ACS OMEGA

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

Article

Public Data Set of Protein–Ligand Dissociation Kinetic Constants for Quantitative Structure–Kinetics Relationship Studies

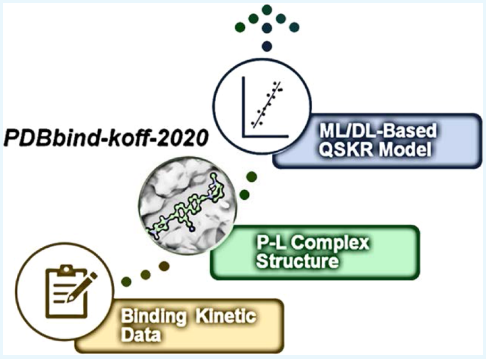
Huisi Liu, Minyi Su, Hai-Xia Lin,* Renxiao Wang,* and Yan Li*

Cite This: *ACS Omega* 2022, 7, 18985–18996

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Protein–ligand binding affinity reflects the equilibrium thermodynamics of the protein–ligand binding process. Binding/unbinding kinetics is the other side of the coin. Computational models for interpreting the quantitative structure–kinetics relationship (QSKR) aim at predicting protein–ligand binding/unbinding kinetics based on protein structure, ligand structure, or their complex structure, which in principle can provide a more rational basis for structure-based drug design. Thus far, most of the public data sets used for deriving such QSKR models are rather limited in sample size and structural diversity. To tackle this problem, we have compiled a set of 680 protein–ligand complexes with experimental dissociation rate constants (k_{off}), which were mainly curated from the references accumulated for updating our PDBbind database. Three-dimensional structure of each protein–ligand complex in this data set was either retrieved from the Protein Data Bank or carefully modeled based on a proper template. The entire data set covers 155 types of protein, with their dissociation kinetic constants (k_{off}) spanning nearly 10 orders of magnitude. To the best of our knowledge, this data set is the largest of its kind reported publicly. Utilizing this data set, we derived a random forest (RF) model based on protein–ligand atom pair descriptors for predicting k_{off} values. We also demonstrated that utilizing modeled structures as additional training samples will benefit the model performance. The RF model with mixed structures can serve as a baseline for testifying other more sophisticated QSKR models. The whole data set, namely, *PDBbind-koff-2020*, is available for free download at our PDBbind-CN web site (http://www.pdbbind.org.cn/download.php).



1. INTRODUCTION

Optimizing the affinity and selectivity of the candidate drug to the target is still the mainstream strategy in current drug discovery and development programs. However, the success rate of final approval of the drug is still disappointingly low despite the tremendous progress made in both experimental and computational techniques. Some studies suggest that in vivo activity of a drug is strongly correlated with the binding/unbinding kinetic properties rather than the equilibrium thermodynamic properties in the drug-target binding process.[1−4] For example, Guo et al. studied the relationships between the intracellular efficacy and the binding affinity, the residence time of 10 adenosine A2A receptor agonists.[2] It was found that the intracellular efficacy of these agonists correlated much better with the residence time ($R^2 = 0.90$) than the inhibition constant K_i did ($R^2 = 0.13$). Doornbos et al. reported a series of 7-aryl-1,2,4-triazolo[4,3-*a*]pyridines with mGlu2 positive allosteric modulator activity and affinity.[4] Evaluations of the compounds with short, medium, and long residence time in a label-free assay demonstrated a correlation between the residence time and their pharmaceutical effect. They also pointed out that affinity-only driven selection would lead to compounds with high k_{on} values but not necessarily optimized residence time, which is the key indicator for predicting optimal efficacy in vivo. All these results indicate that improving the protein–ligand binding/unbinding kinetics has become a new trend in lead optimization.[5,6]

Kinetic properties of protein–ligand binding (i.e., residence time and association and dissociation rate constants) can be measured by some experimental methods, such as surface plasmon resonance (SPR), radioligand binding, time-resolved fluorescence resonance energy transfer, and so on. However, all these experimental techniques are not only expensive and time-consuming but also insufficient to characterize the kinetics of protein–ligand interactions in detail. Therefore, it is necessary to develop computational methods for elucidating the protein–ligand binding/unbinding process. To answer this demand, molecular dynamics (MD) simulations have been widely applied to explore the dissociation pathway of the ligand from the protein in full atom detail.[7−10] For example, Platter and Noe combined an extensive set of 150 μs MD simulation and Markov model-based analysis to investigate the interplay of conforma-

Received: April 6, 2022
Accepted: May 13, 2022
Published: May 26, 2022

https://doi.org/10.1021/acsomega.2c02156

tional change and ligand-binding kinetics for the serine protease trypsin and its competitive inhibitor benzamidine.[9] Seven metastable conformations with different binding pocket structures were found to interconvert at timescales of tens of microseconds, which all had corresponding crystal structures released in PDB. Miao et al. developed a new computational method, ligand Gaussian accelerated MD, which provides a powerful enhanced sampling approach for characterizing ligand binding thermodynamics and kinetics simultaneously.[10] It was applied to the protein−ligand binding model system, that is, trypsin complexed with a benzamidine inhibitor. The calculated ligand binding free energy and kinetic rate constants were well consistent with the experimental data. Obviously, MD simulation is gradually playing a pivotal role in investigating the protein−ligand binding kinetics. However, MD-based methods usually require a lot of computational resources and thus are not suitable for high-throughput applications.

In recent years, with the increasing availability of protein binding kinetics data, data-driven modeling provides an alternative and efficient solution for studying the protein−ligand interaction kinetics. For example, the typical quantitative structure−kinetics relationship (QSKR) strategy has been adopted to build models for predicting protein−ligand dissociation kinetic constants and give guidance for structure-based designs.[11−13] Mei H. et al. adopted partial least squares and support vector machine to build a QSKR model on a training set composed of 37 HIV-1 protease inhibitors.[12] The predictive model showed that several descriptors derived from water and hydrophobic probes are dominant factors for the kinetic and thermodynamic properties. Wade et al. applied a similar protocol to the data set of 66 HSP and 33 HIV-1 protease inhibitors by using the comparative binding energy analysis.[13] Target-specific scoring functions were derived by correlating k_{off} with a subset of weighted interaction energy components, which were promising for binding kinetics-guided lead optimization.[13]

Obviously, a data set of protein−ligand complexes with reliable kinetics data and three-dimensional (3D) structures is a prerequisite for the development of QSKR models. However, current public collections of protein−ligand kinetic data are rather rare. Some well-known public databases (such as ChEMBL,[14,15] Binding MOAD,[16] etc.) that collect the protein−ligand binding constants do not include kinetic-related data. This is why the published studies have focused on data sets with very few targets. Even though there are some databases that contain kinetic data (such as KDBI,[17,18] KOFFI,[19] BindingDB,[20] etc.), they are either small in size or lack the 3D structures of the protein−ligand complexes. Recently, Fedorov et al. have described a set of 501 protein−ligand complexes with dissociation rate constants collected from the public literature.[21] This data set was the largest data set of this kind to the best of our knowledge. In addition, Fedorov et al. utilized their data set to develop a random forest (RF) model for predicting dissociation rate constants. Their work probably represents the state-of-the-art of QSKR models.

To overcome the limitations in the development of QSKR models, we have attempted to compile an even larger data set of protein−ligand binding kinetic data and the corresponding protein−ligand complex structures. Our data set covers more diverse target proteins and ligand molecules, which can serve as a benchmark for developing or validating computational QSKR models. For this purpose, we utilized the 38000+ references accumulated during our work of updating the PDBbind database[22,23] to collect protein−ligand dissociation rate constants (i.e., k_{off}) and relevant information (e.g., protein−ligand complex structure, experimental methods, etc.). In addition, we integrated two data sets described in the public literature[12,13] into our data set. Finally, our data set contained experimental k_{off} values for a total of 680 protein−ligand complexes. These complexes were formed by 155 types of protein and a wide range of ligand molecules. The 3D structure of each protein−ligand complex in our data set was either retrieved from the Protein Data Bank (PDB)[24] or carefully modeled based on a proper template, making our data set more suitable for QSKR studies. Based on our data set, we also provided an RF model for predicting k_{off} values, which is in concept similar to Fedorov's model.[21] Our model may serve as a baseline model for evaluating other QSKR models. The whole data set, namely, *PDBbind-koff-2020*, is freely available from our PDBbind-CN web site (http://www.pdbbind.org.cn/download.php).

2. METHODS

2.1. Collection of Protein−Ligand Complexes with Kinetic Constants.

The protein−ligand binding kinetic data in our data set were mainly collected from the references that we have processed for the regular update of the PDBbind database.[22,23] As on January 2020, we had accumulated a total of over 38,000 references for this purpose. All those references were first filtered by a few keywords (e.g., k_{on}, k_{off}, off-rate, and dissociation rate). This step was conducted with a computer program and resulted in ∼3400 references. Then, those references were manually read to curate k_{off} values of protein−ligand complex, the applied experimental methods and settings, and other relevant information. During this process, kinetic data of covalently bound protein−ligand complexes were ignored. If the ligand molecule contained unusual elements (such as B, Be, etc.), the data of this complex was not considered, either. For the collected complexes, their corresponding target information (i.e., PDB code and UniProt ID) was also recorded. After preliminary collection, we found that the kinetic data collected were redundant in several situations. The first case was that for the same protein−ligand complex, multiple sets of kinetic data and complex structures were obtained from different experimental methods or conditions. In this case, we gave priority to retaining the kinetic data and its corresponding structure measured by the SPR method under normal conditions (such as room temperature). The second case was that multiple sets of kinetic data for a protein−ligand complex were also determined from different research groups, but not each one corresponded to a complex structure. Thus, the complex with a crystal structure and its kinetic data should be retained first. If there were multiple crystal structures to choose from, the priority was given to retaining the one with better resolution. The last case was that the proteins in the complex structures had mutations or were from different species. In this case, we checked whether the residues around the binding site were significantly different. If so, they were considered as different complex structures and their kinetic data were retained as unique. Otherwise, they were considered as the same complexes, which would be handled as previous two cases.

In addition, we also solicited two data sets of protein−ligand dissociation kinetic constants from the literature, one containing 39 HIV-1 protease complexes[12] and the other containing 89 HSP90 complexes.[13] Finally, a total of 680 records of k_{off} with their target and ligand information were included in our data set. The complete information of the data set is provided in the

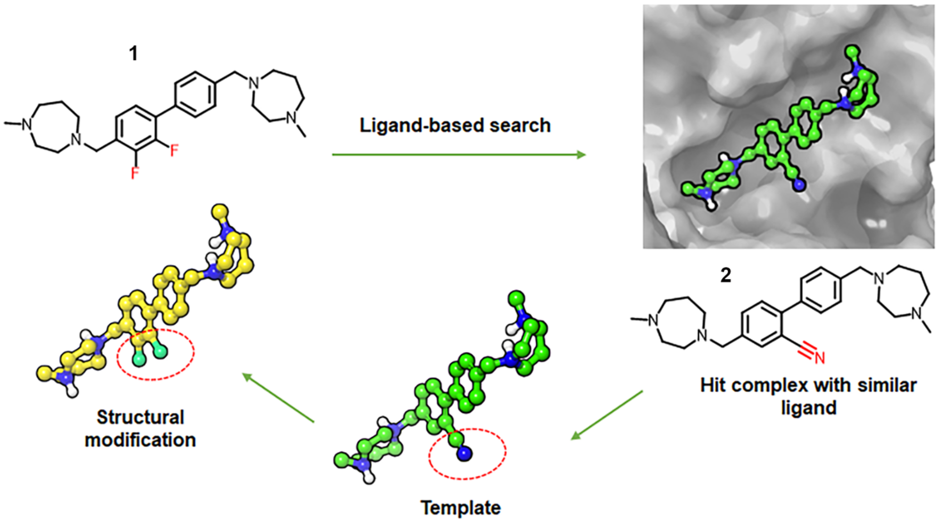


Figure 1. Illustration of the process of constructing a 3D model of a protein−ligand complex without an available crystal structure. The template ligand was selected from the crystal structure in PDB entry 5EQY as an example.

Supporting Information (Table S1). To the best of our knowledge, this data set is the largest collection of protein−ligand dissociation kinetic constants supplied with complex structures, which is publicly reported in the literature.

2.2. Process of Protein−Ligand Complex Structures. We required each protein−ligand complex included in our data set to have a 3D structure. This is obviously an important feature for our data set to become a popular benchmark. In our data set, some samples had known crystal complex structures, while others did not. For those samples with known crystal complex structures, we downloaded the coordinates from the PDB (http://www.rcsb.org/pdb).[24] The original structural files from the PDB were processed so that they could be readily applied to the software for computational tasks. The biological unit of each complex was split into a protein molecule and a ligand molecule. They were prepared using the Protein Preparation Wizard in Schrödinger software (Schrödinger LLC, version 2019). Water molecules and metal ions were removed from the protein structure. In some cases, the coordinates of some residues or heavy atoms were missing from the protein structure. Thus, the "Advanced Homology Modeling" module in Schrödinger was used to deal with this problem. The ligand molecule was also visually examined and corrected. The ProToss tool (https://protein.plus/) developed by Rarey's group[25] was used to add hydrogen atoms to the protein−ligand complex and predict the protonation states. The processed protein molecule and ligand molecule were saved in a PDB file format and a MOL2 file format, respectively.

For those samples without known crystal complex structures, we first searched the PDB to find out whether there were crystal complexes with the same protein and similar ligands. We then constructed the 3D structure of the unknown ligand based on the maximum common substructure (MCS) of the matched ligand. Take compound **1** as an example (Figure 1). Its congeneric compound **2** was found to bind with the target protein choline kinase (PDB code: 5EQY). The structure of compound **2** was extracted directly from the crystal structure and defined as the template. The 3D structure of compound **1** could be obtained by simply replacing the cyano and its adjacent hydrogen atom with two fluorine atoms, keeping the coordinates of the MCS unchanged. Assuming no apparent violations in the binding pocket, we believe that ligands derived based on the same molecular framework should adopt a similar binding pose. Therefore, it is reasonable to some extent to build a 3D ligand based on a common molecular framework. This strategy was also adopted in the previous work.[26,27]

2.3. Modeling Protein−Ligand Complex Structures. In our work, the GLIDE module with the GlideScore-SP scoring function[28,29] implemented in the Schrödinger software was adopted to predict the binding poses of the complexes with unknown crystal structures. Based on the assumption that ligands with similar chemical structures may adopt similar binding modes to the target protein, we visually inspected the docking output results and selected the reasonable binding pose. Here, we defined the binding pose that was most similar to the template ligand as the "best-fit" binding pose. If the "best-fit" binding pose was found, it was selected as the possible reasonable binding pose. However, for a few ligands that could not get a reference to determine the "best-fit" binding pose, we selected 1−3 representatives from the "high-score" binding poses. Both the "best-fit" and "high-score" binding poses would be applied to MD simulations with subsequent binding free energy analysis to determine the final reasonable protein−ligand binding pose.

Short-length MD simulation was applied to further refine the protein−ligand complex structure. Here, we utilized the AMBER software (version 2014)[30] to carry out the MD simulation tasks. The GAFF force field parameters were applied to the ligand molecule. Its atomic partial charges were derived with the RESP method[31] based on the HF/6-31G* computation results given by the Gaussian software (version 09, released by Gaussian Inc).[32] The protein atoms were assigned according to the FF14SB template charges, and all ionizable residues were set at the default protonation states at pH = 7. The complex structure was then soaked in a cubic box of TIP3P water molecules with a margin of 10 Å along each dimension. Counterions (i.e., Na^+ or Cl^-) were added to neutralize the whole system.

A regular MD protocol was applied to the prepared protein−ligand−water complex structure, which includes the following: (1) Minimizing the complex structure using the steepest descent and the conjugate gradient algorithms. (2) Heating from 0 to

https://doi.org/10.1021/acsomega.2c02156

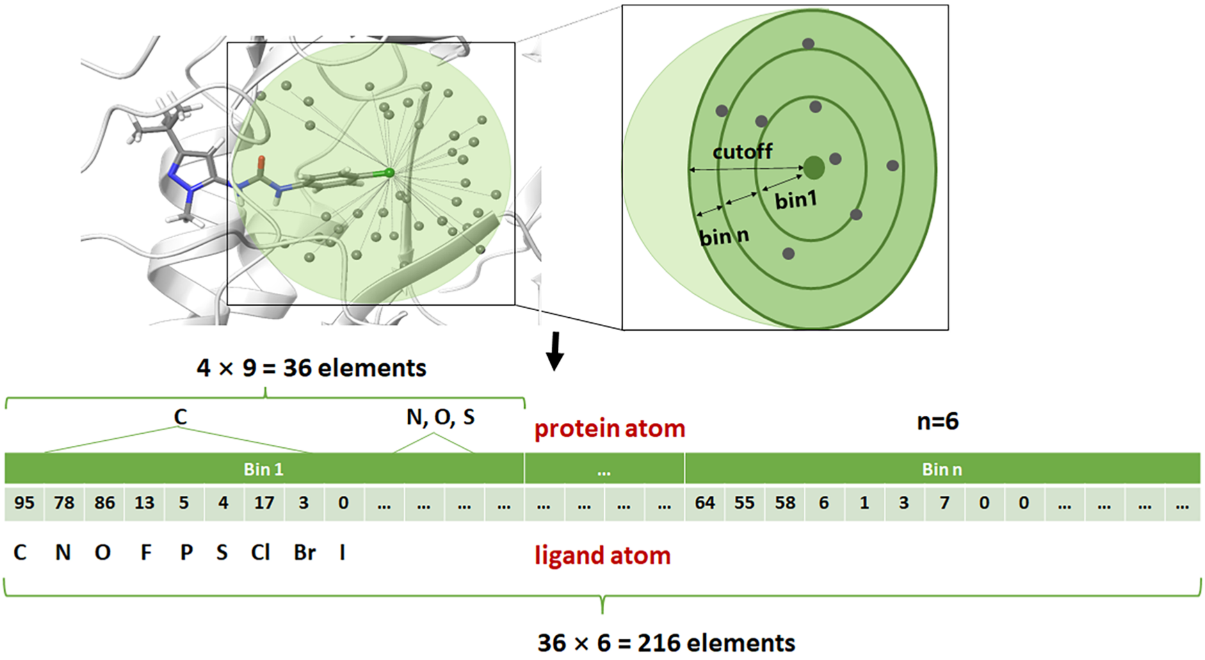


Figure 2. Definition of the protein−ligand atom pair descriptors in our RF model. Each protein−ligand complex was encoded as an 1D vector consisting of 216 elements, which included the occurrences of 36 protein−ligand atom pairs in six even bins (0−12 Å).

300 K in 200 ps followed by a 300 ps equilibration under the NPT ensemble. In this process, the heavy atoms of the protein−ligand complex were restrained with a constant of 10.0 kcal mol^{-1} Å^{-2}. (3) Restraining the protein−ligand complex from the crystal structure for a 2 ns production (restraint_wt = 10.0 kcal mol^{-1} Å^{-2}). For the protein−ligand complex from molecular docking, a production simulation of 2 ns long was performed without restraints.

We first analyzed the conformational fluctuations of the protein−ligand complex during the 2 ns MD simulation. In most systems, the complex structure reached convergence in terms of RMSD. A total of 45 complexes required another 2 ns production to reach convergence. The MM-GB/SA method was used to calculate the binding free energy of the final snapshot on the MD trajectory.[33] As mentioned above, "best-fit" and "high-score" binding poses may exist for the protein−ligand complex from molecular docking. According to the MM-GB/SA results, we selected the one with the more favorable (i.e., more negative) binding energy as the final structural model of the given protein−ligand complex.

2.4. Simple Model for Predicting Protein−Ligand Dissociation Constants.

Fedorov's work demonstrated what an RF model could achieve on their data set.[21] Because our data set is the same in nature as theirs, we also trained a similar RF model[34] for computing the k_{off} value of a given protein−ligand complex on our data set. To achieve this goal, the atom pair descriptors implemented in the RF-Score scoring function[35] were adopted here to construct the RF model. Four types of protein atoms (i.e., C, N, O, and S) and nine types of ligand atoms (i.e., C, N, O, F, P, S, Cl, Br, and I) were defined, resulting in a total of 36 protein−ligand atom pairs. Those atom pair descriptors count the occurrence of each possible protein−ligand atom pair within a distance range of 12 Å. This relatively wide range was intended to cover the protein structure relevant to the ligand dissociation pathway. This range was equally spaced into six bins, that is, [0−2), [2−4), [4−6), [6−8), [8−10), and [10−12]. As result, the final descriptor for encoding a protein−ligand complex was a one-dimensional (1D) vector consisting of 36 × 6 = 216 elements (Figure 2).

In order to evaluate the predictive power of this model, we selected the complexes formed by the three most populated proteins in our data set, that is, HSP90 (89 complexes), HIV-1 protease (39 complexes), and FAK (33 complexes), as the external test sets. Besides, 10 complexes including a large stapled peptide as the ligand were defined as the fourth test set, because those peptides had very different physiochemical properties from common small-molecule ligands. Based on the distance between two complexes calculated by the atom pair descriptors, the Kennard−Stone algorithm[36] was applied to split the remaining 509 complexes into a training set (407 complexes) and a validation set (102 complexes) in a ratio of 8:2. To examine whether the modeled structures can be used as an extended data set in the development of machine learning prediction models, we also trained and validated the RF models by using two sets of 312 crystal structures and 197 modeled structures separated from the 509 complexes. The two data sets were also split into training and validation sets in a ratio of 8:2 by using the Kennard−Stone algorithm. The RF models trained and tested on the crystal structures, modeled structures, and mixed structures were named "RF_crystal", "RF_model", and "RF_mixed", respectively.

We used the Scikit-learn (version 0.21.3) software[37] to construct the RF regression model. Note that three hyperparameters significantly affected the performance of an RF model. They were the number of data used in the model (*n_estimators*), the maximal depth of the decision trees (*max_depth*), and the criterion of feature selection (*criterion*). The "grid search" method was applied to sample those hyperparameters. It first determined the range of each hyperparameter and then attempted all combinations of those three hyperparameters to select the optimal one. During the sampling process, the value of *n_estimators* gradually increased from 10 to 600 at a step size of 10, and the value of *max_depth* increased from 2 to 30 at a step size of 2. With a certain set of

https://doi.org/10.1021/acsomega.2c02156

hyperparameters, the RF model was first generated by considering all samples in the training set and then evaluated on the validation set. The optimal set of hyperparameters was recorded when the Pearson correlation coefficient (R_p) between the predicted values and the experimental data in the validation set reached the maximum.

3. RESULTS AND DISCUSSION

3.1. Basic Features of the Data Set. Our data set mainly comes from two resources. Most of the data are kinetic constants manually curated from the literature included in the PDBbind database. The rest part of the data was taken from public literature. Basic information of the protein−ligand complexes in this data set, including PDB codes, experimental kinetic data, detection method, temperature, protein names, UniProt IDs, ligand SMILEs, and the related citations, are summarized in Table S1 in the Supporting Information. In all 680 records of the protein−ligand dissociation kinetic constants, over half of them (54.7%, 372) have the corresponding complex structures resolved by X-ray crystal diffraction method. For the remaining ones, their complex structures were derived through molecular modeling (i.e., molecular docking followed by short-length MD simulation). This data set, including the experimental k_{off} values and the 3D structures of all complexes, is free available from our PDBbind-CN web site (http://www.pdbbind.org.cn/download.php). This data set is named *PDBbind-koff-2020* because it is basically an extension of PDBbind version 2020.

The curated binding kinetic data k_{off} stem from a variety of detection methods (Figure 3). Overall, SPR took up the largest share, with 69.6% of all detection methods. Fluorescence-based methods ranked second, accounting for approximately 15.5%. Other methods were used less frequently, such as biolayer interferometry, isothermal titration calorimetry, radioligand binding analysis, and so on. The dissociation rate constants (k_{off}) of the protein−ligand complexes in this data set range from −3 to 7 (in log units), spanning over nearly 10 magnitudes (Figure 4). The Shapiro−Wilk normal distribution test[38] was performed on the data set, and the result showed that the distribution conformed to the normal distribution at the 95% confidence level (p-value = 0.144).

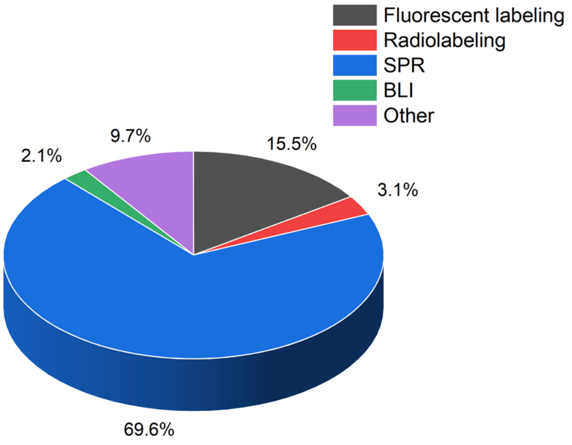


Figure 3. Distribution of the detection methods used for measuring the protein−ligand dissociation rate constants (k_{off}) included in our data set.

The entire data set was further clustered by the sequence of the protein molecule in each complex using the CD-hit program (v 4.5.4),[39] with the similarity threshold setting to 90%. The clustering results indicated that all samples in the data set could be divided into 155 groups, covering different protein families. The top 10 protein families are HSP90 (89 complexes), HIV-1 protease (39 complexes), focal adhesion kinase (33 complexes), p38α (26 complexes), soluble epoxide hydrolase (sEH) (19 complexes), thermolysin (17 complexes), lipid transfer protein (11 complexes), thrombin (11 complexes), FimH protein (11 complexes), and eukaryotic translation initiation factor (10 complexes). Distributions of six key “drug-likeness” properties of the ligand molecules in this data set, including molecular weight (MW), the number of rotatable bonds (RBs), the number of hydrogen bond donor atoms (HBDs), the number of hydrogen bond acceptor atoms (HBAs), the polar surface area (PSA),[40] and the computed log P value,[41] are also given in Figure 5. A total of 314 ligands satisfied the Lipinski’s rule of 5,[42] accounting for 46.2% of the entire set. However, there are also a number of ligand molecules with MW or other properties far beyond Lipinski’s rule of 5. In general, our data set shows significant diversity, whether it is from a variety of protein types or the wide distribution of the physiochemical properties of ligands.

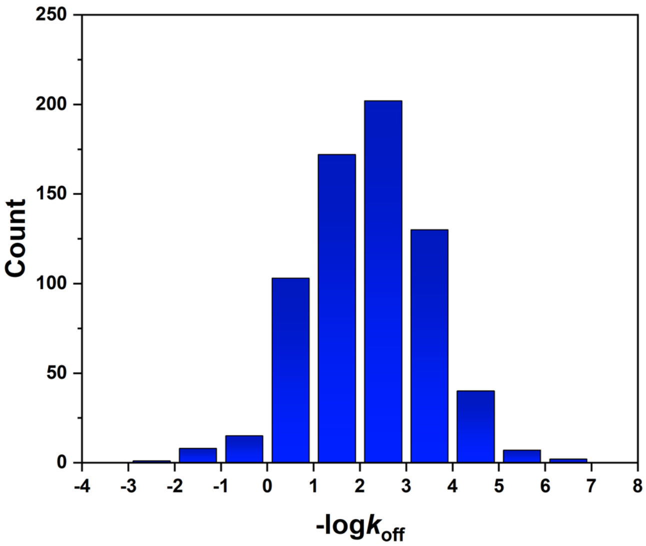


Figure 4. Distribution of the dissociation rate constants (k_{off} in s^{-1}) of the 680 ligands in our data set.

3.2. Baseline Model for Predicting Dissociation Rate Constants. Once publicly released, we expect that our data set will be employed by other researchers to develop computational models for predicting ligand dissociation rate constants. Using the ligand “drug-likeness” properties as the simplest molecular descriptors, we found that they were basically uncorrelated with experimental unbinding kinetics measured by −log k_{off} (Figure S1). It indicates that it is inappropriate to predict k_{off} by the linear regression of simple descriptors. Fedorov’s work proposed that the RF model could be used as a baseline to well correlate the protein−ligand interactions with their dissociation kinetics.[21] Therefore, we have derived a simple RF model based on our data set as described in the Methods section.

The purpose of this RF model is to set a baseline to testify the true power of other predictive models derived on the same data set. This RF model was constructed by using protein−ligand atom pairs as the input descriptors. Those descriptors are used to encode the protein environment relevant to protein−ligand interaction as well as ligand dissociation kinetics. The perform-

https://doi.org/10.1021/acsomega.2c02156

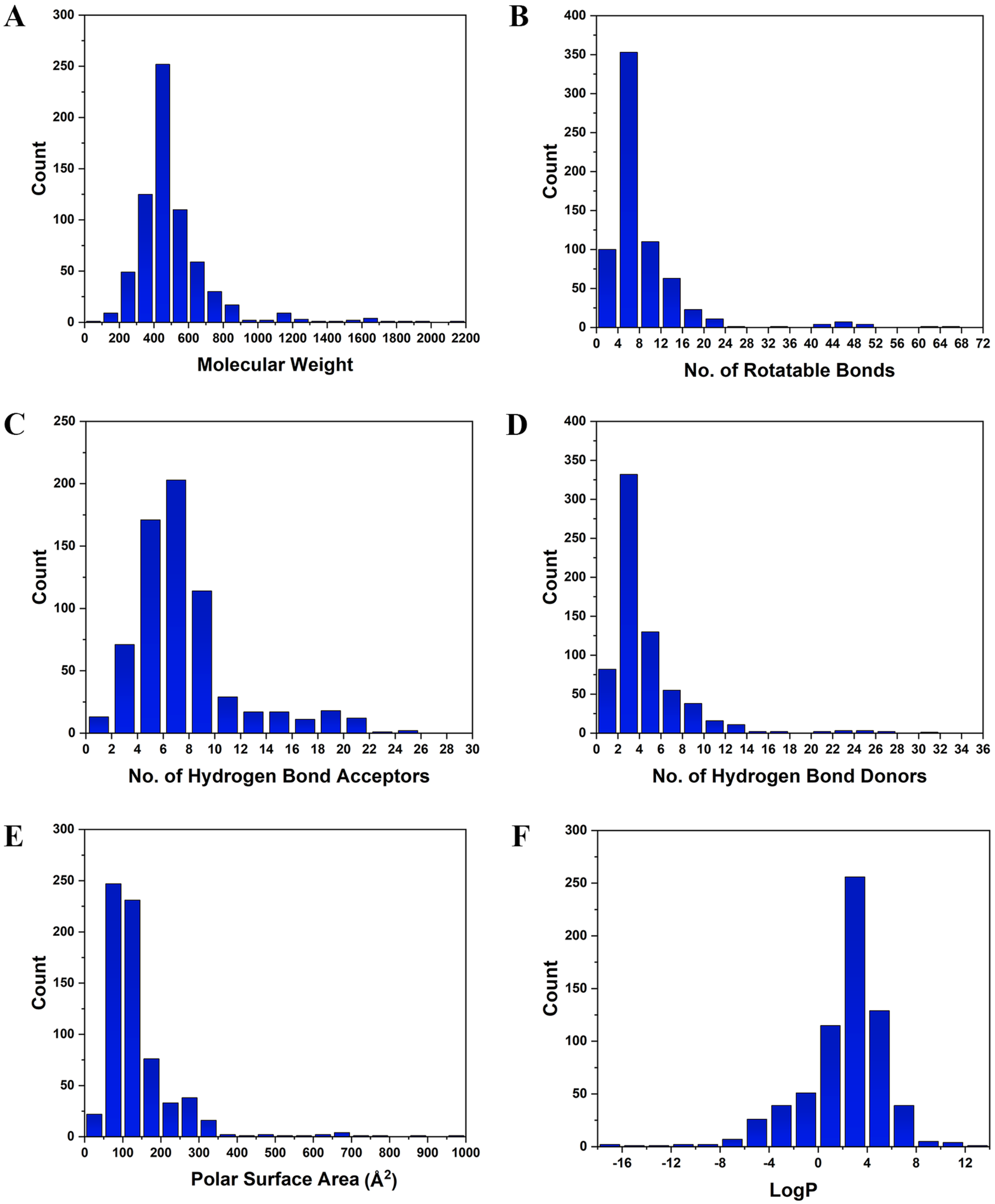


Figure 5. Distribution of six "drug-likeness" properties of the 680 ligands in our data set. (A) MW; (B) number of RBs; (C) number of HBAs; (D) number of hydrogen bond donors; (E) PSA; and (F) octanol−water partition coefficient (log *P*).

ance of all tested RF models, as evaluated on the validation set of 102 diverse protein−ligand complexes, is illustrated as a function of the hyperparameters including "*n_estimators*", "*max_depth*", and "*criterion*" in Figure 6. The final optimal RF model was selected when *n_estimators* = 50, *max_depth* = 14, and *criterion* = "*mae*". Under this setting, the RF model produced a Pearson correlation coefficient (R_p) of 0.968 and an RMSE of 0.474 on the training set (Figure 6A) and a Pearson correlation

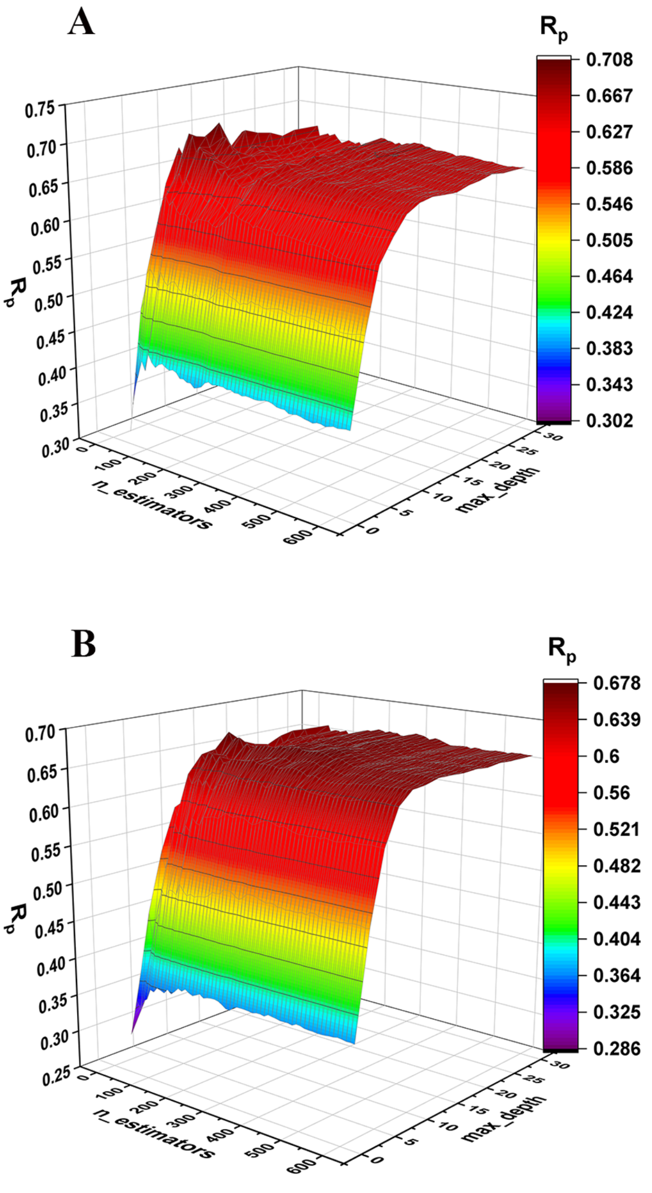


Figure 6. Pearson correlation coefficients between experimentally measured dissociation rate constants (−log k_{off}) in the validation set and the corresponding predicted values produced by the RF model. The correlation coefficients are plotted as the function of "*n_estimators*" and "*max_depth*" when the criterion was "*mae*" (A) or "*mse*" (B).

coefficient (R_p) of 0.706 and an RMSE of 0.986 (in −log k_{off} units) on the validation set (Figure 6B). These results were generally consistent with Fedorov's RF models. We also tested the ability of an empirical scoring function X-Score[43] to predict the k_{off} values. The Pearson correlation coefficients on the training set and test set were 0.227 and 0.101, respectively (Figure S2). It again demonstrates that these scoring models developed based on protein−ligand equilibrium thermodynamics cannot be used to predict ligand kinetics at all.

The RF model was tested on four external sets. On three of them, that is, HSP90, HIV-1 protease, and FAK, the Pearson correlation coefficients produced by this RF model were 0.501, 0.306, and 0.241, respectively, with corresponding RMSE values of 0.891, 1.602, and 1.044 (Figure 7C−E). Thus, this model exhibited moderate predictive power on the HSP 90 test set but obviously failed on the HIV-1 protease and FAK test sets. We have applied the t-SNE software[44] to visualize the protein−ligand complexes in our data set after dimension reduction conducted on atom-pair descriptors. The outcome reveals that the HSP90 complexes are relatively dispersed in the descriptor space compared to the HIV-1 protease complexes and FAK complexes in our data set (Figure 8). In other words, it indicates that the ligand molecules in the HIV-1 protease complexes and FAK complexes in our data set represent certain limited types of structure, and those types of structure are not well understood by our RF model. This RF model was also evaluated on the test set of large stapled peptides. Here, one can see a good correlation coefficient (R_p = 0.834), but the model does not reproduce the absolute values of experimental data well (Figure 7F).

Our results described above indicate that it might be too naïve to predict the protein−ligand dissociation rate constant based merely on a static protein−ligand complex structure. Again, our purpose is to showcase how well a baseline model can perform in those cases, and hopefully one will have a more realistic interpretation of other computational models, either machine-learned or not, derived for dissociation rate prediction.

To verify whether utilizing modeled structures as additional training samples will benefit the model performance, we also trained and validated the RF models by using two separate data sets containing only crystal structures and only modeled structures. The splitting of training and validation sets followed the same workflow as the previous one. Crystal structures and modeled structures in the HSP set and HIV-1 set were used as external test sets for these new models. FAK set and stapled peptide set were not considered for validations due to the limited sample size. The performances of three models (i.e., "RF_crystal", "RF_model", and "RF_mixed") are summarized in Table 1. Evaluation results showed that the model trained on modeled structures had comparable performance to the model trained on crystal structures. "RF_model" and "RF_crystal" produced Pearson correlation coefficients of 0.971 and 0.975 on the training sets and 0.677 and 0.635 on the validation sets, respectively. Compared to "RF_crystal", "RF_model" had a slightly higher RMSE value on the validation set. These results were a bit worse compared to the model with mixed structures, that is, "RF_mixed". It is actually understandable, because fewer samples were used for training these two separate models, leading to some key features probably missing in model learning. Thinking in turn, it also confirmed that using the modeled structures as an extended set is beneficial for the development of machine learning models.

Furthermore, we found that "RF_model" performed poorly on the external HSP and HIV-1 sets. It was partially due to the smaller size of the training set compared to the crystal structures. On the other hand, all modeled structures were constructed based on the MCS of the crystal ligands, which could be considered as a subset of the chemical space covered by the crystal structures. Consequently, a model trained on the modeled structures would miss some types that were present only in crystal structures and thus performed worse when it was extrapolated to an external test set. When the crystal structures and modeled structures were combined for training, the model achieved improved performance on both the validation set and external test sets. This further supports the use of modeled structures to augment training samples for developing machine learning models.

3.3. Comparison to Other Data Sets of Kinetic Data.

A few resources provide extensive collections of the thermodynamic binding affinity data of protein−ligand complexes, such as

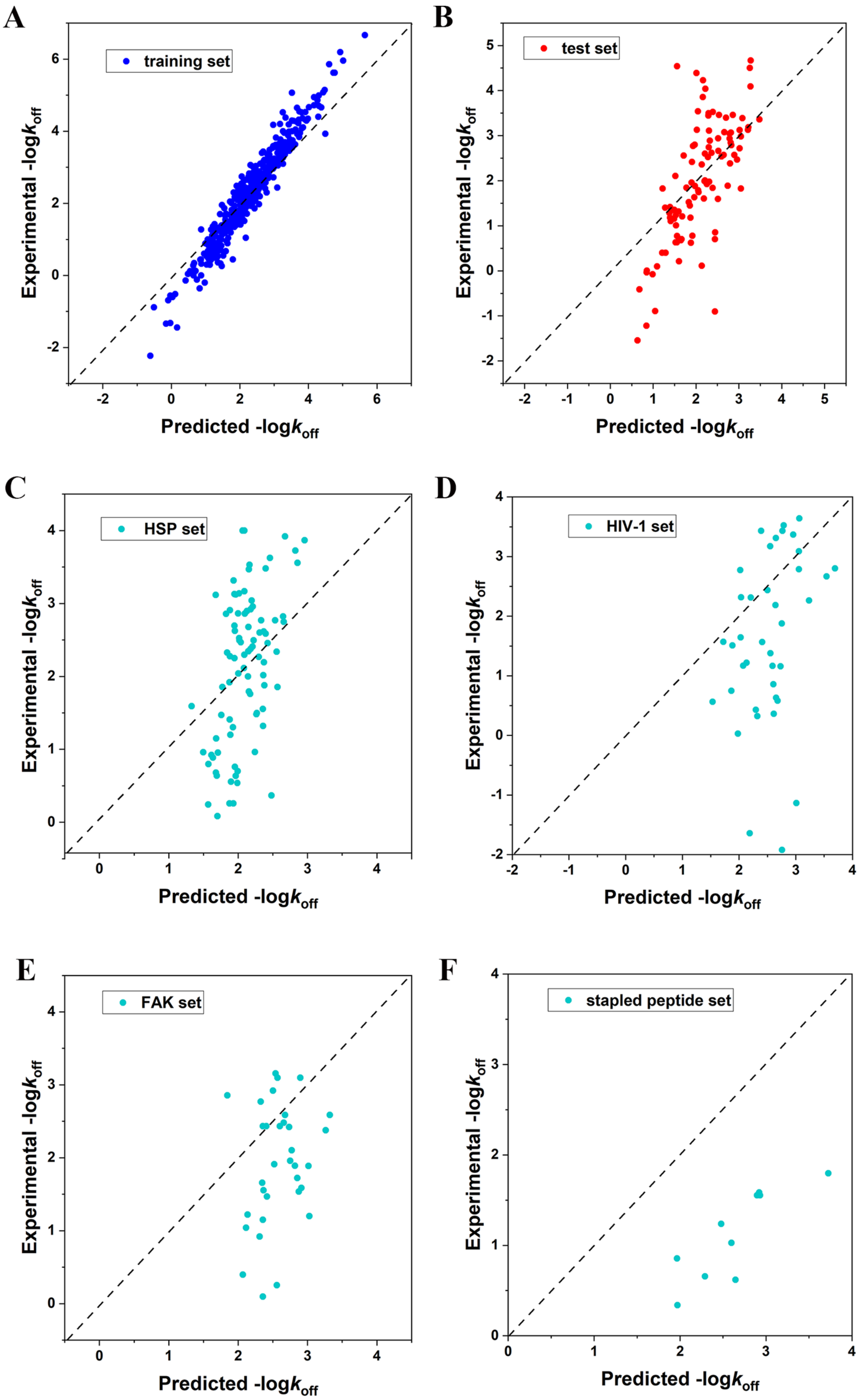


Figure 7. Correlation between experimentally measured dissociation rate constants (−log k_{off}) and the corresponding predicted values for: (A) training set, where $N = 407$, $R_p = 0.968$, and RMSE = 0.474; (B) validation set, where $N = 102$, $R_p = 0.706$, and RMSE = 0.986; (C) test set of HSP90 complexes, where $N = 89$, $R_p = 0.501$, and RMSE = 0.891; (D) test set of HIV-1 protease complexes, where $N = 39$, $R_p = 0.306$, and RMSE = 1.602; (E) test set of FAK complexes, where $N = 33$, $R_p = 0.241$, and RMSE = 1.044; and (F) test set of complexes containing stapled peptide ligands, where $N = 10$, $R_p = 0.834$, and RMSE = 1.546.

PDBbind,[22,23] ChEMBL,[14,15] Binding MOAD,[16] and BindingDB.[20] However, public collections of kinetic data are rather rare. KDBI[17,18] provides experimentally measured kinetic data for protein−protein, protein−nucleic acid, and protein−small molecule interactions. Its latest release provides a total of 19,263 entries of kinetic data. However, those kinetic data are not

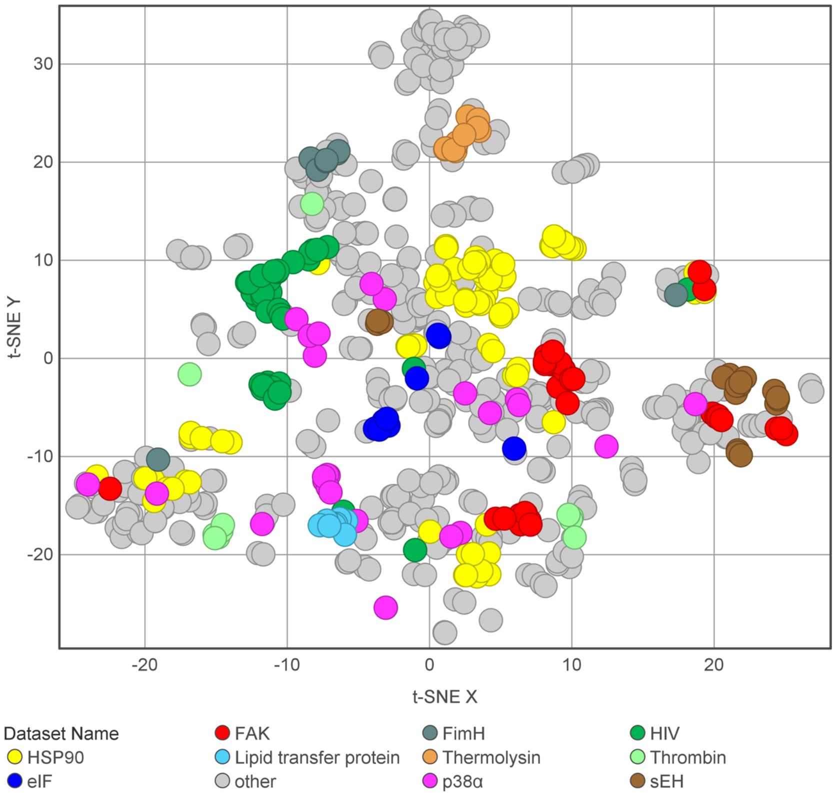


Figure 8. t-SNE visualization of the protein–ligand complexes in our data set based on protein–ligand atom pair descriptors. Complexes of the 10 most populated proteins are rendered as colored circles, while other complexes are rendered as gray circles.

Table 1. Performances of Models Trained on Crystal Structures, Modeled Structures, and Mixed Structures[a]

data sets	training set			validation set			HSP set			HIV-1 set		
	N	R_p	RMSE	N	R_p	RMSE	N	R_p	RMSE	N	R_p	RMSE
RF_crystal	250	0.975	0.457	62	0.635	0.943	39	0.463	0.968	12	0.337	1.782
RF_model	158	0.971	0.458	39	0.677	1.003	50	−0.348	1.012	27	−0.023	1.374
RF_mixed	407	0.968	0.474	102	0.706	0.986	89	0.501	0.891	39	0.306	1.602

[a] The validation results for the FAK set and stapled peptide set are not presented here due to the small sample size.

associated with corresponding 3D complex structures, which limited their applications to molecular modeling. KOFFI is a similar database.[19] It added more information, such as the assay method, device, chip, and so on. It collects a total of 1705 individual entries, most of which are kinetic data of protein–protein and protein–nucleic acid complexes, with very few records of protein–small molecules. Note that KOFFI does not provide the linkage between kinetic data and 3D complex structures either.

BindingDB is a public, web-accessible database of binding affinity data,[20] focusing chiefly on the interactions of protein considered to be drug targets with small, drug-like molecules. It also collects the binding kinetic data of some targets and small molecules. The data set established in this work was compared with the kinetic data included in BindingDB in terms of sample size, protein type, and so on. As of May 10, 2021, BindingDB contained a total of 301 non-redundant k_{off} records of 12 protein types. Each protein has an average of nearly 25 records of its ligands. In fact, the distribution of the kinetic data in different protein types is extremely unequal. The protein human leukocyte elastase had 194 records of kinetic data, accounting for about two-thirds of all data. In contrast, our data set includes kinetic data for 680 protein–ligand complexes, which cover 155 different proteins. The size of our data set is about twice the collection available in BindingDB. Besides, our data set obviously presents a wider of structural diversity.

The most attractive feature of our data set is that a protein–ligand complex structure is provided for each record of kinetic data. Here, crystal structures are available for 372 complexes, which account for 54.7% of the entire data set. Other similar database rarely collects related crystal structures for these protein–ligand complexes, such as KDBI and KOFFI. BindingDB also contains only 14 crystal structures for their k_{off} data. For the protein–ligand complex with no available crystal structure, we constructed the initial binding model based on a similar ligand template through molecular docking. The binding models derived from crystal structures and molecular docking were further refined by a short-time MD simulation. With these 3D structures and the corresponding kinetic data, we can use machine learning methods to explore the QSKR between the targets and their ligands, which cannot be achieved by other databases such as KDBI, BindingDB, and so on.

When we were preparing this article, Fedorov et al. reported a similar collection of kinetic data.[21] Their data set consisted of 501 protein–ligand complexes with experimentally measured dissociation rate constants. A comprehensive comparison of the two data sets was carried out from the aspects of the k_{off} data distribution, protein types, ligand structural diversity, and complex structures. Although the −log k_{off} data of the two sets are mainly located in the range of −2 to 6, they have quite different distributions (Figure S3A,B). The peak value is biased toward the interval of 3−4 in Fedorov's set, which does not pass the Shapiro−Wilk normal distribution test (p-value < 0.05). The k_{off} data collected in our set generally obey the Shapiro−Wilk normal distribution (p-value = 0.144), with the peak falling between 2 and 3. Setting the protein sequence identity above 90% as the threshold, Fedorov's set contains 53 different protein families, with an average of 9.5 protein−ligand complexes per family. Our set covers 155 protein families, with an average of 4.4 complexes. We further compared the top 10 protein types that appear frequently in the two data sets. In Fedorov's set, the top 10 proteins accounted for more than two-thirds of the total sample size, while in our data set, this number is only 39%, reflecting more abundant protein types (Figure S4A,B). It is worth mentioning that there are four common protein types that frequently appear in the two data sets, including HSP90, HIV-1 protease, p38α, and sEH. It indicates that they are the targets of most interest in kinetics-guided drug discovery research. The physiochemical properties of the ligands were also compared (Figure S5). Broader distributions of six physiochemical properties were indicated for the ligands in our set. In summary, our data set is more diverse in terms of kinetic data, target types, and ligand properties. In addition, more than 50% of the protein−ligand complexes in our set have crystal structures resolved by X-ray crystallography, while this percentage in Fedorov's set is only 33%. It demonstrates the advantages of our data set in developing structure-based prediction models.

4. CONCLUSIONS

A reliable computational model for predicting protein−ligand binding/unbinding kinetic properties can hopefully provide a more rational basis for structure-based drug design. Apparently, a high-quality data set is the basis for developing such QSKR models, especially for machine-learning models. In this work, we have compiled a data set of experimental dissociation rate constants and the corresponding protein−ligand complex structures. This data set was mostly curated from over 38,000 references accumulated during our work of updating the PDBbind database. It includes 680 records of protein−ligand dissociation rate constants (k_{off}), ranging from −3 to 7 (in log units). The target proteins included in this data set could be clustered into 155 groups at a sequence similarity cutoff of 90%, covering a range of enzymes, receptors, transporters, and so on. The physiochemical properties of the ligand molecules included in this data set also had a wide distribution. This data set has exceeded those described in the literature in terms of sample size and structural diversity. In addition, we have derived a simple RF model based on this data set for predicting k_{off} values. It is of course naïve to expect that binding kinetics can be depicted from a static protein−ligand complex structure. Thus, this model should be considered as a baseline for testifying the true power of other more sophisticated QSKR models. The whole data set, namely, *PDBbind-koff-2020*, is freely available from the PDBbind-CN web site (http://www.pdbbind.org.cn/download.php).

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.2c02156.

Correlation analysis between six ligand properties and experimental −log k_{off}, prediction performance of the empirical scoring function X-Score on our data set, and comparisons of Fedorov's data set with ours (PDF)

Complete information of the protein−ligand complexes in our data set, including PDB codes, experimental dissociation rate constant (k_{off}), detection method, temperature, protein names, UniProt IDs, ligand SMILEs and the related citations; index files of training sets, validation sets, and external test sets used in this article (XLSX)

Source code for generating protein−ligand atom pair descriptors and RF models (ZIP)

■ AUTHOR INFORMATION

Corresponding Authors

Hai-Xia Lin − *Department of Chemistry, College of Sciences, Shanghai University, Shanghai 200444, People's Republic of China*; Email: haixialin@staff.shu.edu.cn

Renxiao Wang − *Department of Medicinal Chemistry, School of Pharmacy, Fudan University, Shanghai 201203, People's Republic of China*; orcid.org/0000-0003-0485-0259; Email: wangrx@fudan.edu.cn

Yan Li − *Department of Medicinal Chemistry, School of Pharmacy, Fudan University, Shanghai 201203, People's Republic of China*; orcid.org/0000-0002-8259-2470; Email: li_yan@fudan.edu.cn

Authors

Huisi Liu − *Department of Chemistry, College of Sciences, Shanghai University, Shanghai 200444, People's Republic of China; State Key Laboratory of Bioorganic and Natural Products Chemistry, Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences, Shanghai 200032, People's Republic of China*

Minyi Su − *State Key Laboratory of Bioorganic and Natural Products Chemistry, Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences, Shanghai 200032, People's Republic of China;* Present Address: Parc Cientific de Barcelona, Carrer de Baldiri Reixac, 4-8, Torre R, 04A05, 08028 Barcelona

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.2c02156

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This study was financially supported by the Ministry of Science and Technology of China (National Key Research Program, grant no. 2016YFA0502302), the National Natural Science Foundation of China (grant nos 81725022, 81430083, 21472227, 21661162003, 21673276, and 21472226), and the Chinese Academy of Sciences (Strategic Priority Research Program, grant no. XDB20000000).

■ REFERENCES

(1) Swinney, D. C. The role of binding kinetics in therapeutically useful drug action. *Curr. Opin. Drug Discov. Dev* **2009**, *12*, 31−39.

(2) Guo, D.; Mulder-Krieger, T.; IJzerman, A. P.; Heitman, L. H. Functional efficacy of adenosine A2A receptor agonists is positively correlated to their receptor residence time. *Br. J. Pharmacol.* **2012**, *166*, 1846−1859.

(3) Kruse, A. C.; Hu, J.; Pan, A. C.; Arlow, D. H.; Rosenbaum, D. M.; Rosemond, E.; Green, H. F.; Liu, T.; Chae, P. S.; Dror, R. O.; Shaw, D. E.; Weis, W. I.; Wess, J.; Kobilka, B. K. Structure and dynamics of the M3 muscarinic acetylcholine receptor. *Nature* **2012**, *482*, 552−556.

(4) Doornbos, M. L. J.; Cid, J. M.; Haubrich, J.; Nunes, A.; van de Sande, J. W.; Vermond, S. C.; Mulder-Krieger, T.; Trabanco, A. A.; Ahnaou, A.; Drinkenburg, W. H.; Lavreysen, H.; Heitman, L. H.; IJzerman, A. P.; Tresadern, G. Discovery and Kinetic Profiling of 7-Aryl-1,2,4-triazolo[4,3-a]pyridines: Positive Allosteric Modulators of the Metabotropic Glutamate Receptor 2. *J. Med. Chem.* **2017**, *60*, 6704−6720.

(5) Copeland, R. A.; Pompliano, D. L.; Meek, T. D. Drug-target residence time and its implications for lead optimization. *Nat. Rev. Drug Discovery* **2006**, *5*, 730−739.

(6) Schuetz, D. A.; de Witte, W. E. A.; Wong, Y. C.; Knasmueller, B.; Richter, L.; Kokh, D. B.; Sadiq, S. K.; Bosma, R.; Nederpelt, I.; Heitman, L. H.; Segala, E.; Amaral, M.; Guo, D.; Andres, D.; Georgi, V.; Stoddart, L. A.; Hill, S.; Cooke, R. M.; De Graaf, C.; Leurs, R.; Frech, M.; Wade, R. C.; de Lange, E. C. M.; IJzerman, A. P.; Müller-Fahrnow, A.; Ecker, G. F. Kinetics for drug discovery: an industry-driven effort to target drug residence time. *Drug Discov. Today* **2017**, *22*, 896−911.

(7) Dickson, A.; Tiwary, P.; Vashisth, H. Kinetics of ligand binding through advanced computational approaches: A review. *Curr. Top. Med. Chem.* **2017**, *17*, 2626−2641.

(8) Bruce, N. J.; Ganotra, G. K.; Kokh, D. B.; Sadiq, S. K.; Wade, R. C. New approaches for computing ligand-receptor binding kinetics. *Curr. Opin. Struct. Biol.* **2018**, *49*, 1−10.

(9) Platter, N.; Noe, F. Protein conformational plasticity and complex ligand-binding kinetics explored by atomistic simulations and Markov models. *Nat. Commun.* **2015**, *6*, 7653.

(10) Miao, Y.; Bhattarai, A.; Wang, J. Ligand Gaussian accelerated molecular dynamics (LiGaMD): characterization of ligand binding thermodynamics and kinetics. *J. Chem. Theory Comput.* **2020**, *16*, 5526−5547.

(11) De Benedetti, P. G.; Fanelli, F. Computational modeling approaches to quantitative structure-binding kinetics relationships in drug discovery. *Drug Discov. Today* **2018**, *23*, 1396−1406.

(12) Qu, S.; Huang, S.; Pan, X.; Yang, L.; Mei, H. Constructing interconsistent, reasonable, and predictive models for both the kinetic and thermodynamic properties of HIV-1 protease inhibitors. *J. Chem. Inf. Model.* **2016**, *56*, 2061−2068.

(13) Ganotra, G. K.; Wade, R. C. Prediction of drug-target binding kinetics by comparative binding energy analysis. *ACS Med. Chem. Lett.* **2018**, *9*, 1134−1139.

(14) Davies, M.; Nowotka, M.; Papadatos, G.; Dedman, N.; Gaulton, A.; Atkinson, F.; Bellis, L.; Overington, J. P. ChEMBL web services: streamlining access to drug discovery data and utilities. *Nucleic Acids Res.* **2015**, *43*, W612−W620.

(15) Mendez, D.; Gaulton, A.; Bento, A. P.; Chambers, J.; De Veij, M.; Félix, E.; Magariños, M. P.; Mosquera, J. F.; Mutowo, P.; Nowotka, M.; Gordillo-Marañón, M.; Hunter, F.; Junco, L.; Mugumbate, G.; Rodriguez-Lopez, M.; Atkinson, F.; Bosc, N.; Radoux, C. J.; Segura-Cabrera, A.; Hersey, A.; Leach, A. R. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.* **2019**, *47*, D930−D940.

(16) Smith, R. D.; Clark, J. J.; Ahmed, A.; Orban, Z. J.; Dunbar, J. B., Jr.; Carlson, H. A. Updates to binding MOAD (Mother of All Databases): polypharmacology tools and their utility in drug repurposing. *J. Mol. Biol.* **2019**, *431*, 2423−2433.

(17) Ji, Z. L.; Chen, X.; Zhen, C. J.; Yao, L. X.; Han, L. Y.; Yeo, W. K.; Chung, P. C.; Puy, H. S.; Tay, Y. T.; Muhammad, A.; Chen, Y. Z. KDBI: Kinetic Data of Bio-molecular Interactions database. *Nucleic Acids Res.* **2003**, *31*, 255−257.

(18) Kumar, P.; Han, B. C.; Shi, Z.; Jia, J.; Wang, Y. P.; Zhang, Y. T.; Liang, L.; Liu, Q. F.; Ji, Z. L.; Chen, Y. Z. Update of KDBI: Kinetic data of bio-molecular interaction database. *Nucleic Acids Res.* **2009**, *37*, D636−D641.

(19) Norval, L. W.; Krämer, S. D.; Gao, M.; Herz, T.; Li, J.; Rath, C.; Wöhrle, J.; Günther, S.; Roth, G. KOFFI and Anabel 2.0-a new binding kinetics database and its integration in an open-source binding analysis software. *Database* **2019**, *2019*, baz101.

(20) Gilson, M. K.; Liu, T.; Baitaluk, M.; Nicola, G.; Hwang, L.; Chong, J. BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res.* **2016**, *44*, D1045−D1053.

(21) Amangeldiuly, N.; Karlov, D.; Fedorov, M. V. Baseline model for predicting protein-ligand unbinding kinetics through machine learning. *J. Chem. Inf. Model.* **2020**, *60*, 5946−5956.

(22) Liu, Z.; Li, Y.; Han, L.; Li, J.; Liu, J.; Zhao, Z.; Nie, W.; Liu, Y.; Wang, R. PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics* **2015**, *31*, 405−412.

(23) Liu, Z.; Su, M.; Han, L.; Liu, J.; Yang, Q.; Li, Y.; Wang, R. Forging the basis for developing protein-ligand interaction scoring functions. *Acc. Chem. Res.* **2017**, *50*, 302−309.

(24) 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.

(25) Bietz, S.; Urbaczek, S.; Schulz, B.; Rarey, M. Protoss: a holistic approach to predict tautomers and protonation states in protein-ligand complexes. *J. Cheminf.* **2014**, *6*, 12.

(26) Duan, R.; Xu, X.; Zou, X. Lessons learned from participating in D3R 2016 Grand Challenge 2: compounds targeting the farnesoid X receptor. *J. Comput.-Aided Mol. Des.* **2018**, *32*, 103−111.

(27) Alekseenko, A.; Kotelnikov, S.; Ignatov, M.; Egbert, M.; Kholodov, Y.; Vajda, S.; Kozakov, D. ClusPro LigTBM: Automated template-based small molecule docking. *J. Mol. Biol.* **2020**, *432*, 3404−3410.

(28) 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.

(29) Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L. Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J. Med. Chem.* **2004**, *47*, 1750−1759.

(30) Case, D. A.; Babin, V.; Berryman, J. T.; Betz, R. M.; Cai, Q.; Cerutti, D. S.; Cheatham, T. E.; Darden, T. A.; Duke, R. E.; Gohlke, H.; Goetz, A. W.; Gusarov, S.; Homeyer, N.; Janowski, P.; Kaus, J.; Kolossvary, I.; Kovalenko, A.; Lee, T. S.; LeGrand, S.; Luchko, T.; Luo, R.; Madej, B.; Merz, K. M.; Paesani, F.; Roe, D. R.; Roitberg, A.; Sagui, C.; Salomon-Ferrer, R.; Seabra, G.; Simmerling, C. L.; Smith, W.; Swails, J.; Walker, R. C.; Wang, J.; Wolf, R. M.; Wu, X.; Kollman, P. A. *AMBER 14*; University of Carlifornia: San Francisco, 2014.computer program

(31) Bayly, C. I.; Cieplak, P.; Cornell, W.; Kollman, P. A. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: The RESP model. *J. Phys. Chem.* **1993**, *97*, 10269−10280.

(32) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams-Young, D.; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery, J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Keith, T.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo,

C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J.. *Gaussian 09*, Revision A.02; Gaussian, Inc.: Wallingford CT, 2016.computer program

(33) Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A.; Cheatham, T. E. Calculating Structures and Free Energies of Complex Molecules: Combining Molecular Mechanics and Continuum Models. *Acc. Chem. Res.* **2000**, *33*, 889−897.

(34) Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5−32.

(35) Ballester, P. J.; Schreyer, A.; Blundell, T. L. Does a More Precise Chemical Description of Protein-Ligand Complexes Lead to More Accurate Prediction of Binding Affinity? *J. Chem. Inf. Model.* **2014**, *54*, 944−955.

(36) Kennard, R. W.; Stone, L. A. Computer Aided Design of Experiments. *Technometrics* **1969**, *11*, 137−148.

(37) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.

(38) Royston, J. P. An extension of Shapiro and Wilk's W test for normality to large samples. *J. R. Stat. Soc., C: Appl. Stat.* **1982**, *31*, 115−124.

(39) Niu, B.; Fu, L.; Sun, S.; Li, W. Artificial and natural duplicates in pyrosequencing reads of metagenomic data. *BMC Bioinf.* **2010**, *11*, 187.

(40) Ertl, P.; Rohde, B.; Selzer, P. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *J. Med. Chem.* **2000**, *43*, 3714−3717.

(41) Cheng, T.; Zhao, Y.; Li, X.; Lin, F.; Xu, Y.; Zhang, X.; Li, Y.; Wang, R.; Lai, L. Computation of octanol-water partition coefficients by guiding an additive model with knowledge. *J. Chem. Inf. Model.* **2007**, *47*, 2140−2148.

(42) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **2001**, *46*, 3−26.

(43) Wang, R.; Lai, L.; Wang, S. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 11−26.

(44) Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579−2605.