Published in final edited form as:
J Comput Chem. 2017 January 30; 38(3): 169–177. doi:10.1002/jcc.24667.

Improving Scoring-Docking-Screening Powers of Protein-Ligand Scoring Functions using Random Forest

Cheng Wang[1] and **Yingkai Zhang**[1,2,*]

[1]Department of Chemistry, New York University, New York, New York 10003

[2]NYU-ECNU Center for Computational Chemistry at NYU Shanghai, Shanghai 200062, China

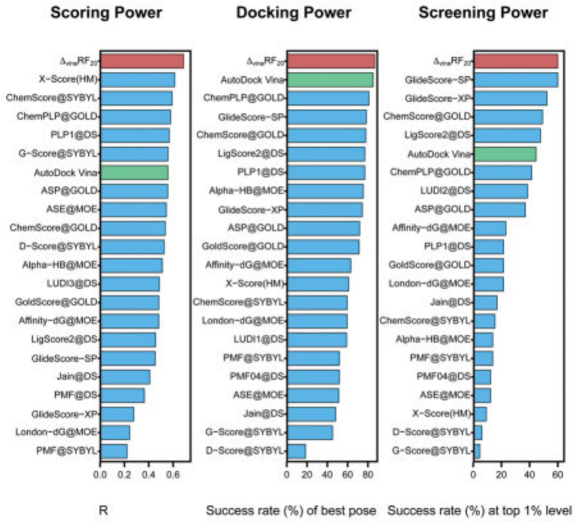
Abstract

The development of new protein-ligand scoring functions using machine learning algorithms, such as random forest, has been of significant interest. By efficiently utilizing expanded feature sets and a large set of experimental data, random forest based scoring functions (RFbScore) can achieve better correlations to experimental protein-ligand binding data with known crystal structures; however, more extensive tests indicate that such enhancement in scoring power comes with significant under-performance in docking and screening power tests compared to traditional scoring functions. In this work, in order to improve scoring-docking-screening powers of protein-ligand docking functions simultaneously, we have introduced a Δ_{vina}RF parameterization and feature selection framework based on random forest. Our developed scoring function $\Delta_{vina}RF_{20}$, which employs twenty descriptors in addition to the AutoDock Vina score, can achieve superior performance in all power tests of both CASF-2013 and CASF-2007 benchmarks compared to classical scoring functions. The $\Delta_{vina}RF_{20}$ scoring function and its code are freely available on the web at: https://www.nyu.edu/projects/yzhang/DeltaVina.

Graphical Abstract

A new machine-learning protein-ligand scoring function $\Delta_{vina}RF_{20}$ has been developed to achieve better performance for a variety of protein-ligand docking tasks compared to classical scoring functions.

[*]To whom correspondence should be addressed: yingkai.zhang@nyu.edu.



1. Introduction

Protein ligand docking is a computational approach that attempts to predict the binding mode between a protein receptor and a small molecule ligand as well as their binding affinity. It plays an increasingly important role in structure-based drug design as well as in functional studies of proteins. The most critical component of docking is the scoring function, which is needed to determine binding site and binding mode of a ligand on a protein,[1] to screen virtual small-molecule libraries to identify potential leads for further inhibitor development,[2–8] and to explicitly estimate the binding affinity between a protein and a ligand given their complex structure. Correspondingly, in order to assess performance of scoring functions for these different important tasks, several key metrics have been developed and adopted, including: (i) a docking power test, which evaluates the ability of the scoring function to identify the native binding site and binding mode among a set of computer generated decoys; (ii) a screening power test to evaluate the ability of the scoring function to identify a true binder for a given target from a pool of random molecules; and (iii) a scoring power test, which assesses the linear correlation between predicted and experimental binding affinities.[9–13] Extensive retrospective and comparative studies[10,11,14–18] indicate that some widely used scoring functions, such as GlideScore[19–21], can perform relatively well in docking and screening power tests, but most perform less satisfactorily in the scoring power test. Thus, the accuracy of scoring functions remains a central limitation of protein-ligand docking.

In the last few years, machine learning approaches have proven useful for many technologies in modern society, such as computer vision and natural language processing.[22–25] In the field of biomolecular modeling, there has been significant interest to develop new protein-ligand scoring functions using state-of-the-art machine learning methods,[26–40] such as the random forest (RF) algorithm. By efficiently utilizing expanded feature sets and a large set of experimental data, random forest based scoring functions (RFbScore)[26–32] have achieved significantly better correlations with experimental protein-ligand binding data for known crystal structures; however, more extensive testing indicates that this enhancement in scoring power comes with significant under-performance in docking and screening power tests compared to traditional scoring functions.[41,42]

Random forest is an ensemble learning method based on the aggregation of numerous decision trees.[43,44] In RFbScore, every regression tree is a non-parametric predictive model to relate structural features to binding affinities, the predicted values of which are bounded by the learning set. Thus random forest can do interpolation but not extrapolation.[45] Without a predetermined function form, random forest has the ability to learn complicated interactions directly from a large set of experimental data based on numerous input features. Up to now, almost all published RFbScores that predict binding affinity have used experimental protein-ligand binding data with known crystal structures as the training set alone. Thus, in retrospect, it is not surprising that RFbScores can achieve success in scoring power tests, which mostly rely on interpolation—i.e. to estimate binding affinities given experimentally determined structures.[29] On the other hand, numerous tasks in docking and screening tests depend on extrapolation—i.e. to estimate binding affinities for computationally generated structures which should have weaker binding affinities. Thus, it is understandable that RFbScores would falter when applied to such decoy structures, leading to significant underperformance of RFbScores in docking and screening power tests.[41,42]

From the above discussion, we can see that the inferior performance of RFbScores in docking tests may reflect two problems: 1. Random forest is restricted to do interpolation; 2. Use of experimental protein-ligand binding data with known crystal structures as the training set alone limits the applicability of RFbScores. In this work, in order to improve scoring-docking-screening powers of scoring functions simultaneously, we employ a two-pronged strategy:

- One is to expand the training set. Besides enlarging the experimental data set to include crystal structures with weak binding affinities, we have added a similar amount of computationally generated structures (decoy data) into the training set. The idea of including decoy data in the training set has been previously employed in the development of several other scoring functions.[34,37]

- The other is to employ a Δ_{vina}RF approach, in which random forest is employed to parameterize corrections to the AutoDock Vina scoring function. This is partly inspired by the recent development of the Δ-machine learning approach to predict enthalpies of organic molecules.[46] AutoDock Vina is one of most widely used open-source docking programs, which has been successfully employed in numerous docking and screening tasks.[47] Our Δ_{vina}RF parameterization framework aims to combine the excellent docking power of the Vina docking function with the strength of random forest for improving scoring accuracy.

Furthermore, by employing random forest for feature selection and introducing a pharmacophore-based solvent-accessible surface area (SASA) feature set, our developed $\Delta_{vina}RF_{20}$, which employs twenty descriptors in addition to the AutoDock Vina score, can achieve superior performance compared to traditional scoring functions in all tests of both

CASF-2013 and CASF-2007 benchmarks, including scoring, ranking, docking and screening power tests. It should be noted that our new scoring function $\Delta_{vina}RF_{20}$ has not been incorporated into AutoDock Vina for ligand sampling, and currently it can only be used for post-scoring.

2. Methods

2.1 Training Set of Protein-ligand Complexes

The training set of protein-ligand complexes for this work consists of two subsets: one is an experimental subset, which includes 3336 crystal complex structures with experimentally measured binding affinities; the other is a decoy subset, which includes 3322 computer generated decoy structures with computationally estimated binding affinities (Table S1). These are obtained, respectively, from the PDBbind database,[12,48–50] which is a collection of protein-ligand complex PDBs with experimental binding affinities, and the CSAR decoy set, which is a collection of computer generated binding poses as well as native poses for structures in the CSAR-NRC HiQ benchmark release.[51,52] Any structure in the CASF-2007 or CASF-2013 benchmark sets,[12,13] which will be used for the test set, is excluded from the training set.

The experimental subset consists of data from three sources: the PDBbind refined set (v2014), native poses in the CSAR decoy data set, and weak-binding protein-ligand crystal structures (pK_d between 0.4 to 3) in PDBbind v2014 general set. Any entry in the CASF-2007 or CASF-2013 benchmark set is excluded.

The decoy subset contains decoy data for 302 protein-ligand complexes in the CSAR decoy set, excluding 41 complexes that are also in the CASF-2007 or CASF-2013 benchmark set. For each native pose, 11 decoys are selected from up to 500 decoys in the original CSAR decoy set based on the rank of AutoDock Vina score at 0%, 10%, 20%, …, 90%, 100%, respectively. Thus the decoy subset has a similar number of data entries as in the experimental subset.

The binding affinity for each complex in the training set is denoted as pK_d(train). For each entry in the experimental subset, pK_d(train) is the experimental binding affinity pK_d(exp). The binding affinity for each entry in the decoy subset should not be larger than the experimental binding affinity for the corresponding native pose. For each decoy, first we calculate pK_d(Vina) based on the AutoDock Vina score: if the calculated pK_d(Vina) of a decoy structure is less than pK_d(exp) of the corresponding native pose, we assign pK_d(Vina) as pK_d(train) for this decoy structure; otherwise pK_d(train) of the decoy structure is assumed to be at the upper limit, which is the pK_d(exp) of the corresponding native pose.

2.2 The Δ_{vina}RF Parameterization Approach

Our main idea is to employ random forest to parameterize corrections to the AutoDock Vina scoring function, and thus to take advantage of both the excellent docking power of the Vina docking function and the strength of random forest in improving scoring accuracy. The Vina scoring function consists of six components: two Gaussian steric terms, one repulsion term, one hydrogen bonding (HB) term, and one torsion count factor, and has been parameterized to improve both binding pose and affinity prediction.[47] The original score calculated by the AutoDock Vina program is in the unit of kcal/mol, and can be converted into pK_d unit with the following formula: $pK_d(Vina) = -0.73349\ E(Vina)$. Thus, our overall $\Delta_{vina}RF$ scoring function can be cast into the following form:

$$pK_d(\Delta_{vina}RF) = pK_d(Vina) + \Delta pK_d(RF)$$

where $\Delta pK_d(RF)$ is the correction term trained by the random forest (RF) algorithm using $\Delta pK_d(train)$, i.e., $pK_d(train) - pK_d(Vina)$.

Given a learning set $L = \{(X^{(1)}, y^{(1)}), \ldots, (X^{(N)}, y^{(N)})\}$, which contains N pairs of input feature vectors $X = (x_1, x_2, \ldots, x_p)$ and output values y, each regression tree in a random forest model can be grown as follows: (1). Sample the learning set. Prior to growing a decision tree *k*, a bootstrap learning subset L_k^* is drawn at random from L with replacement, and the left-out data ($L - L_k^*$) constitutes the (OOB) out-of-bag subset OOB_k; (2). Grow a single decision tree T_k. Based on the bootstrap learning subset L_k^*, T_k is constructed by recursively splitting each terminal node of the tree into two child nodes until the minimum node size is reached. For each splitting, it picks the best feature from a pool of m_{try} features. The m_{try} features are randomly selected from all p features. (3). The prediction error of the T_k is estimated using the out-of-bag subset OOB_k. After repeating steps 1–3 to grow M regression trees, the collection of all regression trees (T_k, $k = 1, \ldots, M$) is considered as a predictive RF model. To make a prediction with a new input feature vector $X^{(new)}$, its predicted value is the average of predictions from all trees: [44]

$$y^{(new)} = \frac{1}{M}\sum_{k=1}^{M} T_k(X^{(new)})$$

In our development of $\Delta_{vina}RF_{20}$ parameterization, the learning set L is derived from our training set that has been described above: N is 6658; the output value y is $\Delta pK_d(train)$, i.e., $pK_d(train) - pK_d(Vina)$; the input feature vector has p = 20 features, which are calculated based on the corresponding protein-ligand structure in the training set. The randomForest package in R is used to build random forest models.[53] The final $\Delta_{vina}RF_{20}$ model is built by using M = 500 regression trees with $m_{try} = 4$, selected based on the OOB performance of the learning set.

The twenty features in $\Delta_{vina}RF_{20}$ are listed in Table 1. There are 10 terms from the AutoDock Vina source code and 10 terms related to buried solvent-accessible surface area (bSASA). In the AutoDock Vina source code[47], there are a total of 58 terms implemented as listed in Table S2, among which 6 terms have been selected for the AutoDock Vina scoring function, including: two Gaussian steric terms, one repulsion term, one hydrogen bonding (HB) term, and one torsion count factor. These 58 Vina-implemented terms have been explored in the development of smina and user-specified custom scoring functions with linear regression.[54] During our development of $\Delta_{vina}RF_{20,}$ we first ranked 58 Vina-implemented terms based on the permutation variable importance indices %IncMSE, which is OOB mean square error (MSE) increase as a result of feature i being permuted (values are randomly shuffled). For a given feature i, it is calculated by

$$\%\mathrm{IncMSE}_i=\frac{\mathrm{MSE}_i^{\mathrm{OOB}}-\mathrm{MSE}^{\mathrm{OOB}}}{\mathrm{MSE}^{\mathrm{OOB}}}\times 100\%$$

where $\mathrm{MSE}^{\mathrm{OOB}}$ is the OOB MSE and $\mathrm{MSE}_i^{\mathrm{OOB}}$ is the OOB MSE when feature i is permuted. More important features have higher %IncMSE values. Then we have employed a backward feature selection approach, in which the least important features are removed one by one to build random forest models, to choose the least number of features with a comparable top performance. The selected 10 Vina-implemented terms in Table 1 include 5 polar interaction terms and 5 ligand-dependent terms.

The bSASA terms are calculated using atomic SASA changes between the unbound and bound structures: for an atom i, $\mathrm{bSASA}_\mathrm{i} = SASA_{\mathrm{i,\ unbound}} - SASA_{\mathrm{i,\ complex}}$, where atomic SASAs are calculated by the MSMS program using a probe radius of 1.0 Å.[55] As shown in Table S3, nine pharmacophore types are defined for the atoms in the protein and ligand based on SYBYL atom types and neighboring atoms as in DOCK.[56] The SYBYL atom types[57] are converted by Pybel from the structures with hydrogen atoms added.[58] Thus in Table 1, there are 9 pharmacophore-based bSASA terms and 1 total bSASA term.

2.3 Testing Set and Evaluation Methods

Both CASF-2013 and CASF-2007 benchmark sets[12,13] are used as testing sets so that the results can be directly compared with other docking functions. Both datasets consist of 195 protein-ligand complexes selected from a refined dataset in their respective year's PDBbind database.[12,48] Scoring, ranking and docking powers have been evaluated for 16 scoring functions (in Table S4) for the CASF-2007 benchmark set[12], while scoring, ranking, docking and screening power tests have been carried out for 20 scoring functions (in Table S5) for the CASF-2013 benchmark set.[13] In our current work, all power tests for AutoDock Vina and $\Delta_{\mathrm{vina}}\mathrm{RF}_{20}$ are carried out in the same way as those described in comparative assessment articles for CASF-2007 and CASF-2013,[13] which are summarized below.

Scoring Power—The scoring power test evaluates the linear correlation between predicted binding affinity and experimental binding affinity. It is evaluated by the Pearson's correlation coefficient (R) between predicted binding affinity and experimental binding affinity and the standard deviation (SD) in regression:

$$R=\frac{\sum(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum(x_i-\overline{x})^2}\sqrt{\sum(y_i-\overline{y})^2}},\ \mathrm{SD}=\sqrt{\frac{\sum[y_i-(a+bx_i)]^2}{N-1}}$$

where x_i is the predicted binding affinity for ith complex; y_i is the experimental binding affinity for ith complex; a and b are the intercept and the slope of linear regression between experimental binding affinity and predicted binding affinity.

Ranking Power—The ranking power test assesses the ability of a scoring function to correctly rank the known ligands of the same target protein based on their predicted binding affinity given the poses from the crystal structures. For each benchmark, there are 65 target proteins and 3 known ligands for each protein. Two levels of success, namely high-level and low-level, are evaluated in CASF-2013. For the high-level, the three ligands for target protein should be ranked by predicted score as the best > the median > the poorest, while the low-level only needs to pick the best one out of three. The success rate is calculated by the number of the correctly ranked targets among all 65 targets. In CASF-2007, only the high-level success rate is evaluated.[12]

Docking Power—The docking power test evaluates the ability of a scoring function to identify native binding poses among computer generated decoys. In CASF-2007, 100 decoy binding poses are selected from the poses generated by LigandFit,[60] GOLD,[61,62] Surflex[63,64] and FlexX[65] for each ligand. Success is defined as one pose from the top one, the top two, or the top three poses ranked by predicted scores is within 2 Å RMSD from the native pose. The RMSDs of decoys relative to the native are provided in CASF-2007 benchmark and used directly. In CASF-2013, up to 100 decoy binding poses are selected from the poses generated by GOLD, Surflex and MOE. The property-matched RMSDs ($\mathrm{RMSD}^{\mathrm{PM}}$) of decoys, are calculated by considering the symmetry of the molecule in CASF-2013, however, $\mathrm{RMSD}^{\mathrm{PM}}$ is not provided in CASF-2013. The symmetry-corrected RMSDs used here are calculated by Pybel.[58] The native poses are included in the decoy set for success rate calculation.

Screening Power—The screening power test assesses the ability of a scoring function to identify a true binder from a pool of random molecules for a given target. The test set for screening in CASF-2013 is designed by cross docking 195 ligands on 65 target proteins. For each protein, there are at least 3 true binders, as defined in Ref. 13. The remaining 192 ligands are searched through the ChEMBL database for possible cross-binders and 12 target proteins have more than 3 true binders in the dataset from the search. There are 12,675 (65 × 195) protein-ligand pairs from docking 195 ligands to 65 target proteins and up to 50 poses are selected for each protein-ligand pair. For a given target protein, 195 ligands are ranked based on the best-scored pose for a given protein-ligand pair. Screening power is measured by two metrics-enhancement factors and success rates—both based on the counts of the total number of true binders among the 1%, 5% and 10% top-ranked molecules. Enhancement factors are computed for each target by

$$\mathrm{EF}_{1\%}=\frac{\mathrm{NTB}_{1\%}}{\mathrm{NTB}_{\mathrm{total}}\times 1\%},\ \mathrm{EF}_{5\%}=\frac{\mathrm{NTB}_{5\%}}{\mathrm{NTB}_{\mathrm{total}}\times 5\%},\ \mathrm{EF}_{10\%}=\frac{\mathrm{NTB}_{10\%}}{\mathrm{NTB}_{\mathrm{total}}\times 10\%}$$

$\mathrm{NTB}_{1\%}$, $\mathrm{NTB}_{5\%}$ and $\mathrm{NTB}_{10\%}$ are the number of true binders among the 1%, 5% and 10% top-ranked molecules. $\mathrm{NTB}_{\mathrm{total}}$ is the total number of true binders for a given target protein. The average EFs over 65 targets are calculated for each scoring function. Success rates are calculated as the number of targets that have true binders in 1%, 5% and 10% of the top-ranked molecules among the total 65 targets.

3. Results

Based on a learning set which consists of 3336 experimental crystal structures and 3322 computer generated decoy structures, we have developed a new scoring function $\Delta_{vina}RF_{20}$ by employing random forest to parameterize a correction term to the original AutoDock Vina score.[47] The overall scoring function $\Delta_{vina}RF_{20}$ has the following form: $pK_d(\Delta_{vina}RF_{20}) = pK_d(Vina) + \Delta pK_d(RF_{20})$, where pK_d(Vina) is the pK_d value calculated by multiplying the original Vina score by a factor of −0.73349. $\Delta pK_d(RF_{20})$ is the correction term parameterized with random forest using 20 features as listed in Table 1. The $\Delta_{vina}RF_{20}$ model and code are available at: https://www.nyu.edu/projects/yzhang/DeltaVina. We have carried out all power tests on both CASF-2013 and CASF-2007 benchmark sets[12,13] for $\Delta_{vina}RF_{20}$ as well as the AutoDock Vina scoring function,[47] and compared with, respectively, 20 other and 16 other docking functions that were tested in the original 2013 and 2007 comparative assessment articles. The results are presented in Figures 1–3 and Tables S6–S13. The new scoring function $\Delta_{vina}RF_{20}$, which employs twenty descriptors in addition to the AutoDock Vina score, has achieved superior performance compared to classical scoring functions in all tests of both CASF-2013 and CASF-2007 benchmarks, including scoring, ranking, docking and screening power tests.

Scoring Power—The scoring power of $\Delta_{vina}RF_{20}$ significantly outperforms AutoDock Vina as well as all scoring functions that have been tested in the original 2013 and 2007 comparative assessment articles, as shown in Figure 1 and 2. It achieves the best Pearson's correlation coefficients of 0.686 and 0.732 for the CASF-2013 and CASF-2007 benchmarks respectively, and significantly improves upon AutoDock Vina, which has corresponding Pearson's correlation coefficients of 0.557 and 0.566 respectively.

Ranking Power—The ranking power of $\Delta_{vina}RF_{20}$ is improved over AutoDock Vina, and is among the top 3 for both benchmarks. In CASF-2013, $\Delta_{vina}RF_{20}$ has a ranking power of 55% for high-level (in Figure 1) and 74% for low-level (in Table S10), which places it third for high-level successes and second for low-level successes. In CASF-2007, the ranking power of $\Delta_{vina}RF_{20}$ is 57%, following the best X-Score::HSScore's success rate of 58%.[66]

Docking Power—The docking power of $\Delta_{vina}RF_{20}$ is among the top rank for both benchmarks. Its success rate to identify the top pose as the native pose is 87% in CASF-2013 (see Figure 1), which outperforms all other scoring functions. In CASF-2007 (in Figure 2), the success rate of $\Delta_{vina}RF_{20}$ is 80%, which ranks second following the best one GOLD::ASP (82%).[67] We can see that the docking power of $\Delta_{vina}RF_{20}$ has improved upon AutoDock Vina by about 2% in both benchmarks. Recently, a new SPA-SE score function was developed by combining knowledge-based atom-pair potential with the atomic solvation energy of a charge-independent implicit solvent model.[68] It shows excellent performance in scoring (with a Pearson's correlation coefficient of 0.662), ranking (60.0% for high level and 75.4% for low level) and docking power (83.1% success rate for the best pose) for the CASF-2013 benchmark, while the screening power of SPA-SE was not reported. We can see that $\Delta_{vina}RF_{20}$ is still slightly better than SPA-SE in both scoring and docking power tests for the CASF-2013 benchmark.

Screening Power—The screening power of $\Delta_{vina}RF_{20}$ is the best as shown in Figure 3 for both metrics: enrichment factor and success rate at top 1% level. The average enrichment factor of $\Delta_{vina}RF_{20}$ is 21 at top 1% level, which is slightly better than GlideScore-SP's 20, and the success rate is 60% (39 out of 65), which is the same as GlideScore-SP at top 1% level.[19,20] It should be noted that $\Delta_{vina}RF_{20}$ has significantly improved upon AutoDock Vina, which has an enrichment factor of 15.6 and success rate of 45% at top 1% level respectively.

4. Discussion

Scoring functions play a central role in protein-ligand docking. An ideal, robust scoring function should perform well across different important tasks, including scoring, docking, and screening power tests. Extensive retrospective and comparative studies[10,11,14–18] indicate that although some widely used scoring functions can do relatively well in docking and screening power tests, most of them are weaker in performance in the scoring power test. Furthermore, it is very challenging for a docking function to achieve superior performance on all three power tests simultaneously.[41,42] For example, X-Score(HM)[66] is the top performer in the scoring power test in the original CASF-2013 comparative study, with a Pearson's correlation coefficient of 0.614, but its performance in the docking power test only ranks in the middle among about 20 tested scoring functions. Its success rate is 61% in predicting the best pose, which is significant lower than the value of 81% for ChemPLP@GOLD.[69] These disparities in performance for different power tests become significantly worse for recently developed machine learning-based scoring functions (MLbScores).[41,42] For example, in a recent comparative assessment of a dozen MLbScores[42] for the CASF-2013 benchmark, RF@ML, which is parameterized with random forest using more than one hundred features, achieved the best scoring power of 0.704 in Pearson's correlation coefficient among all 12 scoring functions developed using different machine learning algorithms. However, the docking and screening powers of these 12 MLbScores are all significantly worse. The docking power of RF@ML is only 12.2% for success in predicting the best pose, while the screening power of RF@ML is just 6.45% for the success rate in finding the best ligand molecule. These values are significantly lower than those of classical scoring functions, whose top performances are 81% (ChemPLP@GOLD)[69] and 60% (GlideScore-SP)[19,20] for docking and screening power respectively. On the other hand, as presented in the above results section, our newly developed $\Delta_{vina}RF_{20}$ scoring function, using 20 features, achieved superior performance compared to traditional scoring functions in all power tests for both CASF-2013 and CASF-2007 benchmarks.

The main idea of our $\Delta_{vina}RF$ approach is to use random forest to parameterize a correction term to the AutoDock Vina score so that it can combine Vina's excellent docking power with RF's ability to significantly improve scoring accuracy. We tested both RF and $\Delta_{vina}RF$ approaches using the same twenty features and training set for the development of the $\Delta_{vina}RF_{20}$ scoring function, and compared them with Vina in Figure 4 for the scoring, docking and screening power tests in CASF-2013. In addition, we have also trained a RF model using experimental data alone and the same twenty features. The results clearly demonstrate that the RF approach can only improve the scoring power by significantly sacrificing docking and screening powers, while the Δ_{vina}RF approach, with a combined experimental and decoy training set, can achieve the improvement over AutoDock Vina in all three tests simultaneously.

One attractive capability of the random forest algorithm is that it can efficiently utilize a large set of training data. It has been previously demonstrated that the larger the training data, the better the resulting RFbScore's performance in the scoring power test[28,29,70], even when low-quality structures are included[30]. For the Δ_{vina}RF approach, as shown in Figure S1, we also find that a larger experimental learning set can significantly improve the scoring power, but it does not necessarily improve the docking power. By expanding the learning set to include decoy structures, both docking and screening power of the scoring function can be improved over AutoDock Vina with the Δ_{vina}RF approach.

Besides the training set, another critical component of a random forest based scoring function (RFbScore) is the feature set. For a random forest model, both the number of features and feature relevance will affect its performance. Numerical experiments show that increasing the fraction of relevant features can improve the performance of the random forest model by increasing the chance that important features will be selected at each tree splitting.[44] In this work, by taking advantage of random forest in ranking features, we employ a strategy that includes both feature selection and aggregation, to yield the twenty features in Δ_{vina}RF$_{20}$. As listed in Table 1, the feature set of Δ_{vina}RF$_{20}$ consists of 5 interaction terms and 5 ligand-dependent terms, which are selected from 58 terms implemented in the AutoDock Vina source code and 10 terms related to buried solvent-accessible surface area (bSASA). Interestingly, all 5 interaction terms in the Δ_{vina}RF$_{20}$ are related to polar interactions, and there is only one overlap term between Δ_{vina}RF$_{20}$ and the original Vina score, which is the number of torsions in the ligand. In comparison with using only 6 terms in the original Vina score and all 58 terms implemented in the AutoDock Vina source code, the Δ_{vina}RF model, developed using 10 selected features, performs better in all scoring, docking and screening power tests for the CASF-2013 benchmark test as shown in Figure S2.

Among 5 ligand-dependent terms, rotors and torsions have been previously used to approximate the entropic change in several empirical scoring functions.[19,66,71,72] Number of heavy atoms and ligand length could also be viewed as entropy-related features since both of them are highly correlated with rotors (Pearson's correlation coefficients are 0.802 and 0.906, respectively) in the crystal structure training set. Number of hydrophobic atoms is related to the hydrophobic interaction and the Pearson's correlation coefficient between number of hydrophobic atoms and the hydrophobic term defined in AutoDock Vina source code is around 0.85 for the crystal structure training set.

Surface area and related features are widely used in protein-ligand scoring function development due to their relation to solvation.[66,68,71–74] The buried solvent-accessible surface area of a ligand (bSASA) has also been tested as a naive scoring function in CASF-2013 scoring comparison list[13] and outperforms most of other scoring functions in the scoring power test but ranks the worst in both docking and screening power tests. Since no bSASA term has been implemented in the AutoDock Vina source code, we have explored

9 pharmacophore-based bSASA terms and 1 total bSASA term as the feature set. As shown in Figure S3, the Δ_{vina}RF model developed using 10 bSASA terms alone would also perform quite well in all three power tests, but not as good as when combined with 10 selected AutoDock Vina terms. By combining two feature sets, the resulting Vina10-bSASA, the feature set used in $\Delta_{vina}RF_{20}$, performs better than either Vina10 or bSASA in all three power tests for the CASF-2013 benchmark. We have tested the importance of features measured by percentage of increased mean squared error (%IncMSE) for the 20 features in $\Delta_{vina}RF_{20}$ as shown in Figure S4. Except for the halogen bSASA, which may be limited by the rarity of halogens in the training set, each of the other 19 features has a %IncMSE value significantly larger than 20%, indicating their general importance and justifying their inclusion in the feature set of $\Delta_{vina}RF_{20}$. Meanwhile, the additional test results in Figures S1–S4 further indicate the robustness of the Δ_{vina}RF approach.

5. Conclusion

A major challenge in developing a robust protein-ligand scoring function is to improve scoring, docking and screening performances simultaneously. In this work, we have made advances in overcoming this challenge by introducing a new Δ_{vina}RF parameterization and feature selection framework based on random forest. Our new scoring function $\Delta_{vina}RF_{20}$ employs twenty features in addition to the AutoDock Vina score, and can achieve superior performance compared to classical scoring functions in all tests of both CASF-2013 and CASF-2007 benchmarks, including scoring, ranking, docking and screening power tests. This work suggests that Δ-machine learning is a promising approach to systemically improve the performance and robustness of docking functions by employing larger diverse experimental/decoy data sets of high quality, developing and selecting physically meaningful features, as well as adapting advanced machine learning algorithms.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to acknowledge the support by NIH (R01-GM079223). We thank NYU-ITS and NYUAD for providing computational resources.

References

1. Huang SY, Grinter SZ, Zou XQ. Phys Chem Chem Phys. 2010; 12:12899. [PubMed: 20730182]
2. Lyne PD. Drug Discovery Today. 2002; 7:1047. [PubMed: 12546894]
3. Shoichet BK. Nature. 2004; 432:862. [PubMed: 15602552]
4. McInnes C. Curr Opin Chem Biol. 2007; 11:494. [PubMed: 17936059]
5. Guido RVC, Oliva G, Andricopulo AD. Curr Med Chem. 2008; 15:37. [PubMed: 18220761]
6. Cheng TJ, Li QL, Zhou ZG, Wang YL, Bryant SH. AAPS J. 2012; 14:133. [PubMed: 22281989]
7. Lavecchia A, Di Giovanni C. Curr Med Chem. 2013; 20:2839. [PubMed: 23651302]
8. Ma DL, Chan DSH, Leung CH. Chem Soc Rev. 2013; 42:2130. [PubMed: 23288298]
9. Wang RX, Lu YP, Wang SM. J Med Chem. 2003; 46:2287. [PubMed: 12773034]

10. Ferrara P, Gohlke H, Price DJ, Klebe G, Brooks CL. J Med Chem. 2004; 47:3032. [PubMed: 15163185]
11. Marsden PM, Puvanendrampillai D, Mitchell JBO, Glen RC. Org Biomol Chem. 2004; 2:3267. [PubMed: 15534704]
12. Cheng TJ, Li X, Li Y, Liu ZH, Wang RX. J Chem Inf Model. 2009; 49:1079. [PubMed: 19358517]
13. Li Y, Han L, Liu ZH, Wang RX. J Chem Inf Model. 2014; 54:1717. [PubMed: 24708446]
14. Halperin I, Ma BY, Wolfson H, Nussinov R. Proteins: Struct, Funct, Genet. 2002; 47:409. [PubMed: 12001221]
15. Perola E, Walters WP, Charifson PS. Proteins: Struct, Funct, Bioinf. 2004; 56:235.
16. Warren GL, Andrews CW, Capelli AM, Clarke B, LaLonde J, Lambert MH, Lindvall M, Nevins N, Semus SF, Senger S, Tedesco G, Wall ID, Woolven JM, Peishoff CE, Head MS. J Med Chem. 2006; 49:5912. [PubMed: 17004707]
17. Kim R, Skolnick J. J Comput Chem. 2008; 29:1316. [PubMed: 18172838]
18. Plewczynski D, Lazniewski M, Augustyniak R, Ginalski K. J Comput Chem. 2011; 32:742. [PubMed: 20812323]
19. Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, Repasky MP, Knoll EH, Shelley M, Perry JK, Shaw DE, Francis P, Shenkin PS. J Med Chem. 2004; 47:1739. [PubMed: 15027865]
20. Halgren TA, Murphy RB, Friesner RA, Beard HS, Frye LL, Pollard WT, Banks JL. J Med Chem. 2004; 47:1750. [PubMed: 15027866]
21. Friesner RA, Murphy RB, Repasky MP, Frye LL, Greenwood JR, Halgren TA, Sanschagrin PC, Mainz DT. J Med Chem. 2006; 49:6177. [PubMed: 17034125]
22. Jordan MI, Mitchell TM. Science. 2015; 349:255. [PubMed: 26185243]
23. LeCun Y, Bengio Y, Hinton G. Nature. 2015; 521:436. [PubMed: 26017442]
24. Lavecchia A. Drug Discovery Today. 2015; 20:318. [PubMed: 25448759]
25. Libbrecht MW, Noble WS. Nat Rev Genet. 2015; 16:321. [PubMed: 25948244]
26. Ballester PJ, Mitchell JBO. Bioinformatics. 2010; 26:1169. [PubMed: 20236947]
27. Ballester PJ, Schreyer A, Blundell TL. J Chem Inf Model. 2014; 54:944. [PubMed: 24528282]
28. Li HJ, Leung KS, Wong MH, Ballester PJ. BMC Bioinf. 2014; 15:291.
29. Li HJ, Leung KS, Wong MH, Ballester PJ. Mol Inf. 2015; 34:115.
30. Li HJ, Leung KS, Wong MH, Ballester PJ. Molecules. 2015; 20:10947. [PubMed: 26076113]
31. Zilian D, Sotriffer CA. J Chem Inf Model. 2013; 53:1923. [PubMed: 23705795]
32. Liu Q, Kwoh CK, Li JY. J Chem Inf Model. 2013; 53:3076. [PubMed: 24191692]
33. Ashtawy HM, Mahapatra NR. BMC Bioinf. 2015; 16(Suppl 4):S8.
34. Durrant JD, McCammon JA. J Chem Inf Model. 2010; 50:1865. [PubMed: 20845954]
35. Durrant JD, McCammon JA. J Chem Inf Model. 2011; 51:2897. [PubMed: 22017367]
36. Wallach, I.; Dzamba, M.; Heifets, A. [accessed Apr 20, 2016] http://arXiv.org/abs/1510.02855
37. Li L, Wang B, Meroueh SO. J Chem Inf Model. 2011; 51:2132. [PubMed: 21728360]
38. Ding B, Wang J, Li N, Wang W. J Chem Inf Model. 2013; 53:114. [PubMed: 23259763]
39. Li GB, Yang LL, Wang WJ, Li LL, Yang SY. J Chem Inf Model. 2013; 53:592. [PubMed: 23394072]
40. Wang W, He WL, Zhou X, Chen X. Proteins: Struct, Funct, Bioinf. 2013; 81:1386.
41. Gabel J, Desaphy J, Rognan D. J Chem Inf Model. 2014; 54:2807. [PubMed: 25207678]
42. Khamis MA, Gomaa W. Eng Appl Artif Intel. 2015; 45:136.
43. Breiman L. Mach Learn. 2001; 45:5.
44. Hastie, T.; Tibshirani, R.; Friedman, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer-Verlag; New York: 2009.
45. Wyner, AJ.; Olson, M.; Bleich, J.; Mease, D. [accessed Apr 20, 2016] http://arXiv.org/abs/1504.07676
46. Ramakrishnan R, Dral PO, Rupp M, von Lilienfeld OA. J Chem Theory Comput. 2015; 11:2087. [PubMed: 26574412]

47. Trott O, Olson AJ. J Comput Chem. 2010; 31:455. [PubMed: 19499576]
48. Li Y, Liu ZH, Li J, Han L, Liu J, Zhao ZX, Wang RX. J Chem Inf Model. 2014; 54:1700. [PubMed: 24716849]
49. Wang RX, Fang XL, Lu YP, Wang SM. J Med Chem. 2004; 47:2977. [PubMed: 15163179]
50. Wang RX, Fang XL, Lu YP, Yang CY, Wang SM. J Med Chem. 2005; 48:4111. [PubMed: 15943484]
51. Dunbar JB, Smith RD, Yang CY, Ung PMU, Lexa KW, Khazanov NA, Stuckey JA, Wang SM, Carlson HA. J Chem Inf Model. 2011; 51:2036. [PubMed: 21728306]
52. Huang SY, Zou XQ. J Chem Inf Model. 2011; 51:2107. [PubMed: 21755952]
53. Liaw A, Wiener M. R News. 2002; 2:18.
54. Koes DR, Baumgartner MP, Camacho CJ. J Chem Inf Model. 2013; 53:1893. [PubMed: 23379370]
55. Sanner MF, Olson AJ, Spehner JC. Biopolymers. 1996; 38:305. [PubMed: 8906967]
56. Jiang LL, Rizzo RC. J Phys Chem B. 2015; 119:1083. [PubMed: 25229837]
57. Clark M, Cramer RD, Vanopdenbosch N. J Comput Chem. 1989; 10:982.
58. O’Boyle NM, Morley C, Hutchison GR. Chem Cent J. 2008; 2:5. [PubMed: 18328109]
59. Huey R, Morris GM, Olson AJ, Goodsell DS. J Comput Chem. 2007; 28:1145. [PubMed: 17274016]
60. Venkatachalam CM, Jiang X, Oldfield T, Waldman M. J Mol Graphics Modell. 2003; 21:289.
61. Jones G, Willett P, Glen RC. J Mol Biol. 1995; 245:43. [PubMed: 7823319]
62. Jones G, Willett P, Glen RC, Leach AR, Taylor R. J Mol Biol. 1997; 267:727. [PubMed: 9126849]
63. Jain AN. J Med Chem. 2003; 46:499. [PubMed: 12570372]
64. Jain AN. J Comput-Aided Mol Des. 2007; 21:281. [PubMed: 17387436]
65. Rarey M, Kramer B, Lengauer T, Klebe G. J Mol Biol. 1996; 261:470. [PubMed: 8780787]
66. Wang RX, Lai LH, Wang SM. J Comput-Aided Mol Des. 2002; 16:11. [PubMed: 12197663]
67. Mooij WTM, Verdonk ML. Proteins: Struct, Funct, Bioinf. 2005; 61:272.
68. Yan ZQ, Wang J. Proteins: Struct, Funct, Bioinf. 2015; 83:1632.
69. Korb O, Stutzle T, Exner TE. J Chem Inf Model. 2009; 49:84. [PubMed: 19125657]
70. Ashtawy HM, Mahapatra NR. IEEE/ACM Trans Comput Biol Bioinf. 2015; 12:335.
71. Bohm HJ. J Comput-Aided Mol Des. 1994; 8:243. [PubMed: 7964925]
72. Cao Y, Li L. Bioinformatics. 2014; 30:1674. [PubMed: 24563257]
73. Krammer A, Kirchhoff PD, Jiang X, Venkatachalam CM, Waldman M. J Mol Graphics Modell. 2005; 23:395.
74. Bohm HJ. J Comput-Aided Mol Des. 1998; 12:309. [PubMed: 9777490]

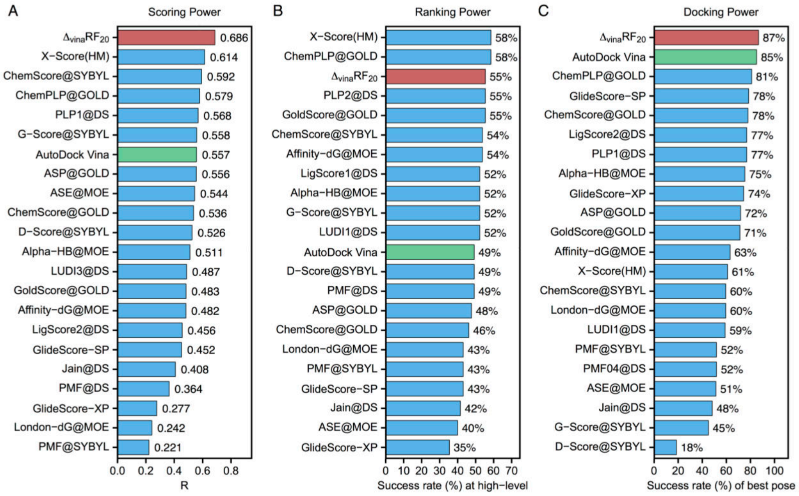


Figure 1.

Performance of 22 scoring functions in (A) scoring power measured by Pearson’s R, (B) ranking power in terms of high-level success rate and (C) docking power measured by the success rate when the best-scored pose is considered to match the native pose in CASF-2013 benchmark. $\Delta_{vina}RF_{20}$ is colored in red and AutoDock Vina is colored in green. All results colored in blue are obtained from reference[13].

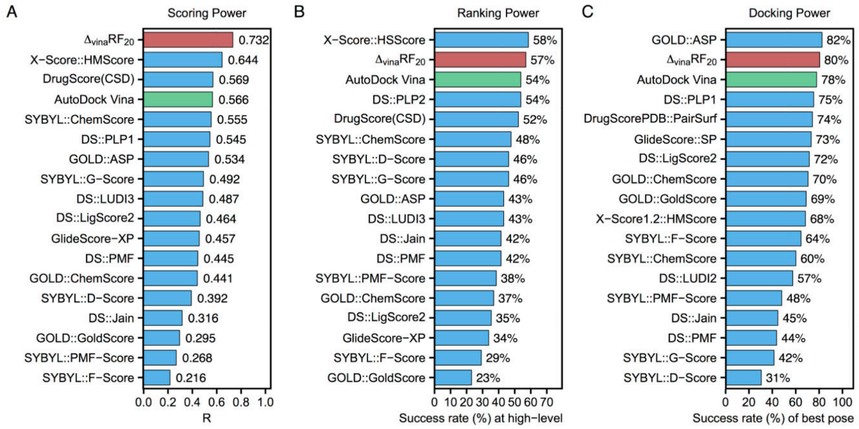


Figure 2.

Performance of 18 scoring functions in (A) scoring power measured by Pearson's R, (B) ranking power in terms of high-level success rate and (C) docking power measured by the success rate when the best-scored pose is considered to match the native pose in CASF-2007 benchmark. $\Delta_{vina}RF_{20}$ is colored in red and AutoDock Vina is colored in green. All results colored in blue are obtained from reference[12].

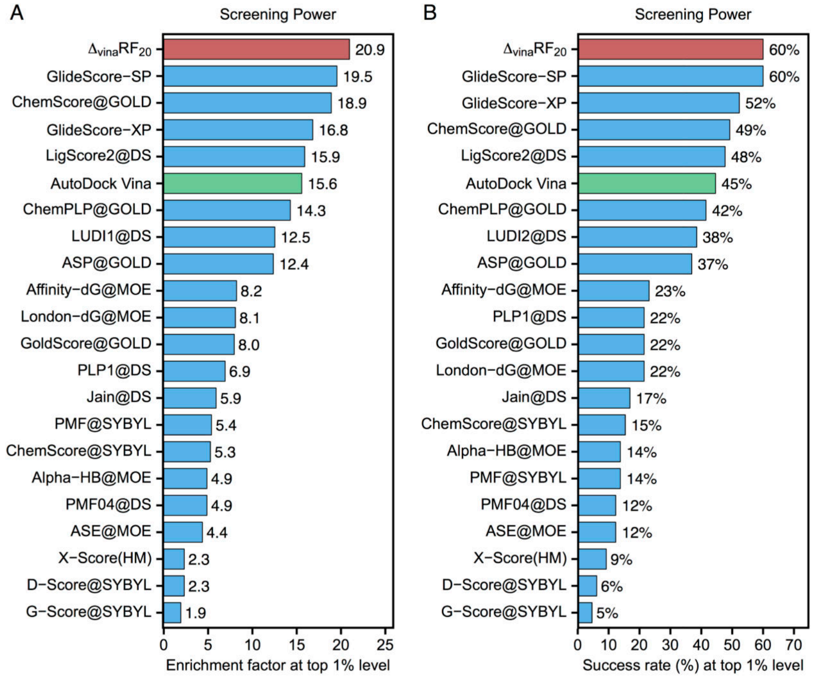


Figure 3.

Performance of 22 scoring functions in screening power measured by (A) enrichment factor and (B) success rate at top 1% level in CASF-2013 benchmark. $\Delta_{vina}RF_{20}$ is colored in red and AutoDock Vina is colored in green. All results colored in blue are obtained from reference[13].

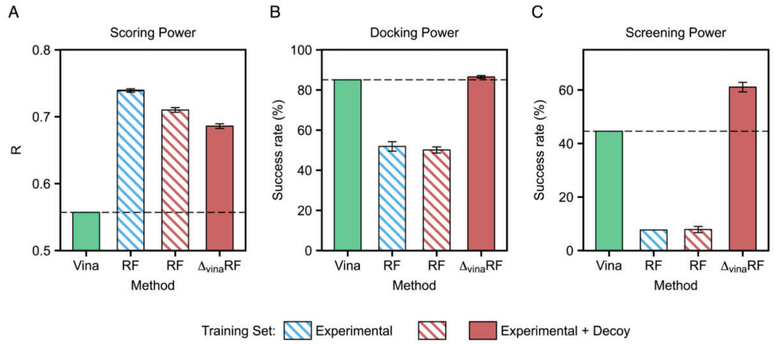


Figure 4.

CASF-2013 benchmark test performance of AutoDock Vina (colored in green), scoring function developed with RF approach (colored in blue) using experimental data alone and the same twenty features in $\Delta_{vina}RF_{20}$ scoring function, and scoring functions developed with RF approach and Δ_{vina}RF approach (colored in red) using the same twenty features and the same training set for the development of the $\Delta_{vina}RF_{20}$ scoring function. (A) Scoring power; (B) Docking power; (C) Screening power. Each set is run 10 times with different random seed for random forest and calculated by averaging over 10 performances except AutoDock Vina. The AutoDock Vina performance is also indicated by dashed line.

Table 1

20 Features in $\Delta_{\text{vina}}\text{RF}_{20}$

No.	Feature Description
AutoDock Vina Interaction Terms[a]	
1	$\text{non\_ hydrohobic}(a_1,a_2,d)=\begin{cases} 0, & a_1 \text{ or } a_2 \text{ is hydrophobic} \\ 1, & d_{\text{diff}}(a_1,a_2)<0.5 \\ 0, & d_{\text{diff}}(a_1,a_2) \geq 1.5 \\ 1.5-d_{\text{diff}}(a_1,a_2), & \text{otherwise} \end{cases}$
2	$\text{hydrogen\_ bond}(a_1,a_2,d)=\begin{cases} 0, & (a_1,a_2) \text{ do not form hydrogen bond} \\ 1, & d_{\text{diff}}(a_1,a_2)<-0.7 \\ 0, & d_{\text{diff}}(a_1,a_2) \geq 0.4 \\ \frac{d_{\text{diff}}(a_1,a_2)-0.4}{-1.1}, & \text{otherwise} \end{cases}$
3	$\text{solvation}(a_1,a_2,d)=\left[\left(\text{ASP}_{a_1}+\text{QASP}\times\lvert q_{\text{a}_1}\rvert\right)V_{a_2}+\left(\text{ASP}_{a_2}+\text{QASP}\times\lvert q_{a_2}\rvert\right)V_{a_1}\right]e^{-\left(\frac{d}{7.2}\right)^2}$
4–5	$\text{electrostatic}(a_1,a_2,d)=\frac{q_{a_1}\times q_{a_2}}{d^x}, \quad x=1 \text{ or } 2$
AutoDock Vina Ligand Dependent Terms	
6	number of heavy atoms
7	number of hydrophobic atoms
8	number of torsion
9	number of rotors
10	ligand length
bSASA Features[b]	
11	positive
12	negative
13	donor-acceptor
14	donor
15	acceptor
16	aromatic
17	hydrophobic
18	polar
19	halogen
20	total bSASA

[a] Interaction terms are from the AutoDock Vina source code.[54] d is the distance between two atoms, a_1 and a_2. d_{diff} is the surface distance calculated by $d_{\text{diff}} = d - R(a_1) - R(a_2)$, where $R(a_1)$ and $R(a_2)$ are the van der Waals radius of atom a_1 and a_2.[47] q is the atomic charge and V is the atomic volume. ASP and QASP refer to atomic solvation parameter and charge-based solvation parameter respectively.[59]

[b] The pharmacophore type definitions are presented in Table S3.