Combining Machine Learning and Pharmacophore-Based Interaction Fingerprint for in Silico Screening

Tomohiro Sato,[†,‡] Teruki Honma,[‡] and Shigeyuki Yokoyama*[,†,‡]

Department of Biophysics and Biochemistry, Graduate School of Science, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan, and RIKEN Systems and Structural Biology Center, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama 230-0045, Japan

Received October 1, 2009

In this study, we developed a new pharmacophore-based interaction fingerprint (Pharm-IF) and examined its usefulness for in silico screening using machine learning techniques such as support vector machine (SVM) and random forest (RF) instead of similarity-based ranking. Using the docking results of PKA, SRC, cathepsin K, carbonic anhydrase II, and HIV-1 protease, the screening efficiencies of the Pharm-IF models were compared to GLIDE score and the residue-based IF (PLIF) models. The combination of SVM and Pharm-IF demonstrated a higher enrichment factor at 10% (5.7 on average) than those of GLIDE score (4.2) and PLIF (4.3). In terms of the size of the training sets, learning more than five crystal structures enabled the machine learning models to stably achieve better efficiencies than GLIDE score. We also employed the docking poses of known active compounds, in addition to the crystal structures, as positive samples of training sets. The enrichment factors of the RF models at 10% using the docking poses for SRC and cathepsin K showed significantly higher values (6.5 and 6.3) than those using only the crystal structures (3.9 and 3.2), respectively.

INTRODUCTION

At the early stage of rational drug discovery, in silico screening based on molecular docking is often used to predict compounds that bind to the target protein. DOCK,[1] AutoDock,[2] and GLIDE[3] are the representative examples of widely used docking programs. The procedure utilized by these docking programs can be roughly divided into two steps. The first step is the exhaustive generation of adequate docking poses of a protein−ligand complex. The second step is the evaluation of the docking poses using scoring functions. According to the existing studies, the main issue with the current docking programs for in silico screening is the accuracy and versatility of the scoring functions.[4−11] Most of the conventional scoring functions use relatively simple formulas for the sake of the calculation speed. For example, with empirical scoring functions, the binding free energy of a protein−ligand complex is decomposed into several energy terms such as hydrogen bond, ionic attraction, and hydrophobic interaction. The coefficients of the terms are defined by regression to fit experimentally determined protein−ligand complexes with binding affinities such as K_d, K_i, and IC_{50}. A disadvantage of these methods is that the coefficients are dependent on the small amount of biased structure−activity relationship data used for the regression, affecting the applicability to targets of different classes. It is difficult to create a completely versatile scoring function for all types of proteins from the limited data reported so far. In addition, these simplified scoring functions cannot properly consider weak interactions such as dispersion forces and π-orbital interactions, solvation−desolvation effects, and entropy.

During the past decade, increasing numbers of complex structures were solved by the development of high throughput X-ray crystallography[12] and NMR[13] methods. As of 2008, 54 996 three-dimensional structures were registered in the Protein Data Bank (PDB).[14] The analysis and exploitation of the PDB data for in silico screening have been continuously investigated for various methods, such as knowledge-based weighting for ligand-based similarity search[15,16] and molecular field analysis of docking poses (AFMoC).[17] Both methods use the structural information to create target-specific scoring methods. This target-specific scoring reportedly performed better than the general scoring functions when a sufficient amount of information was available.[18] More recently, the concept of the interaction fingerprint (IF) was developed, to describe the protein−ligand interactions of complex structures. The first IF, named SIFt, was reported by Deng et al. in 2004.[19] SIFt describes the intermolecular interactions between a ligand and the residues on the binding site of a protein as a bit string by (i) whether the residue is in contact with the ligand; (ii) whether any main chain atom is involved in the contact; (iii) whether any side chain atom is involved in the binding; (iv) whether a polar interaction is involved; (v) whether a nonpolar interaction is involved; (vi) whether the residue provides hydrogen bond acceptor(s); and (vii) whether the residue provides hydrogen bond donors. To date, some variations of IFs have been developed for various purposes.[20−28] For example, PLIF implemented on Molecular Operating Environment (MOE)[29] classifies the interactions between a ligand and the residues on the binding site into various types of weak and strong interactions, using interaction assessment functions built in

* Corresponding author phone: +81-3-5841-4395; fax: +81-3-5841-8057; e-mail: yokoyama@biochem.s.u-tokyo.ac.jp.
† The University of Tokyo.
‡ RIKEN.

10.1021/ci900382e
Published on Web 12/29/2009

MOE. The IFs based on amino acid residues were recently reviewed.[30,31] As another interaction description method, APIF, recently developed by Pèrez-Nueno et al.,[27] calculates the fingerprints based on the relative positions of the interacting protein−ligand atom pairs. APIF is classified as an atom-based IF and showed better screening efficiency performance than the residue-based IF.

In this Article, we developed a new atom-based IF named Pharm-IF. The fingerprint of Pharm-IF is calculated from the distances of pairs of ligand pharmacophore features that interact with protein atoms. The screening efficiency of Pharm-IF was compared to those of GLIDE score and PLIF. Pharm-IF can detect important geometrical patterns of ligand pharmacophores. From a medicinal chemists' point of view, the detected patterns of ligand pharmacophores would facilitate an understanding of the structure−activity relationship of the target protein.

In most cases, IFs have been combined with similarity metrics such as the Tanimoto coefficient (*Tc*), to rank or filter compounds when performing in silico screening. The only exception is the research by Venhorst,[25] which combined the IF with naïve Bayesian classifier (NBC) and demonstrated its usefulness for scaffold hopping. The recent increase in the amount of experimentally determined 3D information on the protein−ligand complexes facilitates the use of more complicated and powerful machine learning algorithms. Machine learning has already been applied to the field of drug discovery, such as the prediction of drug-likeness,[32] molecular and biological activities,[33−43] and ADME/Tox profiles of small compounds,[44] using ligand-based molecular fingerprints or descriptors. In those applications, state-of-the-art nonlinear statistical methods, including artificial neural network (ANN), random forest (RF),[45] and support vector machine (SVM),[46] have been used in addition to NBC and linear discrimination analysis (LDA), and provided more accurate and versatile classification models. In this study, four machine learning algorithms, NBC, RF, SVM, and ANN, were used to build prediction models based on PLIF and our Pharm-IF to improve the screening efficiencies. The structural features, which were important for discrimination between the active and inactive compounds by Pharm-IF models, were also analyzed using the Gini coefficient of decision trees in RF models to understand the structure−activity relationship of each target.

The number of crystal structures of protein−ligand complexes for training is critically important, because, in general, larger and more diverse training sets enable us to build more accurate machine learning models than those achieved with smaller training sets. To clarify the appropriate size of the training set, we assessed the effects of the size of the training sets on the screening efficiencies by building models using 1−80 crystal structures as positive cases of the training sets. To create an effective learning model in the case where only limited numbers of crystal structures of protein−ligand complexes are available, methods to complement the data in the training set are needed. We used the docking poses of the known active compounds as positive samples for learning, in addition to the experimentally determined crystal structures.

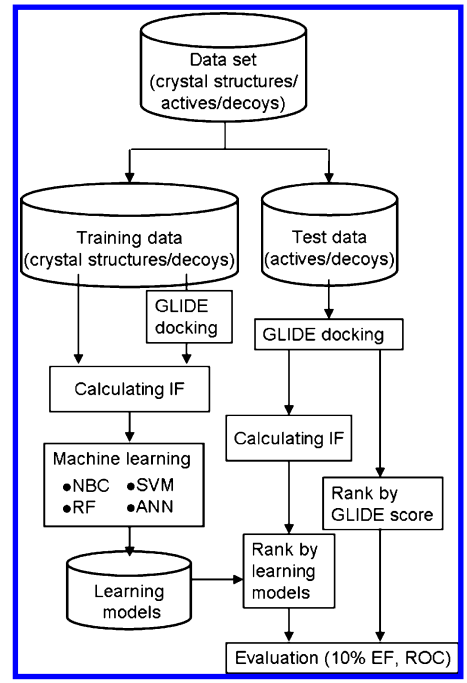


Figure 1. The work flow of the evaluation of GLIDE score and the machine learning models of PLIF and Pharm-IF.

METHODS

Data Set. To evaluate the performance of Pharm-IF and machine learning, the in silico screening of PKA, SRC, carbonic anhydrase II, cathepsin K, and HIV-1 protease inhibitors was performed. PKA is an intracellular serine/threonine kinase mediated by cAMP. Fifty-nine complex structures of PKA were registered in the PDB. SRC is a nonreceptor tyrosine kinase, and nine complex structures of SRC were determined. Carbonic anhydrase is an enzyme associated with the reversible reaction of hydration of CO_2 and dehydration of carbonic acid. There are seven isozymes of human carbonic anhydrase. Carbonic anhydrase II is involved in the mediation of intraocular pressure. Eighty-two complex structures of carbonic anhydrase II were already reported. Cathepsin K is a lysosomal cysteine protease belonging to the papain super family. Cathepsin K is expressed in osteoclasts and is involved in bone resorption. Cathepsin K could be a relatively difficult target protein for drug design, due to its shallow binding site.[47] Twenty-five complex structures of cathepsin K were available. HIV-1 protease is an enzyme required for the precursor processing of viral proteins. HIV-1 protease has a large binding pocket between its two subunits. Two β-strands, located above the binding site, cover the site. One hundred and ninety-seven complex structures of HIV-1 protease were determined.

For each target protein, two compound sets, containing known active compounds and decoys, were created. One is the training set for model building by machine learning, and the other is the test set for the evaluation of the screening efficiencies of GLIDE score, PLIF models, and Pharm-IF models. Figure 1 briefly describes the work flow of the model building using the training sets and the evaluation of the screening efficiencies using the test sets. The training sets used the experimentally determined complex structures of PKA, SRC, carbonic anhydrase II, cathepsin K, and HIV-1 protease as the positive samples, and the docking poses of 2000 decoy compounds randomly selected from the Pub-

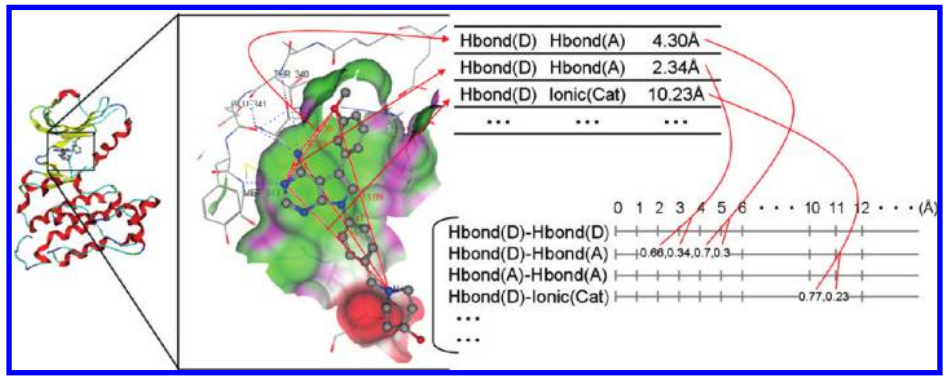


Figure 2. Brief procedure of the calculation of Pharm-IF.

Chem database[48] as the negative samples. Each decoy compound was docked to the target proteins using GLIDE Standard Precision (SP) mode. Five docking poses were generated for each compound. For the positive samples of the test set, active compounds of the target proteins were selected from StARLITe.[49] StARLITe is a database containing biological activity and/or binding affinity data between various compounds and proteins. For each target, the active compounds ($IC_{50} \leq 10\ \mu M$) were divided into 100 clusters using hierarchical clustering by the Ward method, according to the Euclid distances between their 2D structural fingerprints (public MACCS keys). The compound with the highest inhibitory activity was selected from each cluster. The 100 active compounds obtained for each target were docked to their target protein, and five docking poses for each active compound were used as positive samples of the test set. As the decoys for the test set, five docking poses were generated for each of the 2000 compounds randomly selected from the PubChem database, in the same manner as those for the training sets.

Docking. The docking poses of the decoys and the active compounds for which crystal structures were not experimentally determined were generated using the SP mode of GLIDE, developed by Schrödinger, Inc. As the protein structures for docking, high-resolution crystal structures of the protein−inhibitor complexes with high inhibitory activities were selected from the PDB. PDB entry 1re8 was selected for PKA, 2h8h was selected for SRC, 1if7 was selected for carbonic anhydrase II, 1u9w was selected for cathepsin K, and 1pro was selected for HIV-1 protease. The resolutions of the structures are 2.10, 2.20, 1.98, 2.30, and 1.80, respectively. For the preparation of the docking, Protein Preparation Wizard (Schrödinger, Inc.) was used. The hydrogen atoms of the protein were added, and their positions were optimized with Protein Preparation Wizard. Using Pipeline Pilot[50] of SciTegic, tautomers, stereoisomers, and protonation/deprotonation forms at pH 7.4 of the active and decoy compounds were enumerated. The additional ring conformations of the compounds were then generated by LigPrep (Schrödinger, Inc.). From the resulting docking poses, five poses were selected for each compound, using GLIDE score. Other settings of GLIDE were set to the default values.

PLIF. PLIF is a residue-based IF included in MOE ver. 2007.09, distributed by Chemical Computing Group, Inc. The fingerprint of a protein−ligand complex was calculated by the presence or absence of six types of intermolecular interactions: hydrogen bond with side chain donor, hydrogen bond with side chain acceptor, hydrogen bond with backbone donor, hydrogen bond with backbone acceptor, ionic attraction, and surface contact. For each residue, a strong interaction and a weak interaction of the six types of interaction were checked and encoded by 12 bits. As a result, the fingerprint of a complex was defined by 12 times the number of binding site residue bits. The resulting fingerprints were extracted from MOE for the analysis by editing the built-in source code of PLIF.

Table 1. The Machine Learning Algorithms Used To Build Learning Models Based on Pharm-IF and PLIF

algorithm	abbreviation	description
naïve Bayesian classifier	NBC	probabilistic model based on Bayesian statistics
random forest	RF	ensemble of decision trees
support vector machine	SVM	nonlinear discrimination using kernel method
artificial neural network	ANN	decision by multiple layer network model

Pharm-IF. Using Pharm-IF, the fingerprint of a complex was calculated by the following three steps. At first, protein−ligand interactions were detected from the complex structure. Next, pairs of any two interactions (interaction pairs) were created. The interaction pairs were characterized by the pharmacophore features of their ligand atoms and their distance. Finally, a vector was created by counting the number of each type of interaction pair.

Protein−ligand interactions were detected using the functions built in MOE. Interactions were classified into six types: hydrogen bond with ligand acceptor, hydrogen bond with ligand donor, hydrogen bond in which the roles of ligand and protein atoms could not determined, ionic interaction with ligand cation, ionic interaction with ligand anion, and hydrophobic interaction. The presence or absence of the interactions, except for the hydrophobic interaction, was determined in the same manner as PLIF, using the threshold of the weak interactions. Hydrophobic interactions were detected when more than 1% of the van der Waals surfaces of two hydrophobic atoms contacted each other. When the hydroxyl groups of the ligand and the protein formed a hydrogen bond, the roles of the two atoms could not be determined, because a hydroxyl group can be both a hydrogen-bond donor and acceptor. In such cases, PLIF treats the hydrogen bond as two hydrogen bonds with different donor and acceptor pairs. However, Pharm-IF treats it as a different kind of hydrogen bond, because of the slightly superior results of Pharm-IF using this criterion.

Interaction pairs were created by all possible combinations of detected protein−ligand interactions. An interaction pair was classified by the pharmacophore features of its ligand atoms and their distance (Figure 2). APIF by Pèrez-Nueno uses the distances between both the ligand atoms and the protein atoms for the classification. The Pharm-IF procedure focuses on the 3D geometry of the ligand pharmacophores in the docking pose and is an extended method of the pharmacophore fingerprints based on only ligand information, such as CATS 2D,[51] 3D,[52] and Similog keys.[53]

For example, the fingerprint of the complex shown in Figure 2 is calculated as follows. At first, the intermolecular interactions are identified in the same manner as PLIF. Next, all possible interaction pairs are created. Each interaction pair is characterized by the pharmacophore features of the ligand atoms and their distance. To calculate the resulting matrix, each interaction pair is assigned to the corresponding bin. An interaction pair of two hydrogen bonds, whose ligand atoms are a donor and an acceptor that are 4.3 Å apart from each other, is assigned to the vector corresponding to this hydrogen bond pair. The vectors consist of bins corresponding to distances. As the value of this interaction pair, 0.7 was assigned to the bin of 4 Å and 0.3 was assigned to the bin of 5 Å, to describe the distance of 4.3 Å. Finally, the matrix was calculated by the summation of the values of all of the interaction pairs. The following equation describes the detailed definition of the fingerprint of a protein−ligand complex.

$$H_{t,k} = \sum_{i \in I_t} A_k(i) \tag{1}$$

where H is the interaction fingerprint of a protein−ligand complex by Pharm-IF, t is the pair of six types of pharmacophore features such as "H-bond donor - hydrophobic", $k = 1, 2, 3, ...$ stands for the corresponding bins of the distances (Å) between ligand atoms, I_t is the whole set of the interaction pairs classified as type t, and i is a member of I_t.

$$A_k(i) = \begin{cases} 0 & , \text{if} |k - d_i| \geq 1 \\ 1 - |k - d_i| & , \text{otherwise} \end{cases} \tag{2}$$

where d_i represents distances between ligand atoms of i (Å).

When a real valued descriptor such as the distance is converted to a bit of a fingerprint, the value is often assigned to only one predefined bin with a range that covers the value. This type of procedure cannot reflect slight differences of the distances in the range of the same bin, and thus could treat two similar values differently when their distances cross over the threshold of the bin. Pharm-IF uses a simple weighting procedure by the distances, as described by Minai et al.,[54] for counting the number of interaction pairs to avoid this problem.

Machine Learning Algorithms. In this study, the machine learning algorithms shown in Table 1 were used to build the learning models based on Pharm-IF and PLIF. For model building, R component collection and Data Modeling collection of Pipeline Pilot were employed.

Naïve Bayesian Classifier. NBC is a discrimination algorithm based on Bayesian statistics. A discrimination model is built from the distribution of each descriptor of active and inactive samples. NBC assumes that the distributions of all of the descriptors are independent of each other. The resulting probability is calculated by multiplying the posterior probability calculated for each of the descriptors. An NBC model is optimized by using maximum likelihood procedures. The parameters were set to the default values of Pipeline Pilot.

Random Forest. RF creates a lot of decision trees and uses the ensemble of the trees for the discrimination of the samples. The decision trees are created by many data subsets together with the descriptor subset. The data subsets are selected by bootstrap sampling, and the descriptor subsets are chosen by random selection. RF uses the out-of-bag (OOB) method for cross-validation. One-third of the training data is separated and used for validation of the learning model. The overall accuracy of the RF model is calculated by the average of those of all decision trees. In this study, the number of decision trees was set to 1000. The default values of Pipeline Pilot were used for the other parameters.

Support Vector Machine. SVM models nonlinearly discriminate two classes of compounds, by mapping data vectors to a very high-dimensional descriptor space and finding a hyperplane that separates the two classes with the largest margin. The most significant difference between SVM and simple linear discrimination is the so-called "kernel trick". In this study, a radial basis function (RBF) kernel was used for obtaining a complicated nonlinear separating hyperplane. The gamma for the RBF kernel and the "C" value of the constant for the slacks variant were optimized by 5-fold cross validation. Other parameters were set to the default values of Pipeline Pilot.

Artificial Neural Network. ANN is a mathematical model to emulate the procedure of the activity of the human brain. ANN is a network consisting of three layers, the input layer, the hidden layer, and the output layer. In this study, back-propagation neural network (BPNN) based on BFGS algorithm was used. The number of nodes in the hidden layer was set to 1 or 2 and was optimized by 5-fold cross validation. Other parameters were set to the default values of Pipeline Pilot.

Similarity Search. As a standard for the comparison with the performances of the machine learning models, similarity searches using PLIF and the Tanimoto coefficient (Tc), described in eq 3, were performed.

$$Tc = \frac{\sum_i a_i b_i}{\sum_i (a_i^2 + b_i^2 - a_i b_i)} \tag{3}$$

where a_i, b_i are the ith bits of the fingerprint of the two protein−ligand complexes.

After the fingerprints were calculated, the similarities between the docking poses and the crystal structures were calculated by the following two methods. The first method is the procedure normally used for the similarity search, in which the PLIFs of the docking poses are compared to those of all crystal structures, and the maximum similarity was used for the score (max_tc). The second method uses the profile of the fingerprint of the complexes by calculating the average of each bit of the fingerprint of all crystal structures, in the same manner as p-SIFt.[21] The docking poses were

evaluated by the *Tc* between their fingerprints and the averaged fingerprint of the crystal structures (p_tc).

The screening efficiencies of the docking score, the machine learning models, and the similarity searches were evaluated using enrichment factor at 10% (10% EF) and ROC score. EF is one of the most famous measures for evaluating the screening efficiency, and it indicates the ratio of the number of obtained active compounds by in silico screening against that generated by random selection at the predefined sampling percentage. EF is often used to evaluate the early recognition property of screening methods. Usually, only a small fraction of compounds is selected by in silico screening for drug discovery. In general, 0.01−1% of the compounds are selected from a huge compound database (10 000−1 000 000 compounds) in real in silico screening. However, in our test screenings, 0.01−1% EF had a quite large deviation, because 0.01−1% of the test data sets (2100 compounds) is statistically insufficient.[55] To assess the early recognition with low deviation, 10% EF was used in this study. ROC score evaluates the entire range (0−100% sampling) of the screening efficiency, while EF evaluates the screening efficiency of only a particular sampling percentage. ROC score is defined as the area under the receiver operation characteristic curve, which plots the ratio of true positive samples (detected active compounds) on the axis of false positive fractions and ranges from 0 (0%) to 1 (100%). These two measures were used to evaluate the screening efficiency of the methods in this study.

Identification of the Important Interaction Pairs. The important geometrical features of the ligand pharmacophores for the target proteins were identified by calculating the importance of each element of Pharm-IF to the RF models. The importance of a variable was calculated by the difference of the Gini index. The Gini index of a node in a decision tree indicates the impurity of the samples in the node. If many inactive compounds (impurities) are mixed in with active compounds, then the Gini index shows a high value. The value of the Gini index is decreased by adding a new discrimination rule and eliminating the inactive compounds. The difference of the Gini index (ΔGini) of a variable was calculated by comparing the Gini index of the original node and the summation of the Gini indexes of two nodes divided by the variable. The importance of each element of Pharm-IF was calculated by the ΔGini values of all decision trees in the OOB validation of the RF model building.

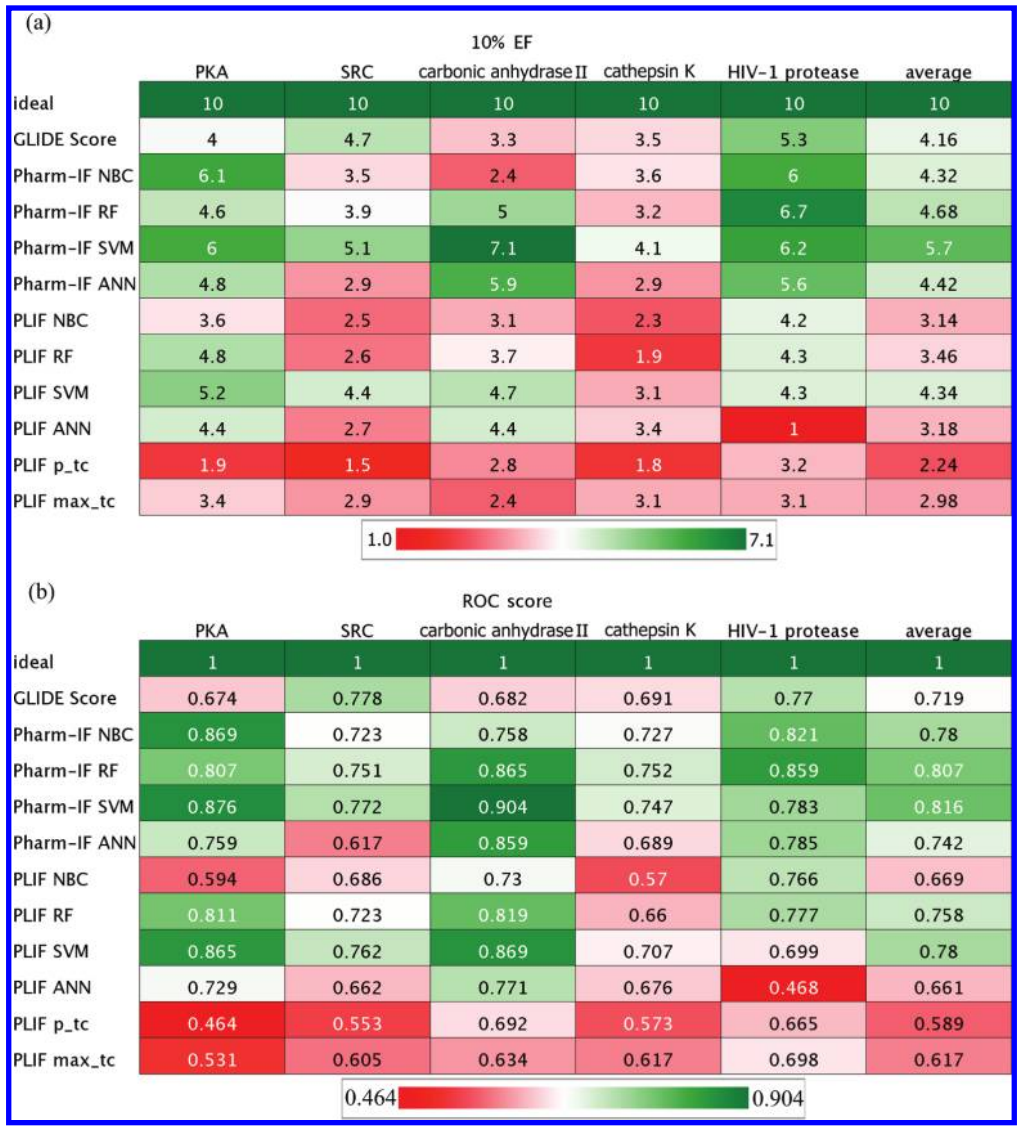
Effects of Training Set Size on Screening Efficiency. For each machine learning algorithm, the effect of the number of experimentally determined protein ligand complex structures on the screening efficiency was analyzed. The data of carbonic anhydrase II were used for the analysis, because the number of determined complex structures (82) is sufficient for making the various sizes of training sets, and the learning models using Pharm-IF recorded the best efficiency among the five target proteins. As the positive samples in the training set, 1, 3, 5, 10, 20, 40, 60, and 80 crystal structures were randomly selected 10 times, respectively. The test set and the negative samples in the training set were not changed from those of the former data set. The screening efficiencies of the learning models using each size of training set were evaluated by the average of the 10% EF and ROC scores of the 10 trials. Using one and three complexes, the cross-validations of the learning models of SVM and ANN were disabled. The RF models using only one complex structure were not created because OOB validation required at least three positive samples.

Machine Learning Using Docking Poses as Positive Samples for the Training Set. To improve the learning models for the target proteins when only a small number of crystal structures are available, the use of the training sets including the docking poses of active compounds as positive samples was tested. SRC and cathepsin K were used for the analysis, because they have fewer crystal structures (9 and 25) than the other three targets. From the known active compounds for which crystal structures with their targets were not determined, 1, 3, 5, 10, 20, 40, 60, and 80 active compounds were randomly selected 10 times, respectively. Five docking poses for each of the selected compounds were generated using GLIDE and were used as positive samples for the training set, in addition to the experimentally determined complex structures. The test set and the negative samples for the training set were not changed from those of the former data set. The screening efficiencies of the learning models using each number of docking poses were evaluated by the averages of the 10% EF and the ROC score of the 10 trials. In this study, we used five docking poses of each active compound as the positive examples, because this procedure generated higher enrichment factors in a preliminary test than those obtained using 1 pose of each active compound. The second to fifth docking poses would include proper complex structures as well as the top-ranked docking poses, and thus would contribute to the learning of interaction patterns.

RESULTS AND DISCUSSION

Evaluation of Screening Efficiency. The results of the in silico screening for PKA, SRC, carbonic anhydrase II, cathepsin K, and HIV-1 protease, using GLIDE score, the PLIF models, and the Pharm-IF models, are shown in Figure 3. In total, the combination of Pharm-IF and SVM recorded a higher 10% EF (5.7 on average) than those of both GLIDE score (4.2) and the PLIF models (4.3 (SVM) and 3.0 (max_tc)) against all five targets. The average of the ROC score by Pharm-IF SVM also generated the highest value (0.816) among all of the ranking methods. Generally, the learning models using Pharm-IF provided better screening efficiencies in the cases of targets with larger numbers of complex structures, such as PKA (56 crystal structures), carbonic anhydrase II (82), and HIV-1 protease (197), recording a 10% EF of 5.5 and a ROC score of 0.829, on average, among all of the Pharm-IF learning models. On the other hand, the learning models displayed relatively worse screening efficiency for SRC (9) and cathepsin K (25), recoding a 10% EF of 3.7 and a ROC score of 0.722, on average. The Pharm-IF-SVM models showed better performances for both types of targets (10% EF, 6.4; ROC score, 0.854 on average for PKA; carbonic anhydrase II and HIV-1 protease and 10% EF, 4.6; ROC score, 0.759 on average for SRC and cathepsin K), as compared to GLIDE score (10% EF, 4.1; ROC score, 0.709 and 10% EF, 4.2; ROC score, 0.734, respectively) and the PLIF models (10% EF, 4.7; ROC score, 0.811 at most and 10% EF, 3.8; ROC score, 0.735 at most, respectively).

In particular, GLIDE score showed poor screening efficiency for carbonic anhydrase II (10% EF, 3.3; ROC score,



(a) 10% EF

	PKA	SRC	carbonic anhydrase II	cathepsin K	HIV-1 protease	average
ideal	10	10	10	10	10	10
GLIDE Score	4	4.7	3.3	3.5	5.3	4.16
Pharm-IF NBC	6.1	3.5	2.4	3.6	6	4.32
Pharm-IF RF	4.6	3.9	5	3.2	6.7	4.68
Pharm-IF SVM	6	5.1	7.1	4.1	6.2	5.7
Pharm-IF ANN	4.8	2.9	5.9	2.9	5.6	4.42
PLIF NBC	3.6	2.5	3.1	2.3	4.2	3.14
PLIF RF	4.8	2.6	3.7	1.9	4.3	3.46
PLIF SVM	5.2	4.4	4.7	3.1	4.3	4.34
PLIF ANN	4.4	2.7	4.4	3.4	1	3.18
PLIF p_tc	1.9	1.5	2.8	1.8	3.2	2.24
PLIF max_tc	3.4	2.9	2.4	3.1	3.1	2.98

1.0 — 7.1

(b) ROC score

	PKA	SRC	carbonic anhydrase II	cathepsin K	HIV-1 protease	average
ideal	1	1	1	1	1	1
GLIDE Score	0.674	0.778	0.682	0.691	0.77	0.719
Pharm-IF NBC	0.869	0.723	0.758	0.727	0.821	0.78
Pharm-IF RF	0.807	0.751	0.865	0.752	0.859	0.807
Pharm-IF SVM	0.876	0.772	0.904	0.747	0.783	0.816
Pharm-IF ANN	0.759	0.617	0.859	0.689	0.785	0.742
PLIF NBC	0.594	0.686	0.73	0.57	0.766	0.669
PLIF RF	0.811	0.723	0.819	0.66	0.777	0.758
PLIF SVM	0.865	0.762	0.869	0.707	0.699	0.78
PLIF ANN	0.729	0.662	0.771	0.676	0.468	0.661
PLIF p_tc	0.464	0.553	0.692	0.573	0.665	0.589
PLIF max_tc	0.531	0.605	0.634	0.617	0.698	0.617

0.464 — 0.904

Figure 3. The screening efficiencies for PKA, SRC, carbonic anhydrase II, cathepsin K, and HIV-1 protease and their averages. (a) 10% EF of the result by each screening method. (b) ROC score of the result by each screening method. The colors of the cells correspond to the efficiencies, from red (the worst result) to white (the average) and green (the best result).

0.682) and cathepsin K (10% EF, 3.5; ROC score, 0.691). The Pharm-IF models achieved the largest improvement for carbonic anhydrase II. In the case of carbonic anhydrase II, most of the active compounds shared the same interaction patterns (i.e., multiple hydrogen bonds by sulfonamide groups), resulting in the high efficiency of the learning models using Pharm-IF. Comparing the results from PKA and SRC, GLIDE recorded better 10% EF and ROC score (4.7 and 0.778) for SRC than those for PKA (4 and 0.674). In contrast, both Pharm-IF and PLIF generated contrary results. Because both PKA and SRC are members of a protein kinase family and share some structural features in the binding site such as important hydrogen bonds at the hinge region of the ATP-binding pocket, the differences between the performances of GLIDE score and the models using both IFs were thought to mainly result from the difference in the number of available crystal structures used for the learning (PKA, 56; SRC, 9). For cathepsin K, none of the screening methods could effectively detect the active compounds (10% EF, 3.0; ROC score, 0.674 on average), as compared to the other targets (10% EF, 4.1; ROC score, 0.734 on average). One possible reason is that the generation of the proper docking pose of cathepsin K seems to be problematic due to the shallow binding site. The difficulty of using the cathepsin K binding site in the docking process affected not only GLIDE score but also the learning models because the docking poses generated by GLIDE were shared with all of the scoring methods.

The combination of the machine learning algorithms with PLIF also improved the screening efficiencies, as compared to the conventional similarity-based ranking with PLIF. The SVM models of PLIF recorded better 10% EF values (5.2, 4.4, 4.7, 3.1, and 4.3) than those using *Tc* (1.9, 1.5, 2.8, 1.8, and 3.2 (p_tc), 3.4, 2.9, 2.4, 3.1, and 3.1 (max_tc)) for all five targets. This result suggested that the machine learning algorithms could also improve the screening efficiencies of the residue-based IFs. Comparing the results of the PLIF models and GLIDE score, the PLIF models provided better screening efficiencies for PKA and carbonic anhydrase II (10% EF, 5.2 and 4.7; ROC score, 0.865 and 0.869) than those of GLIDE score (10% EF, 4.0 and 3.3; ROC score, 0.674 and 0.682). GLIDE score gave comparable or better

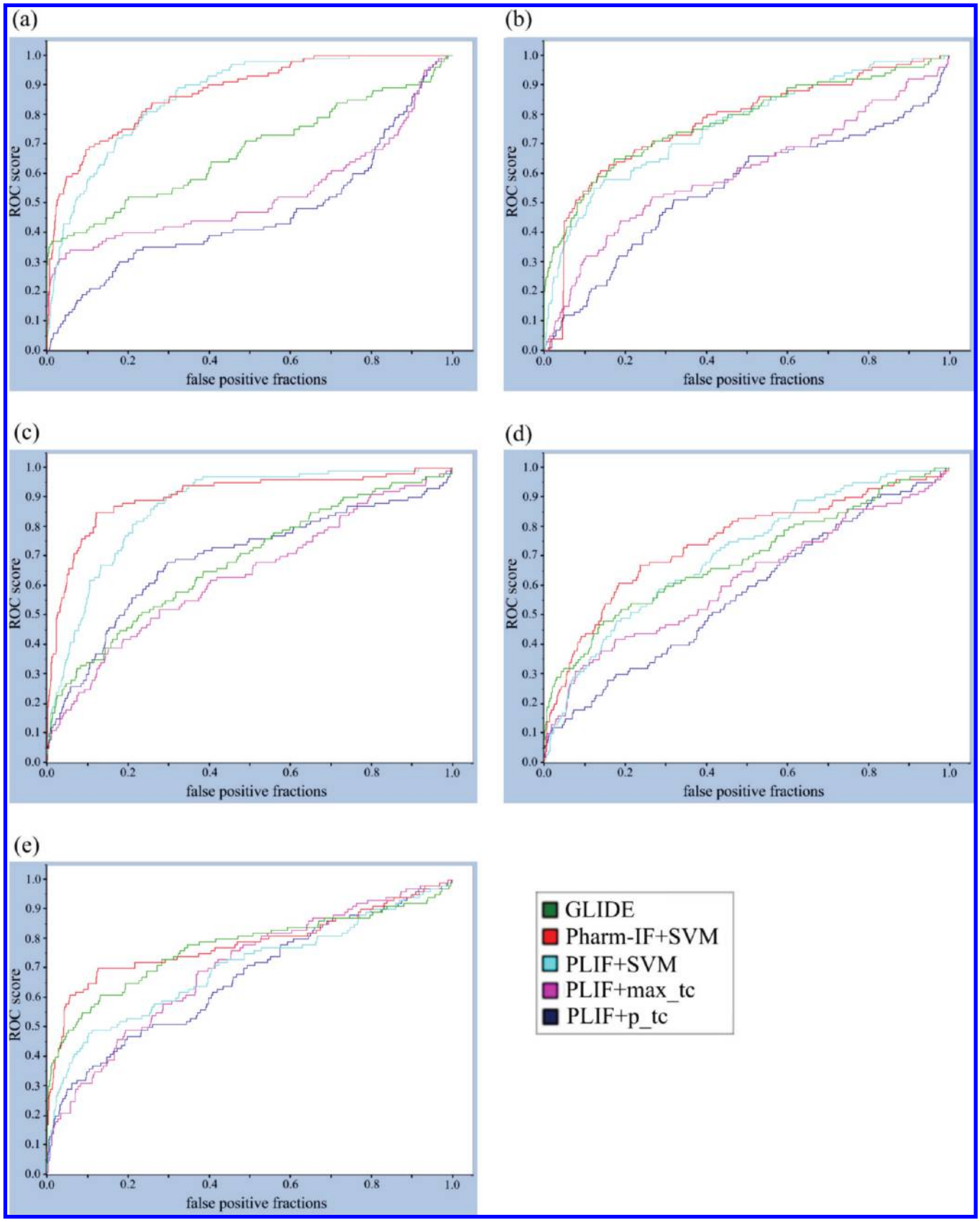


Figure 4. The ROC curves of the screening using GLIDE score, the Pharm-IF-SVM models, the PLIF-SVM models, and similarity search using PLIF for (a) PKA, (b) SRC, (c) carbonic anhydrase II, (d) cathepsin K, and (e) HIV-1 protease.

performance for SRC and cathepsin K, because of the relatively fewer crystal structures.

In terms of the machine learning algorithms, SVM recorded the best performances using both Pharm-IF (10% EF, 5.7; ROC score, 0.816 on average) and PLIF (10% EF, 4.3; ROC score, 0.780 on average) among the four machine learning algorithms. According to the ROC curves, shown in Figure 4, the combination of Pharm-IF and SVM showed

Table 2. The Top 10 Important Interaction Pairs of the RF Model for PKA

pair of interacting ligand atoms	distance (Å)	importance (Gini)
ionic (L = "+") hydrophobic	1	2.16
ionic (L = "+") hydrophobic	3	2.01
ionic (L = "+") hydrophobic	2	1.79
ionic (L = "+") hydrophobic	4	1.77
ionic (L = "+") hydrophobic	6	1.76
ionic (L = "+") Hbond (L = "D")	0	1.70
ionic (L = "+") hydrophobic	5	1.68
ionic (L = "+") hydrophobic	7	1.62
ionic (L = "+") hydrophobic	9	1.57
ionic (L = "+") hydrophobic	8	1.32

excellent early recognition for PKA, carbonic anhydrase II, and HIV-1 protease. The enrichment plots of these results are shown as Supporting Information. Other than SVM, RF outperformed NBC and ANN. The Pharm-IF-RF models recorded a better 10% EF value (4.7) on average for the five targets than that generated by GLIDE score (4.2). The differences between the RF models and the SVM models were especially apparent when the number of crystal structure was relatively small. In fact, the Pharm-IF-RF models recorded inferior efficiencies as compared to GLIDE score for SRC and cathepsin K, while the SVM models outperformed GLIDE score for all five targets. This result suggested that SVM can create efficient prediction models even with relatively few training samples, as compared to the other learning algorithms. The NBC and ANN models based on both Pharm-IF and PLIF resulted in approximately the same level of efficiency. In some cases, these two learning algorithms failed to create efficient models when a sufficient number of crystal structures were available. The ROC score of the PLIF-NBC model for PKA was 0.594. The 10% EF value of the Pharm-IF-NBC model for carbonic anhydrase II was 2.4. The efficiency of the PLIF-ANN model for PKA even fell below random choice. The Pharm-IF-SVM and RF models generated values of at least 3.2 for the 10% EF and 0.751 for the ROC score for all of the targets.

Important Interaction Pairs for Discrimination. PKA. The 10 most important interaction pairs of the Pharm-IF-RF model for PKA are shown in Table 2. The pairs of ionic attractions with ligand cations and hydrophobic interactions with short distances were regarded as the most important features. At the entrance of the ATP binding site of PKA, Glu127, Glu170, and Asp184 expose negatively charged atoms, resulting in the common ionic attractions with the nitrogen cation of the ligands (Figure 5). The hydrophobic environment generated by residues such as Gly50, Gly52, Val57, and Leu173 is located near the negatively charged residues and formed hydrophobic interactions with the ligand carbon atoms. Thus, the ligand cations and the hydrophobic atoms within about 1−4 Å were regarded as the most characteristic motifs to discriminate PKA inhibitors from decoys.

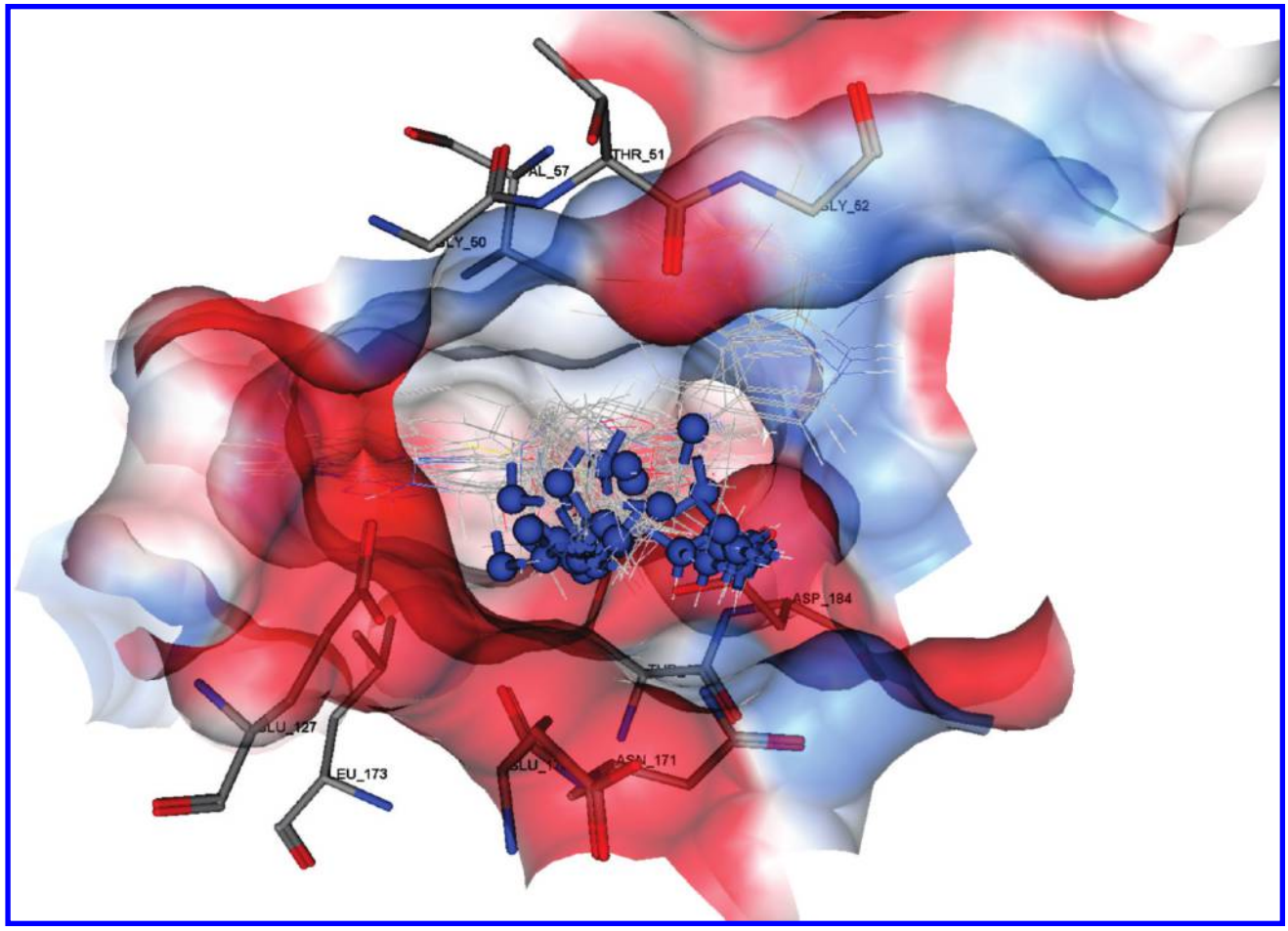


Figure 5. The binding site of PKA. The color of the surface corresponds to the electrostatic properties: red, negative; white, neutral; and blue, positive. The ligands around the negatively charged surface in the complex structures are overlaid. The nitrogen cations of the ligands are highlighted as blue balls.

Table 3. The Top 10 Important Interaction Pairs of the RF Model for SRC

pair of interacting ligand atoms		distance (Å)	importance (Gini)
Hbond (L = "D")	Hbond (L = "A")	3	0.43
Hbond (L = "A")	hydrophobic	7	0.31
hydrophobic	hydrophobic	8	0.31
Hbond (L = "D")	hydrophobic	7	0.30
Hbond (L = "A")	hydrophobic	6	0.30
Hbond (L = "A")	hydrophobic	1	0.27
Hbond (L = "D")	Hbond (L = "A")	2	0.25
Hbond (L = "D")	hydrophobic	6	0.25
Hbond (L = "D")	hydrophobic	2	0.23
Hbond (L = "A")	hydrophobic	4	0.23

Table 4. The Top 10 Important Interaction Pairs of the RF Model for Carbonic Anhydrase II

pair of interacting ligand atoms		distance (Å)	importance (Gini)
Hbond (L = "D")	Hbond (L = "A")	3	6.42
Hbond (L = "D")	Hbond (L = "A")	2	6.35
Hbond (L = "A")	Hbond (L = "A")	0	5.98
hydrophobic	hydrophobic	6	3.72
hydrophobic	hydrophobic	0	3.13
hydrophobic	hydrophobic	1	3.00
hydrophobic	hydrophobic	7	2.98
hydrophobic	hydrophobic	2	2.76
hydrophobic	hydrophobic	4	2.68
hydrophobic	hydrophobic	3	2.47

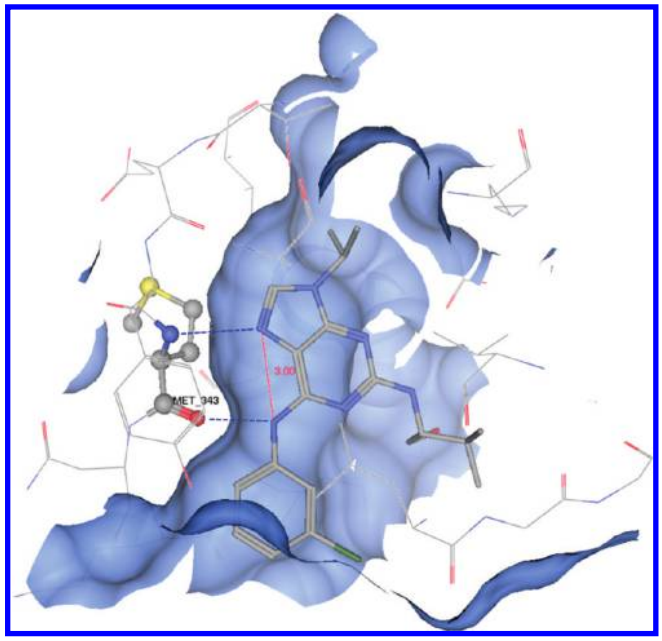


Figure 6. The SRC crystal structure of 1yom. The ligand forms two hydrogen bonds with Met343. The distance between the ligand donor and acceptor is 3.00 Å.

SRC. The most important interaction pair of SRC inhibitors identified by RF was two hydrogen bonds with 3 Å distance (Table 3). Among the nine crystal structures, seven structures formed two hydrogen bonds between the ligand atoms in or next to the ring conformation and Glu341 or Met343 of SRC (Figure 6). The distances between the two ligand atoms forming these hydrogen bonds ranged from 2.28 to 3.15 Å. The RF model recognized these two hydrogen bonds as the most important feature for SRC inhibitors. The interacting residues were located at the hinge region of SRC. In general, hydrogen bonds between the ligands and the hinge region are considered as the common important interactions of the kinase−inhibitor complexes. The RF model of SRC successfully identified the important feature of the active compounds of SRC.

Carbonic Anhydrase II. The RF model for carbonic anhydrase II stressed the importance of the interaction pairs of hydrogen bonds within 2−3 Å (Table 4). Most of the carbonic anhydrase II complex structures formed the hydrogen bonds at the deepest position of the ligand binding site (Figure 7). At this location, the sulfonamide group of each ligand was commonly positioned. The pharmacophore atoms

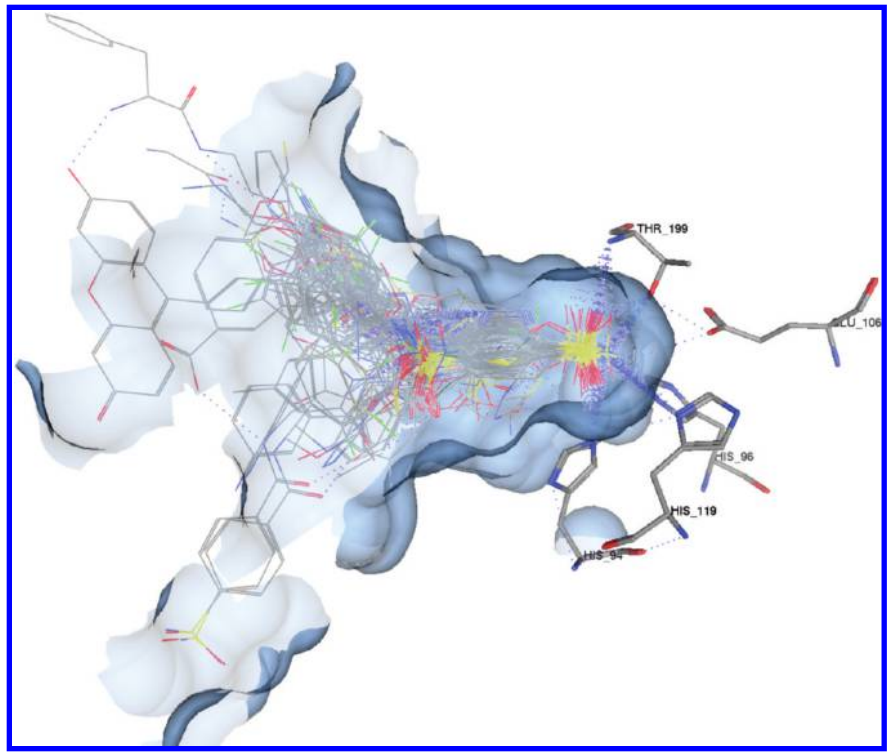


Figure 7. The binding site of carbonic anhydrase II. Most of the ligands positioned their sulfonamide groups at the deepest location in the binding site.

Table 5. The Top 10 Important Interaction Pairs of the RF Model for Cathepsin K

pair of interacting ligand atoms		distance (Å)	importance (Gini)
Hbond (L = "D")	Hbond (L = "A")	3	2.03
Hbond (L = "D")	Hbond (L = "A")	2	1.48
Hbond (L = "D")	Hbond (L = "A")	4	1.45
hydrophobic	hydrophobic	8	0.88
Hbond (L = "D")	Hbond (L = "A")	4	0.87
hydrophobic	hydrophobic	4	0.83
hydrophobic	hydrophobic	9	0.83
Hbond (L = "D")	Hbond (L = "A")	5	0.80
hydrophobic	hydrophobic	6	0.78
hydrophobic	hydrophobic	5	0.75

of the sulfonamide group were 2−3 Å apart. The RF model successfully identified the pharmacophore atom pairs in the sulfonamide group. In some cases, one pharmacophore atom in the sulfonamide formed two hydrogen bonds with the protein simultaneously. The third most important interaction pair of the RF model corresponds to the pharmacophore atoms that form two hydrogen bonds.

Cathepsin K. The interaction pairs of hydrogen-bonding donors and acceptors with a 2−4 Å distance were regarded as the important features of cathepsin K (Table 5). The binding site of cathepsin K is relatively shallow and contains several hydrogen-bonding sites, which recognize the peptide sequences for the proteolysis of its substrates. In the crystal structures of cathepsin K−ligand complexes, 4−5 hydrogen bonds were formed, and the hydrogen-bonding ligand atoms were periodically located at approximately every 2−4 Å. The hydrogen-bonding pharmacophores that were common to more than 33% of the superposed 25 cathepsin K crystal structures were extracted using MOE and are shown in Figure 8. The distances between the neighboring pharmacophores ranged from 2.24 to 4.77 Å. The RF model of cathepsin K recognized the compounds that could effectively use these hydrogen-bonding sites as the active compounds.

HIV-1 Protease. For HIV-1 protease, the pairs of hydrophobic interactions appeared to be the most important features (Table 6). Especially, the distances between the atoms of the top five interaction pairs ranged from 7 to 11 Å, which were the longest distances among the important features detected for all five target proteins in this study. The binding site of HIV-1 protease is covered by two β-strands and forms a tunnel-shaped structure. The distances of the interaction pairs almost matched the entire length of this binding site. Hydrophobic residues such as Gly27, Ala28, Val32, Gly48, Gly49, Ile50, and Leu123 are exposed in the binding site. To validate the long distance interaction pairs, we analyzed the common hydrophobic pharmacophore features in 197 crystal structures of HIV-1 protease. Consequently, we extracted five hydrophobic pharmacophore features (Figure 9) that were located 5.24−11.12 Å apart from each other. These results suggested that the RF model of HIV-1 protease recognizes the compounds that bind to HIV-1 protease by covering the full length of the tunnel-shaped binding site as the active compounds.

Compounds that GLIDE Failed To Detect but Pharm-IF-SVM Saved. The distributions of the molecular weights of all of the active compounds, the active compounds detected in the top 10% of the GLIDE scores, and the active compounds detected in the top 10% of the Pharm-IF-SVM models are shown in Figure 10. Among all of the targets, the combination of Pharm-IF and SVM outperformed GLIDE score in the entire range of molecular weights (43.02−856.8). Especially, the Pharm-IF-SVM models successfully detected the low molecular weight compounds (MW ≤ 400), as compared to GLIDE score. In carbonic anhydrase II, the molecular weights of most of the active compounds are less

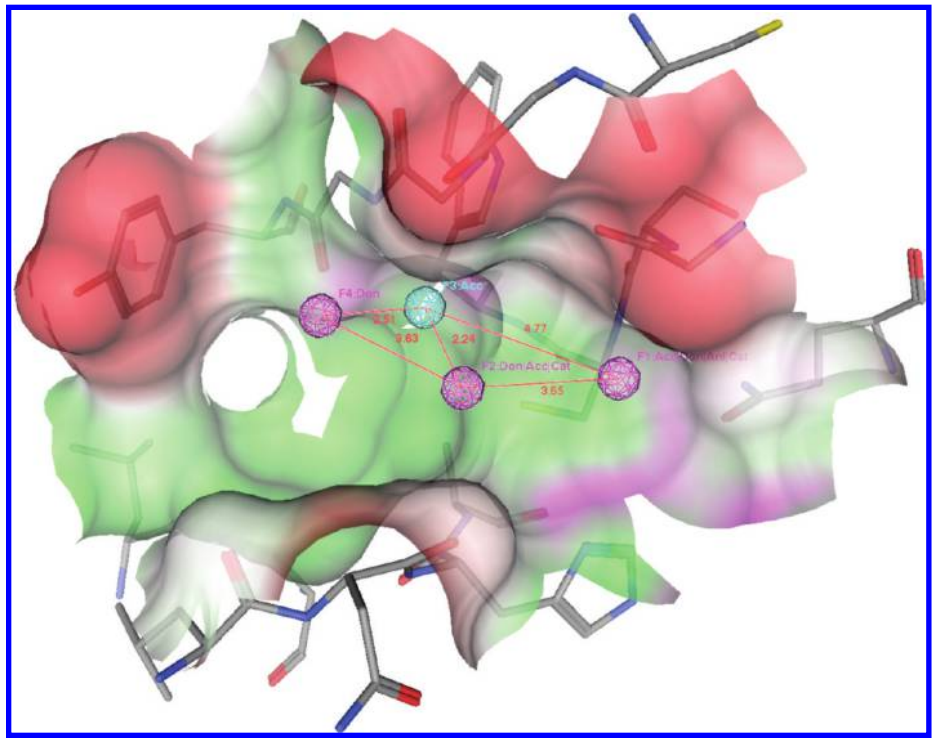


Figure 8. The binding site of cathepsin K. The pharmacophore features of hydrogen-bonding donors and acceptors conserved in more than 33% of the ligands in the crystal structures are shown as meshed balls. The color of the protein surface corresponds to polar (pink), hydrophobic (green) in the binding cavity, and surface exposed to water (red).

Table 6. The Top 10 Important Interaction Pairs of the RF Model for HIV-1 Protease

pair of interacting ligand atoms		distance (Å)	importance (Gini)
hydrophobic	hydrophobic	8	12.09
hydrophobic	hydrophobic	9	11.57
hydrophobic	hydrophobic	10	11.46
hydrophobic	hydrophobic	11	11.29
hydrophobic	hydrophobic	7	10.87
hydrophobic	hydrophobic	2	9.16
hydrophobic	hydrophobic	0	8.77
hydrophobic	hydrophobic	5	8.28
hydrophobic	hydrophobic	3	8.21
hydrophobic	hydrophobic	1	8.06

than 400, resulting in the large improvement of the 10% EF of carbonic anhydrase II screening by the use of the SVM model based on Pharm-IF. The docking poses of the active compounds of carbonic anhydrase II and PKA detected by the Pharm-IF-SVM models were further analyzed to clarify the differences between GLIDE score and the Pharm-IF-SVM models.

PKA. From the analysis, the active compounds were rescued by the Pharm-IF-SVM models for various reasons depending on the molecular weight or the interaction pattern. The smallest PKA inhibitor that GLIDE failed to detect was compound **1**.[56] The docking pose of **1** (Figure 11a) formed an ionic attraction with Glu127 at the entrance of the binding site and a hydrogen bond with Glu121 at the deepest area of the binding site. This binding mode seemed to be reasonable despite the lack of the crystal structure. While the Pharm-IF-SVM model rated the presence of the important ionic attraction and ranked compound **1** as 10th among the 2100 compounds, GLIDE score underestimated the docking pose and ranked it as 1131st. Small compounds that form important interactions and efficiently bind to the target proteins are regarded as promising leads[57] rather than large compounds. However, conventional scoring functions often underestimate such compounds, supposedly because larger compounds can form more protein−ligand interactions than smaller compounds.[58] The machine learning of Pharm-IF would be able to recognize the importance of the ionic attraction with Glu127 and the hydrogen bond with Glu121. The use of these common interaction patterns identified by machine learning could have resulted in the detection of the small active compound **1**.

The majority of the large PKA inhibitors that GLIDE score failed to detect were staurosporine derivatives. Comparing the crystal structure of PKA and staurosporine (1stc) with the docking poses of these compounds, the docking poses were basically predicted to be incorrect, and the incorrect docking poses caused their low GLIDE score ranking. The PKA structure used in this study (1re8) has smaller binding site as compared to 1stc, because Phe327 points toward the inside of the binding site. The difference in the binding site resulted in the van der Waals clash with staurosporine at the position of 1stc and the failures of the GLIDE docking of the staurosporine derivatives. On the other hand, these docking poses formed identical ionic attractions around the entrance of the binding site, which were considered as the important features of the active compounds. The hydrophobic interactions around the ionic attractions were also formed despite the differences in the participating protein atoms. The Pharm-IF-SVM model rescued the large staurosporine derivatives using the incorrect binding mode. This behavior of the model would not be desired from the viewpoint of structure-based drug design. By docking the staurosporine derivatives with 1stc, their proper complex structures were

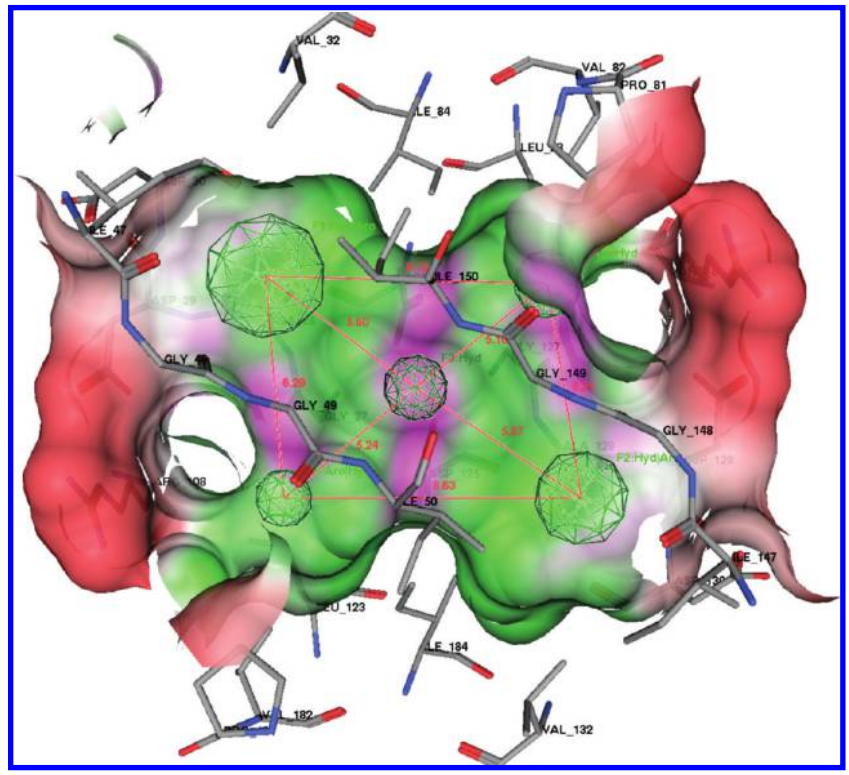


Figure 9. The binding site of HIV-1 protease. The hydrophobic pharmacophore features conserved in more than 33% of the ligands in the crystal structures are shown as green meshed balls. The color of the protein surface corresponds to polar (pink), hydrophobic (green) in the binding cavity, and surface exposed to water (red).

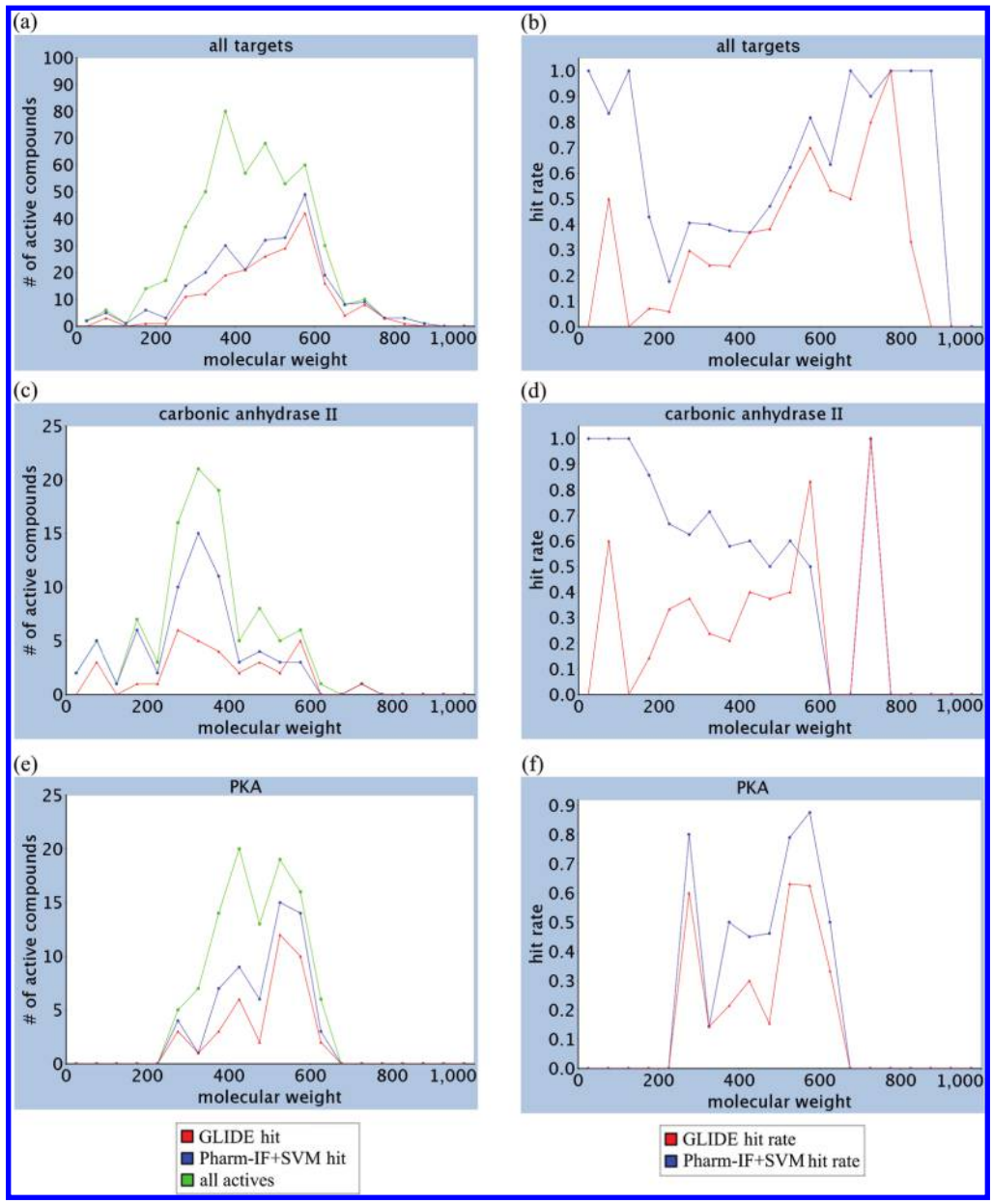


Figure 10. The distributions of the molecular weights of the active compounds. Red lines represent the active compounds detected by GLIDE score within the top 10%. Blue lines represent the active compounds detected by the Pharm-IF-SVM models within the top 10%. Green lines represent all active compounds. (a) The distribution of the molecular weights of the active compounds of all target proteins. (b) The recovery rates of all of the active compounds, using the two methods for each molecular weight bin. (c) The distribution of the active compounds of PKA. (d) The recovery rates of the active compounds of PKA for each molecular weight bin. (e) The distribution of the active compounds of carbonic anhydrase II. (f) The recovery rates of the active compounds of carbonic anhydrase II for each molecular weight bin.

successfully predicted using GLIDE, and the 10% EF of the Pharm-IF-SVM model slightly improved from 6.0 to 6.3. Therefore, accurate and exhaustive docking pose generation is critically important for the interaction pattern analysis of docking poses.

Carbonic Anhydrase II. Most of the active compounds of carbonic anhydrase II that the Pharm-IF-SVM model successfully rescued from the failures by GLIDE score had sulfonamide groups at the deepest area of the binding sites. The carbonic anhydrase II inhibitors are small and formed a limited number of interactions including hydrogen bonds by the sulfonamide groups, while the large decoy compounds could form many interactions. The Pharm-IF-SVM model detected the small active compounds by stressing the presence of the common interactions, such as the hydrogen bonds by the sulfonamide groups. For example, compound **2**,[59] with a molecular weight of 250.27, was ranked 1304th out of the 2100 compounds by GLIDE score and 27th by the Pharm-IF-SVM model. The docking pose of **2** (Figure 11b) formed four hydrogen bonds between the sulfonamide group and His94, His96, and Thr199, and two additional hydrogen bonds with Gln92 and Thr200. The RF model learned the hydrogen-bond pairs that are 3, 2, or 0 Å apart from each other as the three most important features of the carbonic anhydrase II complexes. The subsequently important hydrogen-bond pair in the RF model, ranked as 16th by ΔGini, was a pair of hydrogen donor and acceptor with a 7 Å distance. The distances between the ligand atoms of the additional hydrogen bonds and the sulfonamide group of **2** ranged from 6.47 to 6.64 Å, corresponding to the above type

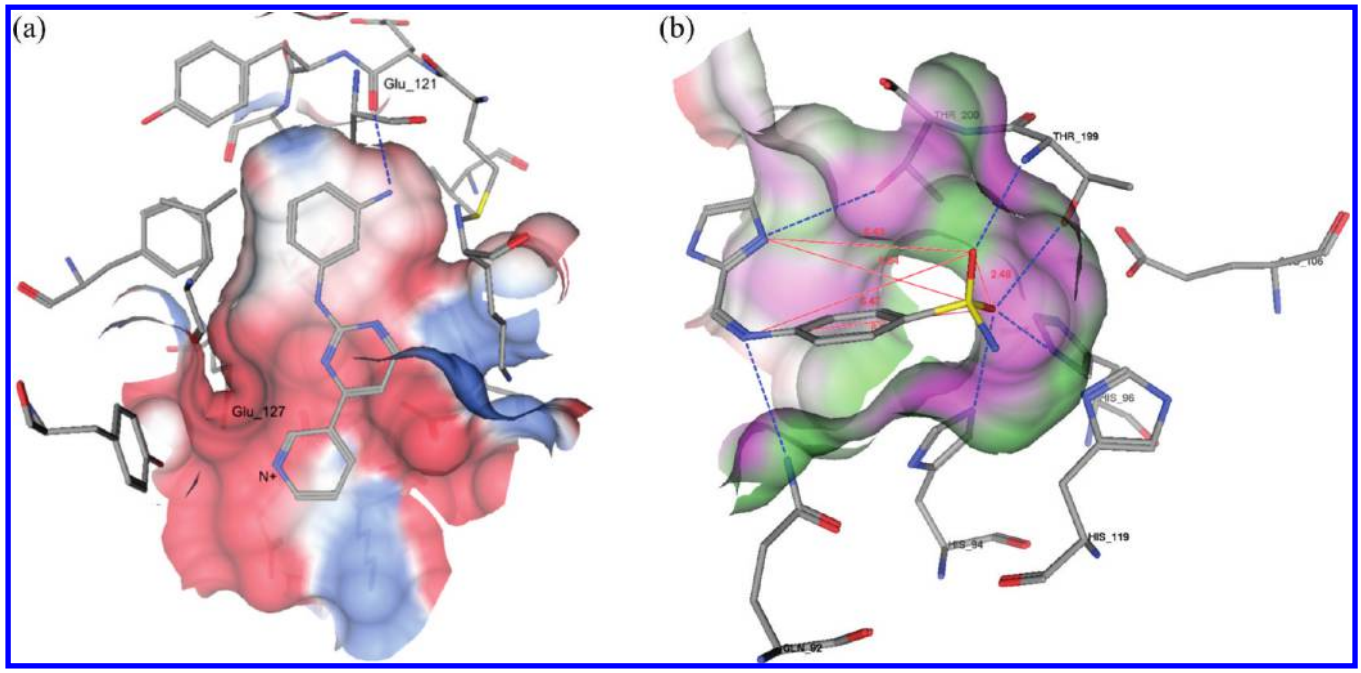


Figure 11. (a) The docking pose of **1** and PKA. The colors represent the electrostatics of the PKA surface. (b) The docking pose of **2** and carbonic anhydrase II, forming six hydrogen bonds.

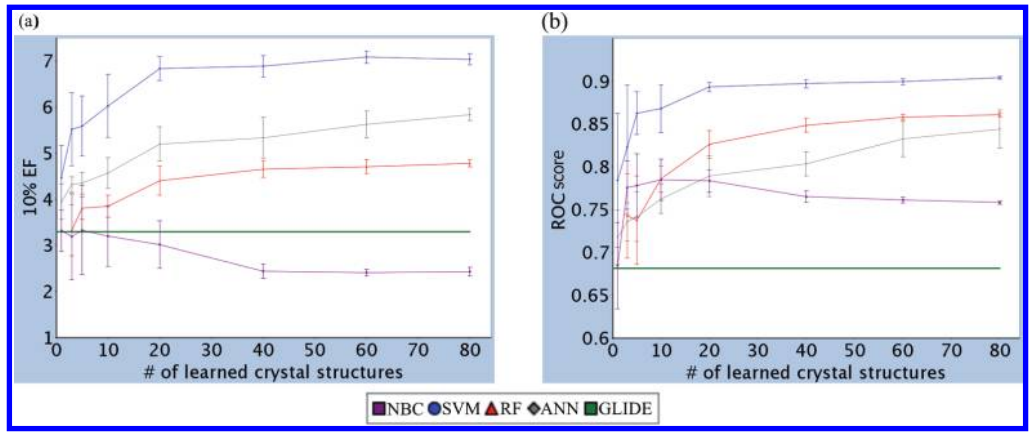


Figure 12. The mean values of (a) 10% EF and (b) ROC score of the learning models for carbonic anhydrase II, using various numbers of crystal structures as positive samples of the training set. The horizontal axis corresponds to the number of crystal structures used for the learning. The bold green lines represent the results of GLIDE score. The blue lines represent the SVM models. The gray lines represent the ANN models. The red lines represent the RF models. The purple lines represent the NBC models. The error bars indicate the standard deviations of 10 trials.

of interaction pair. The docking pose of **2** well matched the geometry of the hydrogen bonds learnt by machine learning and Pharm-IF.

The Effects of Training Set Size on the Performances of the Machine Learning Models. Figure 12 shows the relationship of the screening efficiencies (10% EF and ROC score) and the numbers of positive samples in the training sets for the model buildings for carbonic anhydrase II.

By learning more than five crystal structures, the SVM, ANN, and RF methods stably created learning models that outperformed GLIDE score. These results suggested that the machine learning models based on IFs are useful, even when there are only a few crystal structures available for the target proteins. In particular, the learning effects of the SVM method were excellent with any size of the training set. The screening efficiency of the SVM models reached a plateau after learning approximately 20 crystal structures. On the other hand, the performances of the ANN models improved comparatively slowly, while the ROC score of the ANN models continued to increase, even when large numbers of crystal structures were learnt. The performance of the ANN models strongly depended on the first choice of the learning samples, while the SVM and RF models were more robust. Interestingly, the efficiencies of the NBC models sharply decreased after learning more than 20 crystal structures. Most of the carbonic anhydrase II inhibitors shared common positions of the sulfonamide groups; however, their structures at the entrance of the binding site widely varied. In this case, the prediction models have to learn the many patterns of the active compounds. Because NBC treats all variables independently and builds probabilistic models, and cannot learn the complicated relationship among the variables, it would be difficult to learn the various interaction patterns of the active compounds.

Machine Learning Using Docking Poses of Active Compounds. The screening efficiencies of the learning models for SRC and cathepsin K, built using both the experimentally determined structures and the docking poses

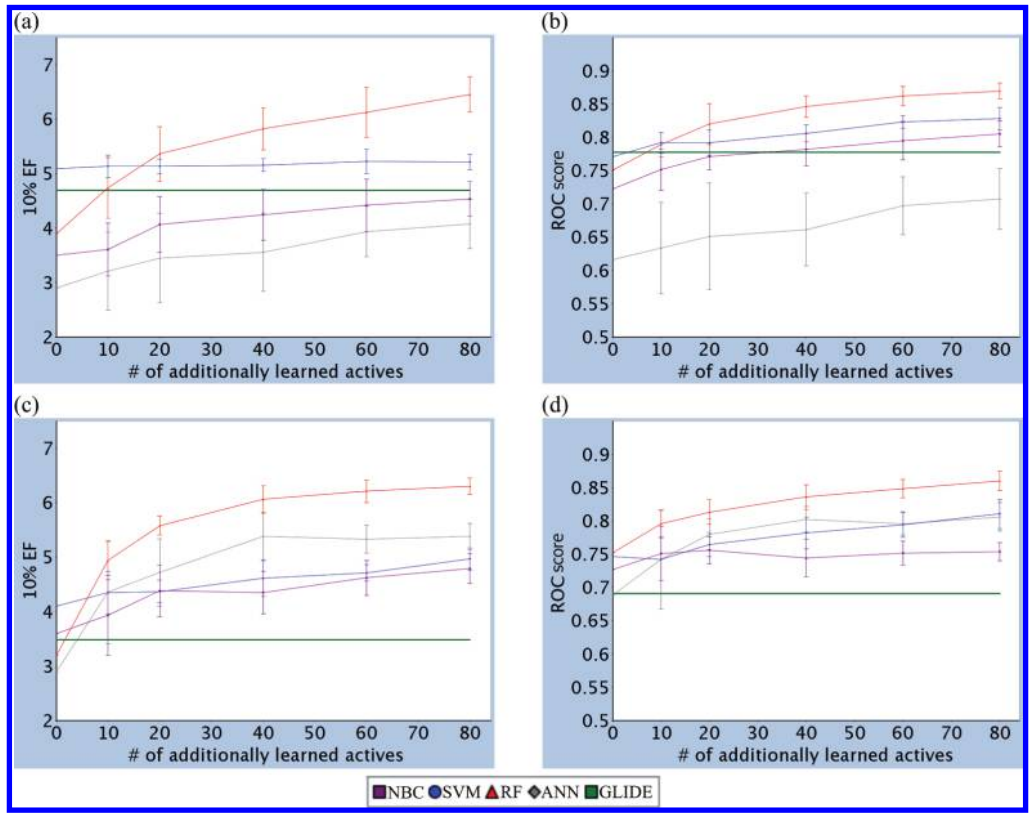


Figure 13. The screening efficiencies of the learning models using the docking poses of the active compounds of SRC (a and b) and cathepsin K (c and d). Panels (a) and (c) show the 10% EF, and (b) and (d) show the ROC scores. The horizontal axes correspond to the number of active compounds used for the learning in addition to the complex structures.

of active compounds as positive samples, are summarized in Figure 13. Using the docking poses, the screening efficiencies of the learning models by RF for both targets surprisingly improved from 3.9 and 3.2 to 6.5 and 6.3 (10% EF), respectively. Although the SVM models recorded the best screening efficiencies when only the experimentally determined crystal structures were learnt, the efficiencies were not improved much by using the docking poses of the active compounds. On the other hand, the performances of the RF models were significantly improved by the additional use of the docking poses and achieved the best efficiencies among all of the machine learning algorithms. The docking poses of the active compounds are supposed to include the incorrect binding modes, as compared to the experimentally determined complex structures. The significant improvement of the screening efficiencies of the RF models supposedly resulted from the robustness of the RF algorithm against noise-containing data.

CONCLUSION

In this study, target-specific discrimination models combining description methods for protein−ligand interactions and machine learning for in silico screening were explored. In terms of PLIF, a residue-based IF, machine learning algorithms, and conventional similarity-based ranking were compared by the screening efficiency, such as 10% EF and ROC score. On average, for five targets, the machine learning models based on PLIF showed higher efficiencies than those of the similarity-based ranking. Especially, the PLIF-SVM models recorded higher 10% EF values and ROC scores than those of the similarity-based ranking for all of the targets. We developed a new atom-based IF (Pharm-IF) to describe the patterns of ligand pharmacophores that interacted with proteins in complex structures and examined its performance. The machine learning models based on Pharm-IF outperformed both GLIDE score and the machine learning models based on PLIF for all five targets. Among the four machine learning algorithms, SVM, RF, ANN, and NBC, SVM recorded the best results for both the Pharm-IF and PLIF models. To validate the machine learning models, the importance of each interaction pair of Pharm-IF was analyzed. As a result, the important interaction pairs were consistent with the common and well-known interactions of each target, suggesting that the combination of machine learning algorithms and Pharm-IF adequately learned the differences among the contributions of various protein−ligand interactions to the binding affinity. Thus, the analysis of important interaction pairs facilitates an understanding of the structure−activity relationship.

The effect of the number of positive samples in the training set on the performance of each model was investigated for all of the algorithms. The machine learning models by SVM, ANN, and RF could outperform GLIDE score when more than five crystal structures were used for model building. Among the four algorithms, SVM showed the maximum learning effects to improve the 10% EF values and the ROC scores. Approximately 20 crystal structures were needed for SVM to create the learning models with the peak performance. Although ANN could create moderate classification models, large standard deviations of the 10% EFs and ROC scores were observed, which suggested that the performance of the ANN models largely depended on the choice of the learned complexes. The efficiency of the NBC models

significantly decreased when various patterns of complex structures were input.

To learn more diverse interaction patterns between a protein and its ligands, we used the docking poses of known active compounds in addition to the crystal structures. The screening efficiencies of the learning models for SRC and cathepsin K, which have relatively few crystal structures available, were improved by adding the docking poses of their active compounds. In particular, the 10% EF values of the RF models were dramatically enhanced to 6.5 (SRC) and 6.3 (cathepsin K), as compared to those of the models using only the crystal structures (3.9 (SRC) and 3.2 (cathepsin K)). In contrast, the SVM models did not show significant learning effects in this case. Because the docking poses are supposed to include incorrect binding modes, SVM would be sensitive to the wrong data. The RF method produces an ensemble of training sets and descriptor sets and is known as a statistical method to build robust learning models against data containing noise. The ensemble methodology of RF is suitable to build models using training sets including noise. Applying machine learning algorithms based on Pharm-IF to in silico screening, SVM would be the choice for learning reliable structures, and RF would be the best algorithm when docking poses are also used as the positive samples.

As a further development, Pharm-IF is expected to be a useful tool for the analysis and classification of proteins according to the interaction patterns with their ligands. The learning models based on Pharm-IF reflected the geometry of the structural features created by the binding sites. As compared to residue-based IFs, such as PLIF and SIFt, Pharm-IF is independent of the protein sequences, and it allows a comparison of the interaction patterns of a target with those of other targets. Using the features of Pharm-IF, applications of machine learning models of a target to in silico screening of other homologous proteins and classifications of proteins from the viewpoint of protein−ligand interactions can be performed. The classification based on protein−ligand interactions can provide a new aspect of protein classification, in addition to the classifications based on amino acid sequences and binding site properties. It would facilitate the identification of inhibitors that simultaneously bind to multiple desired target proteins and the elimination of inhibitors that interact with undesirable targets, causing adverse effects.

Abbreviations. IF, interaction fingerprint; NBC, naïve Bayesian classifier; SVM, support vector machine; RBF, radial basis function; ANN, artificial neural network; RF, random forest; EF, enrichment factor; ROC, receiver operating characteristics.

ACKNOWLEDGMENT

We thank Hitomi Yuki and Akiko Tanaka for discussions and manuscript preparation, and Toshio Furuya and Naoko Inoue, of PharmaDesign, Inc., for providing the StARLITe data.

Supporting Information Available: The enrichment plot of test using GLIDE, Pharm-IF-SVM models, PLIF-SVM models, and PLIF-max_tc for PKA, SRC, carbonic anhydrase II, cathepsin K, and HIV-1 protease. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Ewing, T. J.; 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.
(2) Goodsell, D. S.; Lauble, H.; Stout, C. D.; Olson, A. J. Automated docking in crystallography: analysis of the substrates of aconitase. *Proteins* **1993**, *17*, 1–10.
(3) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004**, *47*, 1739–1749.
(4) Wang, R.; Lu, Y.; Wang, S. Comparative evaluation of 11 scoring functions for molecular docking. *J. Med. Chem.* **2003**, *46*, 2287–2303.
(5) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discovery* **2004**, *3*, 935–949.
(6) Krovat, E. M.; Langer, T. Impact of scoring functions on enrichment in docking-based virtual screening: an application study on renin inhibitors. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1123–1129.
(7) Perola, E.; Walters, W. P.; Charifson, P. S. A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. *Proteins* **2004**, *56*, 235–249.
(8) Leach, A. R.; Shoichet, B. K.; Peishoff, C. E. Prediction of protein−ligand interactions. Docking and scoring: successes and gaps. *J. Med. Chem.* **2006**, *49*, 5851–5855.
(9) Sousa, S. F.; Fernandes, P. A.; Ramos, M. J. Protein−ligand docking: current status and future challenges. *Proteins* **2006**, *65*, 15–26.
(10) Warren, G. L.; Andrews, C. W.; Capelli, A. M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A critical assessment of docking programs and scoring functions. *J. Med. Chem.* **2006**, *49*, 5912–5931.
(11) Moitessier, N.; Englebienne, P.; Lee, D.; Lawandi, J.; Corbeil, C. R. Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go. *Br. J. Pharmacol.* **2008**, *153*, S7–26.
(12) Sharff, A.; Jhoti, H. High-throughput crystallography to enhance drug discovery. *Curr. Opin. Chem. Biol.* **2003**, *7*, 340–345.
(13) Hajduk, P. J.; Gerfin, T.; Boehlen, J. M.; Haberli, M.; Marek, D.; Fesik, S. W. High-throughput nuclear magnetic resonance-based screening. *J. Med. Chem.* **1999**, *42*, 2315–2317.
(14) 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.
(15) Stiefl, N.; Zaliani, A. A knowledge-based weighting approach to ligand-based virtual screening. *J. Chem. Inf. Model.* **2006**, *46*, 587–596.
(16) Crisman, T. J.; Sisay, M. T.; Bajorath, J. Ligand-target interaction-based weighting of substructures for virtual screening. *J. Chem. Inf. Model.* **2008**, *48*, 1955–1964.
(17) Gohlke, H.; Klebe, G. DrugScore meets CoMFA: adaptation of fields for molecular comparison (AFMoC) or how to tailor knowledge-based pair-potentials to a particular protein. *J. Med. Chem.* **2002**, *45*, 4153–4170.
(18) Mooij, W. T.; Verdonk, M. L. General and targeted statistical potentials for protein-ligand interactions. *Proteins* **2005**, *61*, 272–287.
(19) Deng, Z.; Chuaqui, C.; Singh, J. Structural interaction fingerprint (SIFt): a novel method for analyzing three-dimensional protein-ligand binding interactions. *J. Med. Chem.* **2004**, *47*, 337–344.
(20) Kelly, M. D.; Mancera, R. L. Expanded interaction fingerprint method for analyzing ligand binding modes in docking and structure-based drug design. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1942–1951.
(21) Chuaqui, C.; Deng, Z.; Singh, J. Interaction profiles of protein kinase-inhibitor complexes and their application to virtual screening. *J. Med. Chem.* **2005**, *48*, 121–133.
(22) Mpamhanga, C. P.; Chen, B.; McLay, I. M.; Willett, P. Knowledge-based interaction fingerprint scoring: a simple method for improving the effectiveness of fast scoring functions. *J. Chem. Inf. Model.* **2006**, *46*, 686–698.
(23) Deng, Z.; Chuaqui, C.; Singh, J. Knowledge-based design of target-focused libraries using protein-ligand interaction constraints. *J. Med. Chem.* **2006**, *49*, 490–500.
(24) Marcou, G.; Rognan, D. Optimizing fragment and scaffold docking by use of molecular interaction fingerprints. *J. Chem. Inf. Model.* **2007**, *47*, 195–207.
(25) Venhorst, J.; Nunez, S.; Terpstra, J. W.; Kruse, C. G. Assessment of scaffold hopping efficiency by use of molecular interaction fingerprints. *J. Med. Chem.* **2008**, *51*, 3222–3229.
(26) Kumar, A.; Siddiqi, M. I. Virtual screening against Mycobacterium tuberculosis dihydrofolate reductase: suggested workflow for compound prioritization using structure interaction fingerprints. *J. Mol. Graphics Modell.* **2008**, *27*, 476–488.

(27) Perez-Nueno, V. I.; Rabal, O.; Borrell, J. I.; Teixido, J. APIF: a new interaction fingerprint based on atom pairs and its application to virtual screening. *J. Chem. Inf. Model.* **2009**, *49*, 1245–1260.
(28) Nandigam, R. K.; Kim, S.; Singh, J.; Chuaqui, C. Position specific interaction dependent scoring technique for virtual screening based on weighted protein--ligand interaction fingerprint profiles. *J. Chem. Inf. Model.* **2009**, *49*, 1185–1192.
(29) *MOE* (Molecular Operating Environment), version 2007.09; Chemical Computing Group Inc.: Montreal, Quebec, Canada, 2009.
(30) Brewerton, S. C. The use of protein−ligand interaction fingerprints in docking. *Curr. Opin. Drug Discovery Dev.* **2008**, *11*, 356–364.
(31) Singh, J.; Deng, Z.; Narale, G.; Chuaqui, C. Structural interaction fingerprints: a new approach to organizing, mining, analyzing, and designing protein-small molecule complexes. *Chem. Biol. Drug Des.* **2006**, *67*, 5–12.
(32) Muller, K. R.; Ratsch, G.; Sonnenburg, S.; Mika, S.; Grimm, M.; Heinrich, N. Classifying 'drug-likeness' with kernel-based learning methods. *J. Chem. Inf. Model.* **2005**, *45*, 249–253.
(33) Schneider, G.; Wrede, P. Artificial neural networks for computer-based molecular design. *Prog. Biophys. Mol. Biol.* **1998**, *70*, 175–222.
(34) Agrafiotis, D. K.; Cedeno, W.; Lobanov, V. S. On the use of neural network ensembles in QSAR and QSPR. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 903–911.
(35) Kauffman, G. W.; Jurs, P. C. QSAR and k-nearest neighbor classification analysis of selective cyclooxygenase-2 inhibitors using topologically-based numerical descriptors. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1553–1560.
(36) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random forest: a classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947–1958.
(37) Byvatov, E.; Fechner, U.; Sadowski, J.; Schneider, G. Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1882–1889.
(38) Winkler, D. A. Neural networks as robust tools in drug lead discovery and development. *Mol. Biotechnol.* **2004**, *27*, 139–168.
(39) Guha, R.; Jurs, P. C. Interpreting computational neural network QSAR models: a measure of descriptor importance. *J. Chem. Inf. Model.* **2005**, *45*, 800–806.
(40) Plewczynski, D.; Spieser, S. A.; Koch, U. Assessing different classification methods for virtual screening. *J. Chem. Inf. Model.* **2006**, *46*, 1098–1106.
(41) Chen, B.; Harrison, R. F.; Papadatos, G.; Willett, P.; Wood, D. J.; Lewell, X. Q.; Greenidge, P.; Stiefl, N. Evaluation of machine-learning methods for ligand-based virtual screening. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 53–62.
(42) Ehrman, T. M.; Barlow, D. J.; Hylands, P. J. Virtual screening of Chinese herbs with Random Forest. *J. Chem. Inf. Model.* **2007**, *47*, 264–278.
(43) Sato, T.; Matsuo, Y.; Honma, T.; Yokoyama, S. In silico functional profiling of small molecules and its applications. *J. Med. Chem.* **2008**, *51*, 7705–7716.
(44) Sakiyama, Y.; Yuki, H.; Moriya, T.; Hattori, K.; Suzuki, M.; Shimada, K.; Honma, T. Predicting human liver microsomal stability with machine learning techniques. *J. Mol. Graphics Modell.* **2008**, *26*, 907–915.
(45) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32.
(46) Vapnik, V. *The Nature of Statistical Learning Theory*; Springer-Verlag: New York, 1995.
(47) Cheng, A. C.; Coleman, R. G.; Smyth, K. T.; Cao, Q.; Soulard, P.; Caffrey, D. R.; Salzberg, A. C.; Huang, E. S. Structure-based maximal affinity model predicts small-molecule druggability. *Nat. Biotechnol.* **2007**, *25*, 71–75.
(48) Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Bryant, S. H. PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.* **2009**, *37*, W623–W633.
(49) *StARLITe*; Inpharmatica Ltd.: London, UK, 2007.
(50) *Pipeline Pilot*; Accelrys Software Inc.: San Diego, CA, 2007.
(51) Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. Scaffold-hopping by topological pharmacophore search: A contribution to virtual screening. *Angew. Chem., Int. Ed.* **1999**, *38*, 2894–2896.
(52) Renner, S.; Noeske, T.; Parsons, C. G.; Schneider, P.; Weil, T.; Schneider, G. New allosteric modulators of metabotropic glutamate receptor 5 (mGluR5) found by ligand-based virtual screening. *ChemBioChem* **2005**, *6*, 620–625.
(53) Schuffenhauer, A.; Floersheim, P.; Acklin, P.; Jacoby, E. Similarity metrics for ligands reflecting the similarity of the target proteins. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 391–405.
(54) Minai, R.; Matsuo, Y.; Onuki, H.; Hirota, H. Method for comparing the structures of protein ligand-binding sites and application for predicting protein-drug interactions. *Proteins* **2008**, *72*, 367–381.
(55) Truchon, J. F.; Bayly, C. I. Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. *J. Chem. Inf. Model.* **2007**, *47*, 488–508.
(56) Zimmermann, J.; Buchdunger, E.; Mett, H.; Meyer, T.; Lydon, B. Potent and selective inhibitors of the Abl-kinase: phenylaminopyrimidine (PAP) derivatives. *Bioorg. Med. Chem. Lett.* **1997**, *7*, 187–192.
(57) Kuntz, I. D.; Chen, K.; Sharp, K. A.; Kollman, P. A. The maximal affinity of ligands. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 9997–10002.
(58) Pan, Y.; Huang, N.; Cho, S.; MacKerell, A. D., Jr. Consideration of molecular weight during compound selection in virtual target-based database screening. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 267–272.
(59) Vullo, D.; Franchi, M.; Gallori, E.; Antel, J.; Scozzafava, A.; Supuran, C. T. Carbonic anhydrase inhibitors. Inhibition of mitochondrial isozyme V with aromatic and heterocyclic sulfonamides. *J. Med. Chem.* **2004**, *47*, 1272–1279.

CI900382E