu, R.; Lai, M.; Jimenez Rezende, D.; Kavukcuoglu, K. *Advances in Neural Information Processing Systems 29*; Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., Garnett, R., Eds.; Curran Associates, Inc., 2016; pp 4502−4510.

(59) Zadeh, L. A. Fuzzy Sets. *Inf. Control* **1965**, *8*, 338−353.

(60) Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. **2015**, arXiv:1502.03167.

(61) Gabel, J.; Desaphy, J.; Rognan, D. Beware of Machine Learning-Based Scoring FunctionsOn the Danger of Developing Black Boxes. *J. Chem. Inf. Model.* **2014**, *54*, 2807−2815.

(62) Kramer, C.; Gedeck, P. Leave-Cluster-Out Cross-Validation Is Appropriate for Scoring Functions Derived from Diverse Protein Data Sets. *J. Chem. Inf. Model.* **2010**, *50*, 1961−1969.