GB-Score: Minimally Designed Machine Learning Scoring Function Based on Distance-weighted Interatomic Contact Features

Milad Rayka and Rohoullah Firouzi*

Department of Physical Chemistry, Chemistry and Chemical Engineering Research Center of Iran, Tehran, Iran.

Corresponding Author:

* rfirouzi@ccerci.ac.ir and firouzi.chemist@yahoo.com

Abstract

In recent years, thanks to advances in computer hardware and dataset availability, data-driven approaches (like machine learning) have become one of the essential parts of the drug design framework to accelerate drug discovery procedures. Constructing a new scoring function, a function that can predict the binding score for a generated protein-ligand pose during docking procedure or a crystal complex, based on machine and deep learning has become an active research area in computer-aided drug design. GB-Score is a state-of-the-art machine learning-based scoring function that utilizes distance-weighted interatomic contact features, PDBbind-v2019 general set, and Gradient Boosting Trees algorithm to the binding affinity prediction. The distance-weighted interatomic contact featurization method used the distance between different ligand and protein atom types for numerical representation of the protein-ligand complex. GB-Score attains Pearson's correlation 0.862 and *RMSE* 1.190 on the CASF-2016 benchmark test in the scoring power metric. GB-Score's codes are freely available on the web at https://github.com/miladrayka/GB_Score.

Keywords: Molecular docking, Scoring function, Machine learning, Gradient-boosting trees, Scoring power, CASF-2016

1 - Introduction

Ligand-protein docking is one of the widespread tools in computer-aided drug design (CADD), which is used extensively to distinguish potent drug in large molecular libraries[1]. Generally, the docking procedure is constituted of two interconnected steps: pose identification and pose scoring. Various well-known and successful techniques, such as evolutionary and Monte Carlo algorithms, have been proposed to generate poses similar to the actual crystal structures[2-3]. In the scoring step, binding affinity for a specific ligand-protein pose is estimated with a scoring function, which is later used to discriminate between the active and the inactive ligands[4]. The generated poses during the docking procedure have accepted similarity to the native complex structure but attributed binding affinity values by scoring function are imprecise, which leads to undermining the overall docking performance.[5].

A robust scoring function should have four characteristics: scoring, ranking, docking, and screening abilities. Scoring comprises deriving a linear correlation between predicted and experimental binding affinity. Ranking entails arranging the known ligands of a given target protein by their binding affinities. Docking and screening represent the capability of scoring function in identifying the native ligand pose and true binders among decoys and random molecules[6]. In general, most conventional scoring functions (force-field, knowledge, and empirical-based) perform well in docking and screening evaluations but show poor performance in scoring and ranking. This deficiency is assumed associated with the linear regression method, which is utilized in designing these scoring functions. Therefore, one way for improving scoring functions is using approaches based on nonlinear regression and fitting[7-8].

Over the past decade, data-driven approaches, specifically machine and deep learning, have been successfully applied in the drug discovery[9]. Ligand- and structure-based drug discovery exploit this brand-new knowledge and assimilate it to their traditional algorithms to circumvent the well-known obstacles and enhance their performance and capability[10-12]. One burgeoning field of employing machine and deep learning is devising new scoring functions, which harness the nonlinear regression ability of these algorithms[7-8].

In recent decades, various machine learning algorithms are employed for designing data-driven scoring functions. Among these methods, Random Forest[13] (RF), Extremely Randomized Trees[14] (ERT), and Gradient Boosting Trees[15] (GBT) present superior performance over other algorithms like Support Vector Machine[8]. RF-Score[16], $\Delta_{vina}RF_{20}$,[17] the combination of RF-Score v3 and ligand-based features[18], and RI-Score[19] are examples of scoring functions adopted RF for their training algorithm. As examples of GBT-based scoring functions, we can mention TopBP-ML[20], EIC-Score[21], AGL-Score[22], ECIF-GBT[23], and ECIF::LD-GBT[23], which the latter uses Extended Connectivity Interaction Fingerprint (ECIF) and ligand descriptors as features and achieves the highest scoring power on CASF-2016 benchmark. Very recently, a new scoring function called ET-Score[24] has been introduced, which employs the distance-weighted interatomic contact between atom type pairs of the ligand and the protein for featurizing protein-ligand complexes and ERT algorithm for the training process.

Feedforward, convolutional, and graph neural networks are the prominent deep learning architectures, which applied for structure-based protein-ligand binding affinity predictions. Feedforward neural network (FFNN) is composed of several layers, each of which consists of a finite number of neurons. Data input should be converted to a numerical 1-dimensional array. During learning, parameters of FFNN are optimized so that FFNN can predict the target value of unseen data precisely[25]. NNScore[26-27], BgN-Score[28], and very recently Zhu et al[29] paper are among machine and deep learning-based scoring functions, which take advantage of FFNN as a learning algorithm. Convolutional neural network (CNN) is one of the deep learning methods which frequently applied in the computer vision field. In contrast to FFNN, CNN's inputs can be 2 or 3-dimensional grids[25]. So 3-dimensional molecular structure with minimum feature engineering can be employed as inputs for training CNN models. K_{Deep}[30], Pafnucy[31], OnionNet[32], and RosENet[33] are the most famous scoring functions which utilize CNN in their construction. Graph neural network[34] (GNN), specifically message passing neural network (MPNN), gets attention in recent years and becomes one of the most promising neural networks in the chemistry discipline. In MPNN or GNN each input data, i.e. molecule, is represented as a graph in which nodes of the graph are atoms and the bonds between them are considered as edges[35]. PotentialNet[36] and graphDelta[37] can be considered as MPNN scoring functions.

Here, we improve our previous scoring function, ET-Score, by employing better feature selection, an expanded train set, and different learning algorithms. Also, we scrutinize our new generated scoring function, GB-Score, in rigorous circumstances to assess its' generalization ability in the unseen data points. Section 2

provides information about used data sets for training and testing, feature generating based on distance-weighted interatomic contact concept, and a concise description of three learning algorithms. Section 3 consists of 6 parts, which in parts 1 to 3 scoring power of GB-Score is examined in different conditions. Subsections 3-4 and 3-5 deal with analyzing the test set and the importance of used features closely. In subsection 3-6, GB-Score compared to other scoring functions. The last section is devoted to conclusions.

2 - Methods

2-1- Dataset

PDBbind dataset is one of the most widely used datasets in predicting the protein-ligand binding affinity discipline. The PDBbind dataset is composed of protein-ligand complexes whose binding affinities and structures are determined using experimental techniques. The binding affinity values are expressed based on $-\log K_i$, $-\log K_d$, and $-\log IC_{50}$, also protein and ligand structures are saved separately to pdb and mol2 (or sdf) file formats, respectively. Because the PDBbind dataset is updated annually, in this report, we utilized the 2016 and 2019 versions of this dataset for training, validating, and verifying our proposed models. Both versions consist of three sets: general, refined, and core sets. The core set contains 285 high quality protein-ligand structures, which for both versions are the same. Conventionally, the core set is assigned as an external test set for verifying the performance of proposed models, so its complexes have to be excluded from the training set ("hard overlap" exclusion). The quality of structures in the refined and the general sets are lower than the core set, so they are usually used in the training procedure. The refined set 2016, the composition of the refined and general sets of each PDBbind version 2016 and 2019 containing 3772, 12988, and 17366 protein-ligand structures, respectively, are used as training sets[6, 38-39].

In recent years, it has been speculated that the superiority of machine learning-based scoring function is related to the similarity between train and test sets. Recently, Su et al.[40] has attempted to circumvent this obstacle by designing series of new train sets in which "soft overlap" between train and test sets is decreased. This overlap arises from the similarity between proteins, binding sites, and ligands in train and test sets. If for two complexes in the train and the test sets, the aforementioned similarities are above a pre-defined threshold, the complex in the train set is removed. After doing the same procedure iteratively, series of non-redundant train sets are produced.[40] Here, we used new train sets that are compiled of the refined set 2016 with 80%, 85%, 90%, and 95% similarities thresholds.

One of the main challenges in machine learning applications is the failure of the trained model to extrapolate to out-of-distribution data. To investigate the capability of our proposed model in out-of-distribution data, we added the core set structures to the general and the refined sets 2019 then we devised three nonidentical train and test sets by excluding all HIV-1 Protease, Trypsin, and Carbonic Anhydrase from the gathered above data. Details of all aforementioned train and test sets can be found in Table 1.

Table 1- Details of all train and test sets. N_f is the dimension of the generated feature vector by distance-weighted interatomic contacts between ligand and protein atoms method, which during the preprocessing step, all static, quasi-static (variance below 0.01), and correlated (correlation above 95%) features were eliminated.

Name	Train size	Test size	N_f
Refined set (2016)	3772	285	93
General set + Refined set (2016)	12988	285	101
General set + Refined set (2019)	17366	285	104
Refined set 80% (2016)	2105	285	93
Refined set 85% (2016)	2562	285	93
Refined set 90% (2016)	3054	285	93
Refined set 95% (2016)	3570	285	93
HIV-1 Protease	17352	299	104
Trypsin	17429	222	104
Carbonic Anhydrase	17205	446	104

2-2 Feature generation

Recently, we have shown that distance-weighted interatomic contact between ligand and protein atoms can be utilized as features for the mathematical representation of protein-ligand complexes in the machine learning procedure[24]. Besides invariance property, generated features are unique, compact, and computationally affordable which are compatible with Himanena et al.[41] description of ideal features. As before, we considered element-based atom types (H, C, N, O, F, P, S, Cl, Br, I) for ligand. To generate protein atom types, we classified amino acid residues based on their chemical nature of side-chains into four groups (Charged (c), Polar (p), Amphipathic (a), Hydrophobic (h)):

Charged = {Arg, Lys, Asp, Glu}

Polar = {Gln, Asn, His, Ser, Thr, Cys}

Amphipathic = {Trp, Tyr, Met}

Hydrophobic = {Ile, Leu, Phe, Val, Pro, Gly, Ala}

Then the same element-based atom types attributed to each group. Via this process, the resulting protein atom types will reflect the local chemical environment of protein atoms. In the next step, all interatomic distances for a specific atom types pair are calculated. Distances with magnitude below the predefined cutoff (d_{cutoff}) are weighted by an inverse power of a natural number (n) and sum together. In our previous work, we demonstrated that 12 $\dot{A}$ and 2 are appropriate choices for d_{cutoff} and n, respectively[24]. The mentioned algorithm is repeated iteratively for all possible atom types pairs, and a feature vector with 400 dimensions as a representation of a protein-ligand complex is produced[24]:

$$\vec{X} = \left\{X_{H,H_p}, X_{H,C_p}, \dots, X_{I,I_h}\right\}$$

$$X_{i,j} = \sum_{k=1}^{K_j} \sum_{l=1}^{L_i} \frac{1}{d_{lk}^2}$$

where i and j are atom types of ligand and protein, respectively; L_i is the total number of ligand atoms of type i and K_j is the total number of protein atoms of type j, d_{lk} is the Euclidean distance between the l-th ligand atom of type i and the k-th protein atom of type j, which is less than 12 $\dot{A}$.

During the preprocessing step, all static, quasi-static (variance below 0.01), and correlated (correlation above 95%) features were eliminated, which led to different features dimension for different train sets. Also, normalization was applied to the remaining features due to their means and standard deviation. Dimensions of feature vectors (N_f) are represented in Table 1.

2-3 Machine learning algorithms

RF[13], ERT[14], and GBT[15] are among the most widespread machine learning algorithms employed in designing scoring functions[8]. These three algorithms belong to the ensemble learning method, whose objective is to create a model based on a diverse set of predictors. One way to attain this goal is to train each predictor on a different random subset of the train set. RF, ERT, and GBT consist of a collection of decision trees. In RF and ERT, bootstrap aggregating or bagging, a sampling method with the replacement, is employed to devise random subsets of data. Each decision tree is trained separately on the different train subsets and the prediction of the model on a new instance, in regression case, is the mean of all prediction values. Utilizing extra randomness by the ERT algorithm results in more diverse decision trees than RF and faster training. Unlike RF and ERT, GBT incorporates the boosting method to generate a diverse set of predictors. The general idea behind the boosting approach is to train predictors iteratively, each trying to

improve its predecessor[42-43]. Here, Scikit-learn machine learning package is used for training[44]. In RF and ERT number of trees in the forest (*n_estimators*) is set to 500 and only the m_{try} (*max_features*) hyperparameter optimizes concerning out-of-bag criteria[16]. In GBT case, all hyperparameters are set to the values in Sánchez-Cruz et al. paper[23]. All hyperparameters are aggregated in Table 1s. Because of the stochastic nature of the aforementioned algorithms, the training procedure repeats ten times, and the model's root mean square error (*RMSE*) and Pearson's correlation (R_p) are reported by averaging over ten models[6, 16].

3 Results and discussion

3-1 Scoring power CASF-2016

One way for the assessment of scoring function performance is using the CASF-2016 benchmark set. In the CASF-2016, a scoring function is evaluated by four metrics: 1- Scoring power 2- Ranking power 3- Docking power 4- Screening power[6]. Here, we only examine the scoring power of our proposed scoring function. In the scoring power, the capacity of a scoring function to predict binding affinity in a linear correlation with experimental data is evaluated. The core set 2016, as an external set with 285 high-quality crystal structures and experimental binding data, is used to evalute the scoring power.

As mentioned before, after excluding the core set, the refined set 2016, the composition of the refined and the general sets 2016, and the same composition for 2019 are used as the training sets. The distance-weighted interatomic contact featurization method was applied to protein-ligand complexes to generate a numerical representation for them[24]. RF, ET, and GBT were adopted as fast and standard learning algorithms to discern hidden patterns in the training data.

All results for different train sets are aggregated in Table 2. After fine-tuning m_{try}, both RF and ERT learning algorithms presented similar performances on the core set 2016 when different train sets were applied. GBT outperformed RF and ERT slightly on three train sets, specifically when the refined set 2016 was chosen as the train set (R_p improved by 15.8% and 21.9% with respect to RF and ERT). In the machine learning discipline, one approach for improving the performance of the proposed model is the increasing amount of available data[42]. A comparable trend can be noticed in Table 2. By increasing the number of data points from the refined set 2016 to the composition of refined and general sets 2019, R_p for RF, ERT, and GBT improved by 39%, 32.7%, and 28.6%, respectively. The best model was achieved by training GBT on the composition of refined and general sets 2019, which accomplished R_p and *RMSE*, 0.862 and 1.19, respectively. We call this fittest model "GB-Score". Figure 1 illustrates the correlation plot between predicted binding affinities by the GB-Score and experimentally determined values for the core set 2016.

Table 2- Performance of RF, ET, and GBT on the core set 2016. Standard deviation is shown in parenthesis.

ML / Train set	RF		ET		GBT	
	R_p	*RMSE*	R_p	*RMSE*	R_p	*RMSE*
Refined set (2016)	0.820 (0.002)	1.353 (0.003)	0.825 (0.001)	1.344 (0.002)	0.838 (0.002)	1.263 (0.004)
Refined set + general set (2016)	0.850 (0.001)	1.297 (0.004)	0.852 (0.001)	1.287 (0.003)	0.858 (0.002)	1.203 (0.006)
Refined set + general set (2019)	0.852 (0.001)	1.289 (0.002)	0.852 (0.001)	1.287 (0.003)	0.862 (0.001)	1.190 (0.004)

Figure 1 shows that the GB-Score has predicted false values for protein-ligand structures with $pK_{i/d}$ values more than 10. This observation can be attributed to the employed train set because only 1.80% of train set data have *pKa* of more than 10, therefore the GB-Score predictions are biased toward middle range $pK_{i/d}$ values. Increasing data instances with high $pK_{i/d}$ magnitudes can be applied as a solution.

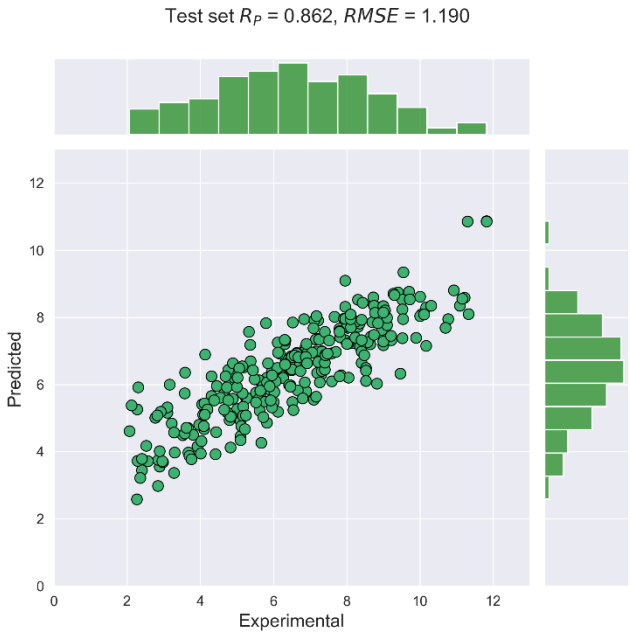

Figure 1- The correlation plot between predicted binding affinities by the GB-Score and experimental values for the core set 2016.

To provide more robust metrics, we devised a new model using GBT and 5-fold cross-validations over all PDBbind 2019v protein-ligand data. The test part holds approximately 3530 instances, which is almost 12.5 times more than the core set 2016. R_p and *RMSE*, in this case, are 0.764 (0.001) and 1.205 (0.007). The decline in model performance is reasonable because the test set size is larger and more diverse concerning the core set 2016.

3-2 Non-redundant train sets

As discussed earlier, one objection against machine learning-based scoring functions is due to the similarity between train and test sets. Therefore, the performance of the proposed scoring function has to be evaluated more rigorously by performing "leave-cluster-out" or composing multiple subsets by filtering out the data similar to those in the test set[45-46].

Lately, Su et al.[40] investigated this problem systematically by employing different machine learning algorithms and various subsets of data as training sets, which were devised by filtering out "hard overlap" and "soft overlap" for different similarity thresholds with respect to test set (Core set 2016). They concluded, even if a train set with low similarity to the test set is applied, some machine learning-based scoring functions achieve better scoring power, which can be attributed to the nonlinear and complex characteristics of the learning algorithms.

Here, Su et al.'s new sets, which are compiled of the refined set 2016 with 80%, 85%, 90%, and 95% similarities thresholds (similarity between protein sequences, binding sites, and ligands), are indicated as train sets and GBT employed as a learning algorithm. Table 3 shows all results for different ranges of sample size and similarity thresholds. Our model achieved acceptable performance in the four different train sets even in the 80% similarity threshold with only 2105 data points. This result allows us to conclude that the distance-weighted interatomic contact featurization and GBT algorithm are relatively robust with respect to decreasing sample size or filtering out "soft overlap".

Table 3- Shows all results for different ranges of sample size and similarity thresholds.

Non-redundant train set similarity (v.2016)	80%	85%	90%	95%
Sample size	2105	2562	3054	3570
R_p	0.729 (0.003)	0.781 (0.001)	0.816 (0.001)	0.838 (0.003)
RMSE	1.546 (0.005)	1.419 (0.002)	1.336 (0.001)	1.271 (0.008)

3-3 Family specific extrapolation

In order to estimate the extrapolation ability of our model, distinct train and test sets were constructed by excluding only the most populated protein subfamilies, which were detected using the search tool on PDBbind website, from the composition of refined and general sets 2019. Excluded subfamily constitutes the test set, and the rest assigns as the train set. HIV-1 Protease, Trypsin, and Carbonic Anhydrase are considered the most populated subfamilies[47]. Train and test set sizes are presented in Table 1 (theirs PDB Ids can be found on Github repository). Table 4 shows that the model can generalize its prediction ability to unseen cases, despite that its performance declines when all instances of a subfamily are excluded from the train set. A hypothetical explanation for this generalization can be attributed to similarities between three subfamilies and other protein-ligand complexes in training sets. The best performance happened at Trypsin subfamily with R_p 0.680 and *RMSE* 1.292. HIV-1 Protease test set with 299 structures is a challenging case in which our model achieved the lowest statistical metrics.

Table 4- Results of model's performance on excluded subfamilies.

Family	HIV-1 Protease	Trypsin	Carbonic Anhydrase
R_p	0.655	0.680	0.673
RMSE	1.335	1.292	1.304

3-4 Core set analysis

The core set 2016 is constituted of 57 clusters or protein families. Predicted R_p and *RMSE* values by GB-Score for each of these clusters are illustrated in Figures 2 and 1s, respectively (the R_p and *RMSE* numerical values can be found in Table 2s in the supporting information). In the Pearson's correlation case, we can notice, the GB-Score can predict trends within the majority of clusters properly, which R_p for 75.4% of all clusters shows an acceptable value (above 0.7). For all clusters, R_p values are positive except one cluster with target ID 33 (Integrase protein family), which presents negative Pearson's correlation (-0.57). Comparing R_p and *RMSE* values affirm which there is no distinct correlation among them (Spearman rank correlation 0.053). The GB-Score, with R_p's average 0.769 (median 0.870) across all clusters, outperforms newly reported scoring function AEScore in per-cluster analysis (average 0.67 and median 0.82)[48].

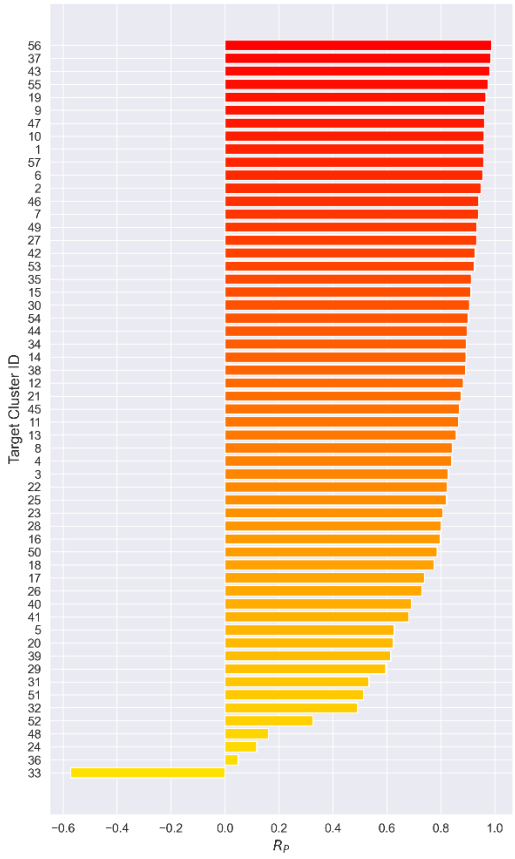

Figure 2- Per-cluster Pearson's correlation coefficient in the core set 2016.

3-5 Feature selection

GBT is a non-linear estimator. Therefore, to obtain the amount of importance of used features on predicted target values, we utilized permutation feature importance[13] (PFI) as a tool for inspecting our model. In PFI, the importance of each feature is quantified by estimating the decrease of model score after random shuffling of the given feature value. This procedure is repeated several times to estimate the mean and the standard deviation for all the features. Here, we adopted the default score of GBT in Scikit learn package[44] (i.e., the coefficient of determination of the prediction) and repeated mentioned procedure 10 times.

Only features whose means of importance are greater than two times their standard deviation are retained. Among 104 used features for numerical representation of protein-ligand complex, 76 features passed our criteria, which constitute 77.04 % of all importance (relevant details shown in Table 3s). The features belonging to hydrophobic amino acids have the most important contributions in binding affinity prediction, which shuffling them decreases model performance by 26.01%, nearly the sum of two charged and polar groups. Amphipathic, polar, and charged amino acid groups impact the model score with 22.46%, 16.45%, and 12.11%, respectively. Our results are almost compatible with our previous work on ET-Score[24], in where we measured feature importance using "mean decrease in impurity"[49].

Figure 3 provides information about the sum of feature importances for protein (right panel) and ligand (left panel) atom types. The features contain the Carbon atom type in both ligand and protein constitute 37.75% and 28.26% of all importances, respectively. This importance can be correlated to the Carbon element abundance in ligands and proteins. N and O are the next important features in the protein case (15% and 15.64%, respectively). The Hydrogen atom type in both protein and ligand has similar importance (both almost 11.8%). The numerical values for other atom types are presented in Table 4s.

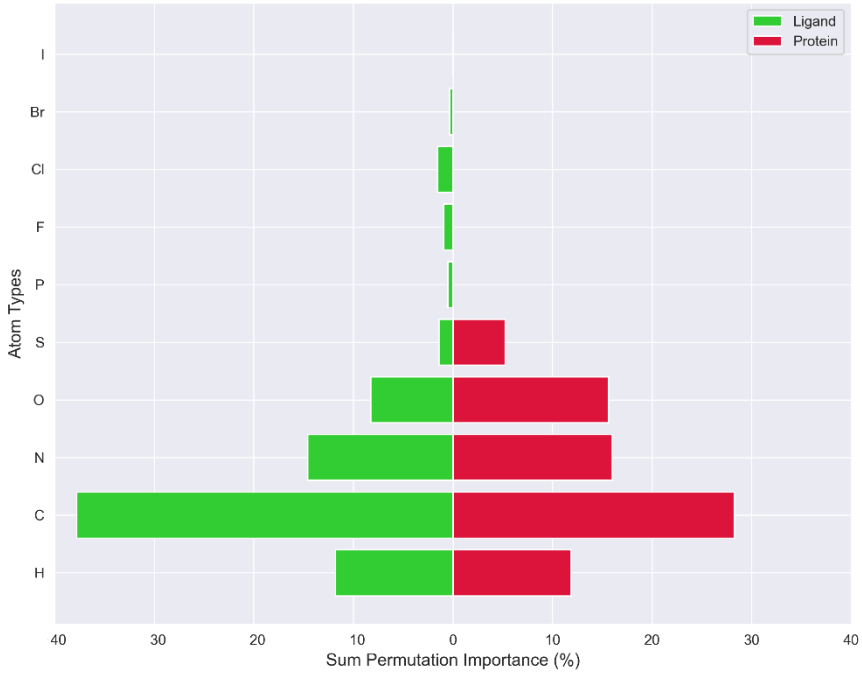

Figure 3 - Illustration of sum of permutation importance for different atom types of ligand (left panel) and protein (right panel).

3-6 Comparison

Figure 4 illustrates the scoring power comparison between the GB-Score (highlighted by red color) and other new scoring functions in terms of R_p. All scoring functions belong to ML-based, e.g. utilizing convolutional neural network or random forest, except X-Score[6], which is chosen as the best representation of the conventional scoring function. As we can clearly see all ML-based scoring functions outperformed X-Score in the scoring power benchmark on the core set 2016. Su et al.[40] demonstrate the superiority of ML-based scoring functions is associated with the non-linear and complex essence of ML algorithms. However, the mentioned ML-based scoring functions in Figure 4 managed different training sets to train their algorithms (different released versions of refined or general sets), which makes the comparison between them unclear and unfair.

All ML-based scoring functions achieve similar and comparable performances (performance metrics are only different in the second and third decimals) in binding score prediction, nevertheless, they used various kinds of feature engineering techniques. ECIF::LD-GBT[23] employed ECIF and ligand descriptors for representing protein-ligand complex and adopted GBT as a learning algorithm. AGL-Score[22] applied multiscale weighted labeled algebraic subgraphs for generating features. ET-Score[24] was developed to utilize distance-weighted interatomic contact features and ERT algorithm to binding affinity prediction. EIC-Score[21] is based on differential geometry representations of the protein-ligand complexes and GBT algorithm. RosENet[33] descriptors employ molecular mechanics energies from Rosetta force field and voxelized representation of protein-ligand complex through CNN architecture. K_{DEEP}[30] as a CNN-based scoring function considers eight pharmacophoric-like properties for featurizing the complex via a voxelized representation of the binding site. PLEC-nn[50], a neural network-based scoring function, represents pairing between ligand and protein atoms and their environment according to hashed fingerprint. OnionNet[32] applied similar features like RF-Score[16], element pair-specific contacts between ligands and protein atoms, but grouped them into different distance ranges to construct grid shape input for its CNN. $\Delta_{vina}RF_{20}$ [17]used RF algorithm and various Autodock Vina[3] energy terms and some molecular descriptors. RI-Score[19] constructed its features based on rigidity index descriptors.

As mentioned above, ML-based scoring functions employed different training sets during training procedures, which makes the comparison ambiguous, however, ECIF::LD-GBT achieves the best scoring power (R_p = 0.866), and our suggested scoring function, GB-Score, stands in the second rank with a slight departure to the first rank scoring function (R_p = 0.862).

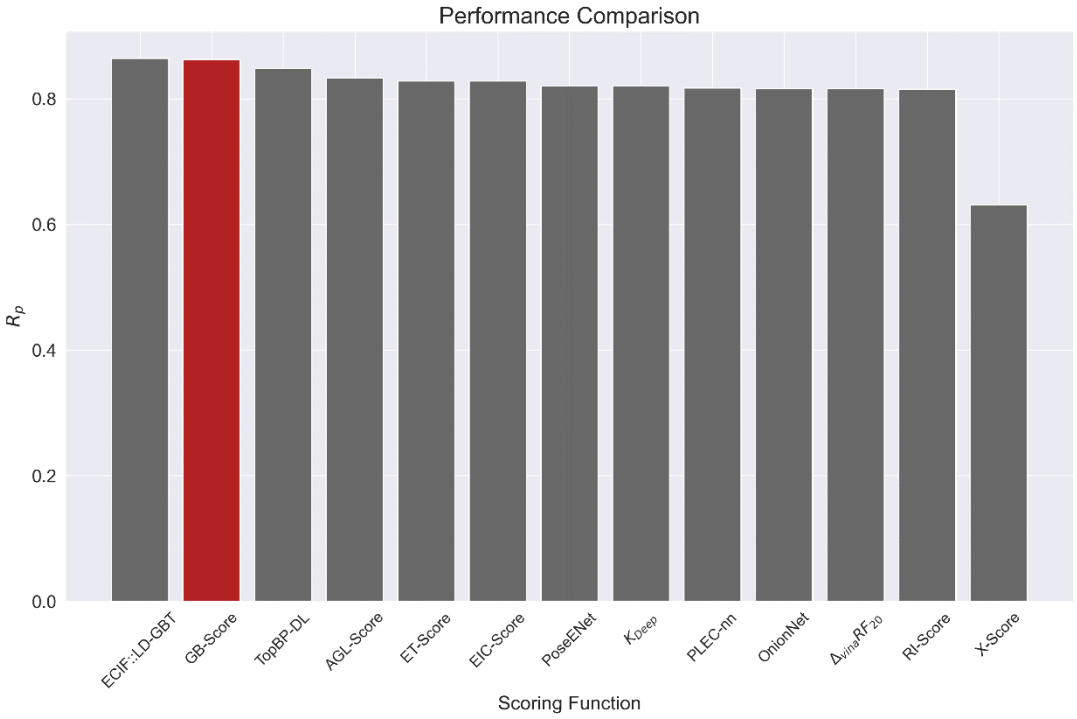

Figure 4- Performance scoring power comparison of the GB-Score with different scoring functions on the core set 2016 in terms of R_p.

4 - Conclusion

In this report, we improve our previous scoring function, ET-Score[24], in several steps to generate GB-Score as a new binding affinity estimator. In the first step, we scrutinized distance-weighted interatomic contact features to eliminate correlated, static, and quasi-static features (section 2.2). Through this step, our features reduced from 189 in ET-Score case to 104 in GB-Score. Thus, we built a simpler model to predict target value than ECIF::LD-GBT with 1710 features which is compatible with Occam's razor principle[43]. Recently developed ML-based scoring functions demonstrated using GBT and the general set as learning algorithm and the training set, respectively, are more suitable in binding score prediction and achieved better performance in the benchmark. As expected, by employing two mentioned enhancements and applying a larger train set, GB-Score accomplished comparable performance to other recently released ML-based scoring functions in the CASF-2016 benchmark test in the scoring power metric with *Rp* and *RMSE*, 0.862, and 1.190, respectively. We investigated GB-Score scoring power capacity by employing newly developed train sets by Su et al.[40], which "soft overlap" between train and the core set 2016 reduced through different similarity thresholds. In this situation, GB-Score achieves acceptable performance (R_p above 0.7). Furthermore, three distinct train sets were devised by excluding HIV-1 Protease, Trypsin, and Carbonic Anhydrase from the collection of the general and refined sets to form test sets for verifying GB-Score capability in the out-of-distribution domain. Although its performance was undermined, in the three mentioned test sets, it performed satisfactorily. The permutation importance technique was used to look under the hood of our black box model for attaining the importance of our proposed features.

In the near future, to thoroughly examining GB-Score and also other ML-based scoring functions, two open problems need to be addressed. The first problem consists of inquiring about the scoring function in the real scenario to verify its capacity in the docking and screening powers to distinguish between wild and decoy poses. It is a challenging problem in developing scoring functions based on ML. One proposed explanation

for this undermining can be related to the unavailability of docking poses in the training set[51]. The second important problem is considering better assessment of ML-based scoring functions through uncertainty quantification[52-53] and domains of applicability[54] because all newly proposed scoring functions are converging to the same performance, which makes them indistinguishable. So analyzing model test error, its uncertainty across test sets, we can spot sub-domains in which different models perform better than the others and practice this assessment for further model comparison. We hope to publish our results on these two challenges in the near future.

5 - Software and data availability

All Python codes, PDB IDs for all training and test sets, and Jupyter notebook for repeating this report are provided in GitHub (https://github.com/miladrayka/GB_Score).

6 - References

1. Lyu, J.; Wang, S.; Balius, T. E.; Singh, I.; Levit, A.; Moroz, Y. S.; O'Meara, M. J.; Che, T.; Algaa, E.; Tolmachova, K., Ultra-large library docking for discovering new chemotypes. *Nature* **2019,** *566* (7743), 224-229.
2. Koes, D. R.; Baumgartner, M. P.; Camacho, C. J., Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise. *Journal of chemical information and modeling* **2013,** *53* (8), 1893-1904.
3. Trott, O.; Olson, A. J., AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry* **2010,** *31* (2), 455-461.
4. Mobley, D. L.; Dill, K. A., Binding of small-molecule ligands to proteins:"what you see" is not always "what you get". *Structure* **2009,** *17* (4), 489-498.
5. Waszkowycz, B.; Clark, D. E.; Gancia, E., Outstanding challenges in protein–ligand docking and structure- based virtual screening. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2011,** *1* (2), 229-259.
6. Su, M.; Yang, Q.; Du, Y.; Feng, G.; Liu, Z.; Li, Y.; Wang, R., Comparative assessment of scoring functions: the CASF-2016 update. *Journal of chemical information and modeling* **2019,** *59* (2), 895-913.
7. Liu, J.; Wang, R., Classification of current scoring functions. *Journal of chemical information and modeling* **2015,** *55* (3), 475-482.
8. Shen, C.; Ding, J.; Wang, Z.; Cao, D.; Ding, X.; Hou, T., From machine learning to deep learning: Advances in scoring functions for protein–ligand docking. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2020,** *10* (1), e1429.
9. Peña-Guerrero, J.; Nguewa, P. A.; García-Sosa, A. T., Machine learning, artificial intelligence, and data science breaking into drug design and neglected diseases. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2021**, e1513.
10. Yang, S. Q.; Ye, Q.; Ding, J. J.; Lu, A. P.; Chen, X.; Hou, T. J.; Cao, D. S., Current advances in ligand-based target prediction. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2021,** *11* (3), e1504.
11. Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T., The rise of deep learning in drug discovery. *Drug discovery today* **2018,** *23* (6), 1241-1250.
12. Mak, K.-K.; Pichika, M. R., Artificial intelligence in drug development: present status and future prospects. *Drug discovery today* **2019,** *24* (3), 773-780.
13. Breiman, L., Random forests. *Machine learning* **2001,** *45* (1), 5-32.
14. Geurts, P.; Ernst, D.; Wehenkel, L., Extremely randomized trees. *Machine learning* **2006,** *63* (1), 3-42.
15. Friedman, J. H., Greedy function approximation: a gradient boosting machine. *Annals of statistics* **2001**, 1189-1232.
16. Ballester, P. J.; Mitchell, J. B., A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. *Bioinformatics* **2010,** *26* (9), 1169-1175.
17. Wang, C.; Zhang, Y., Improving scoring-docking-screening powers of protein–ligand scoring functions using random forest. *Journal of computational chemistry* **2017,** *38* (3), 169-177.

18. Boyles, F.; Deane, C. M.; Morris, G. M., Learning from the ligand: using ligand-based features to improve binding affinity prediction. *Bioinformatics* **2020,** *36* (3), 758-764.
19. Nguyen, D. D.; Xiao, T.; Wang, M.; Wei, G.-W., Rigidity strengthening: A mechanism for protein–ligand binding. *Journal of chemical information and modeling* **2017,** *57* (7), 1715-1721.
20. Cang, Z.; Mu, L.; Wei, G.-W., Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening. *PLoS computational biology* **2018,** *14* (1), e1005929.
21. Nguyen, D. D.; Wei, G. W., DG-GL: Differential geometry-based geometric learning of molecular datasets. *International journal for numerical methods in biomedical engineering* **2019,** *35* (3), e3179.
22. Nguyen, D. D.; Wei, G.-W., AGL-Score: Algebraic graph learning score for protein–ligand binding scoring, ranking, docking, and screening. *Journal of chemical information and modeling* **2019,** *59* (7), 3291-3304.
23. Sánchez-Cruz, N.; Medina-Franco, J. L.; Mestres, J.; Barril, X., Extended connectivity interaction features: Improving binding affinity prediction through chemical description. *Bioinformatics* **2021,** *37* (10), 1376-1382.
24. Rayka, M.; Karimi-Jafari, M. H.; Firouzi, R., ET-score: Improving Protein-ligand Binding Affinity Prediction Based on Distance-weighted Interatomic Contact Features Using Extremely Randomized Trees Algorithm. *Molecular Informatics* **2021,** *40* (8), 2060084.
25. Goodfellow, I.; Bengio, Y.; Courville, A., *Deep learning*. MIT press: 2016.
26. Durrant, J. D.; McCammon, J. A., NNScore: a neural-network-based scoring function for the characterization of protein− ligand complexes. *Journal of chemical information and modeling* **2010,** *50* (10), 1865-1871.
27. Durrant, J. D.; McCammon, J. A., NNScore 2.0: a neural-network receptor–ligand scoring function. *Journal of chemical information and modeling* **2011,** *51* (11), 2897-2903.
28. Ashtawy, H. M.; Mahapatra, N. R., BgN-Score and BsN-Score: bagging and boosting based ensemble neural networks scoring functions for accurate binding affinity prediction of protein-ligand complexes. *BMC bioinformatics* **2015,** *16* (4), 1-12.
29. Zhu, F.; Zhang, X.; Allen, J. E.; Jones, D.; Lightstone, F. C., Binding affinity prediction by pairwise function based on neural network. *Journal of chemical information and modeling* **2020,** *60* (6), 2766-2772.
30. Jiménez, J.; Skalic, M.; Martinez-Rosell, G.; De Fabritiis, G., K deep: protein–ligand absolute binding affinity prediction via 3d-convolutional neural networks. *Journal of chemical information and modeling* **2018,** *58* (2), 287-296.
31. Stepniewska-Dziubinska, M. M.; Zielenkiewicz, P.; Siedlecki, P., Development and evaluation of a deep learning model for protein–ligand binding affinity prediction. *Bioinformatics* **2018,** *34* (21), 3666-3674.
32. Zheng, L.; Fan, J.; Mu, Y., Onionnet: a multiple-layer intermolecular-contact-based convolutional neural network for protein–ligand binding affinity prediction. *ACS omega* **2019,** *4* (14), 15956-15965.
33. Hassan-Harrirou, H.; Zhang, C.; Lemmin, T., RosENet: improving binding affinity prediction by leveraging molecular mechanics energies with an ensemble of 3D convolutional neural networks. *Journal of chemical information and modeling* **2020,** *60* (6), 2791-2802.
34. Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; Philip, S. Y., A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems* **2020,** *32* (1), 4-24.
35. Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. In *Neural message passing for quantum chemistry*, International conference on machine learning, PMLR: 2017; pp 1263-1272.
36. 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 central science* **2018,** *4* (11), 1520-1530.
37. Karlov, D. S.; Sosnin, S.; Fedorov, M. V.; Popov, P., graphDelta: MPNN scoring function for the affinity prediction of protein–ligand complexes. *ACS omega* **2020,** *5* (10), 5150-5159.
38. Li, Y.; Liu, Z.; Li, J.; Han, L.; Liu, J.; Zhao, Z.; Wang, R., Comparative assessment of scoring functions on an updated benchmark: 1. Compilation of the test set. *Journal of chemical information and modeling* **2014,** *54* (6), 1700-1716.
39. Li, Y.; Su, M.; Liu, Z.; Li, J.; Liu, J.; Han, L.; Wang, R., Assessing protein–ligand interaction scoring functions with the CASF-2013 benchmark. *Nature protocols* **2018,** *13* (4), 666-680.
40. Su, M.; Feng, G.; Liu, Z.; Li, Y.; Wang, R., Tapping on the black box: how is the scoring power of a machine-learning scoring function dependent on the training set? *Journal of chemical information and modeling* **2020,** *60* (3), 1122-1136.
41. Himanen, L.; Jäger, M. O.; Morooka, E. V.; Canova, F. F.; Ranawat, Y. S.; Gao, D. Z.; Rinke, P.; Foster, A. S., DScribe: Library of descriptors for machine learning in materials science. *Computer Physics Communications* **2020,** *247*, 106949.
42. Géron, A., *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media: 2019.
43. Marsland, S., *Machine learning: an algorithmic perspective*. Chapman and Hall/CRC: 2011.

44. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V., Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* **2011,** *12*, 2825-2830.
45. Kramer, C.; Gedeck, P., Leave-cluster-out cross-validation is appropriate for scoring functions derived from diverse protein data sets. *Journal of chemical information and modeling* **2010,** *50* (11), 1961-1969.
46. Li, Y.; Yang, J., Structural and sequence similarity makes a significant impact on machine-learning-based scoring functions for protein–ligand interactions. *Journal of chemical information and modeling* **2017,** *57* (4), 1007-1012.
47. Wang, Y.; Guo, Y.; Kuang, Q.; Pu, X.; Ji, Y.; Zhang, Z.; Li, M., A comparative study of family-specific protein–ligand complex affinity prediction based on random forest approach. *Journal of computer-aided molecular design* **2015,** *29* (4), 349-360.
48. Meli, R.; Anighoro, A.; Bodkin, M. J.; Morris, G. M.; Biggin, P. C., Learning protein-ligand binding affinity with atomic environment vectors. *Journal of Cheminformatics* **2021,** *13* (1), 1-19.
49. Louppe, G., Understanding random forests: From theory to practice. *arXiv preprint arXiv:1407.7502* **2014**.
50. Wójcikowski, M.; Kukiełka, M.; Stepniewska-Dziubinska, M. M.; Siedlecki, P., Development of a protein–ligand extended connectivity (PLEC) fingerprint and its application for binding affinity predictions. *Bioinformatics* **2019,** *35* (8), 1334-1341.
51. Shen, C.; Hu, Y.; Wang, Z.; Zhang, X.; Pang, J.; Wang, G.; Zhong, H.; Xu, L.; Cao, D.; Hou, T., Beware of the generic machine learning-based scoring functions in structure-based virtual screening. *Briefings in Bioinformatics* **2021,** *22* (3), bbaa070.
52. Cesar de Azevedo, L.; Pinheiro, G. A.; Quiles, M. G.; Da Silva, J. L.; Prati, R. C., Systematic Investigation of Error Distribution in Machine Learning Algorithms Applied to the Quantum-Chemistry QM9 Data Set Using the Bias and Variance Decomposition. *Journal of Chemical Information and Modeling* **2021,** *61* (9), 4210-4223.
53. Kwon, Y.; Lee, D.; Choi, Y.-S.; Kang, M.; Kang, S., Neural Message Passing for NMR Chemical Shift Prediction. *Journal of chemical information and modeling* **2020,** *60* (4), 2024-2030.
54. Sutton, C.; Boley, M.; Ghiringhelli, L. M.; Rupp, M.; Vreeken, J.; Scheffler, M., Identifying domains of applicability of machine learning models for materials science. *Nature communications* **2020,** *11* (1), 1-9.

TOC graphic

GB-Score: Minimally Designed Machine Learning Scoring Function Based on Distance-weighted Interatomic Contact Features Milad Rayka and Rohoullah Firouzi*