JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

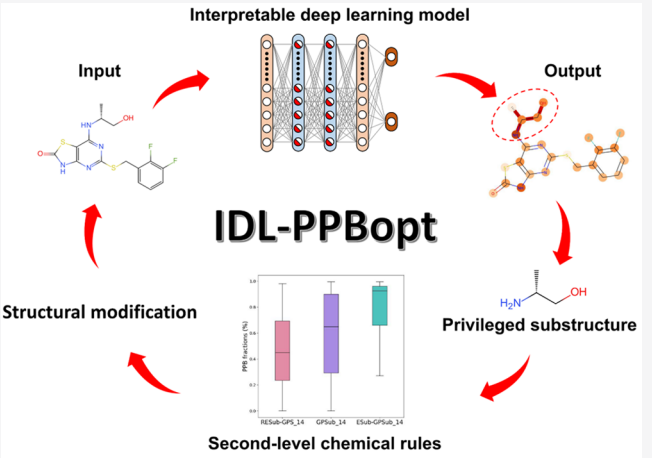
IDL-PPBopt: A Strategy for Prediction and Optimization of Human Plasma Protein Binding of Compounds via an Interpretable Deep Learning Method

Chaofeng Lou, Hongbin Yang, Jiye Wang, Mengting Huang, Weihua Li, Guixia Liu, Philip W. Lee, and Yun Tang*

Cite This: *J. Chem. Inf. Model.* 2022, 62, 2788−2799 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The prediction and optimization of pharmacokinetic properties are essential in lead optimization. Traditional strategies mainly depend on the empirical chemical rules from medicinal chemists. However, with the rising amount of data, it is getting more difficult to manually extract useful medicinal chemistry knowledge. To this end, we introduced IDL-PPBopt, a computational strategy for predicting and optimizing the plasma protein binding (PPB) property based on an interpretable deep learning method. At first, a curated PPB data set was used to construct an interpretable deep learning model, which showed excellent predictive performance with a root mean squared error of 0.112 for the entire test set. Then, we designed a detection protocol based on the model and Wilcoxon test to identify the PPB-related substructures (named privileged substructures, PSubs) for each molecule. In total, 22 general privileged substructures (GPSubs) were identified, which shared some common features such as nitrogen-containing groups, diamines with two carbon units, and azetidine. Furthermore, a series of second-level chemical rules for each GPSub were derived through a statistical test and then summarized into substructure pairs. We demonstrated that these substructure pairs were equally applicable outside the training set and accordingly customized the structural modification schemes for each GPSub, which provided alternatives for the optimization of the PPB property. Therefore, IDL-PPBopt provides a promising scheme for the prediction and optimization of the PPB property and would be helpful for lead optimization of other pharmacokinetic properties.

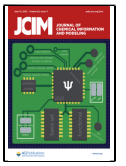


1. INTRODUCTION

Pharmacokinetic properties are closely related to drug safety and efficacy.[1−3] During lead optimization, medicinal chemists seek to design rational structural modification schemes for lead compounds, aiming to obtain favorable pharmacokinetic properties and sufficient bioactivity. Traditional strategies, such as scaffold hopping[4] and bioisosteric replacement,[5] mainly depend on the empirical chemical rules from medicinal chemists.[4,5] However, given the rising amount of data and complexity of chemical and biological systems, it is getting more difficult for medicinal chemists to manually extract related chemical rules.[6] Accordingly, researchers developed many computational methods to automatically learn hidden medicinal chemistry knowledge from large data sets for the prediction and optimization of pharmacokinetic properties, such as machine learning-based quantitative structure−activity relationship (QSAR) models[7−9] and matched molecular pairs analysis (MMPA).[10−12] However, these strategies have certain limitations. For example, QSAR models cannot provide structural modification schemes,[13] and MMPA can only be used to extract practical transformations from molecules sharing the same context.[14] Therefore, more efficient attempts to predict and optimize pharmacokinetic properties should be considered.

As an important pharmacokinetic property, plasma protein binding (PPB) shows the binding affinity of a drug with plasma proteins, which can modulate the effective concentration of the drug at the pharmacological target.[15] When the drug is absorbed into the body, it will selectively bind to plasma proteins. In most cases, such binding is reversible and there is a

Received: March 14, 2022
Published: May 24, 2022



https://doi.org/10.1021/acs.jcim.2c00297
J. Chem. Inf. Model. 2022, 62, 2788−2799

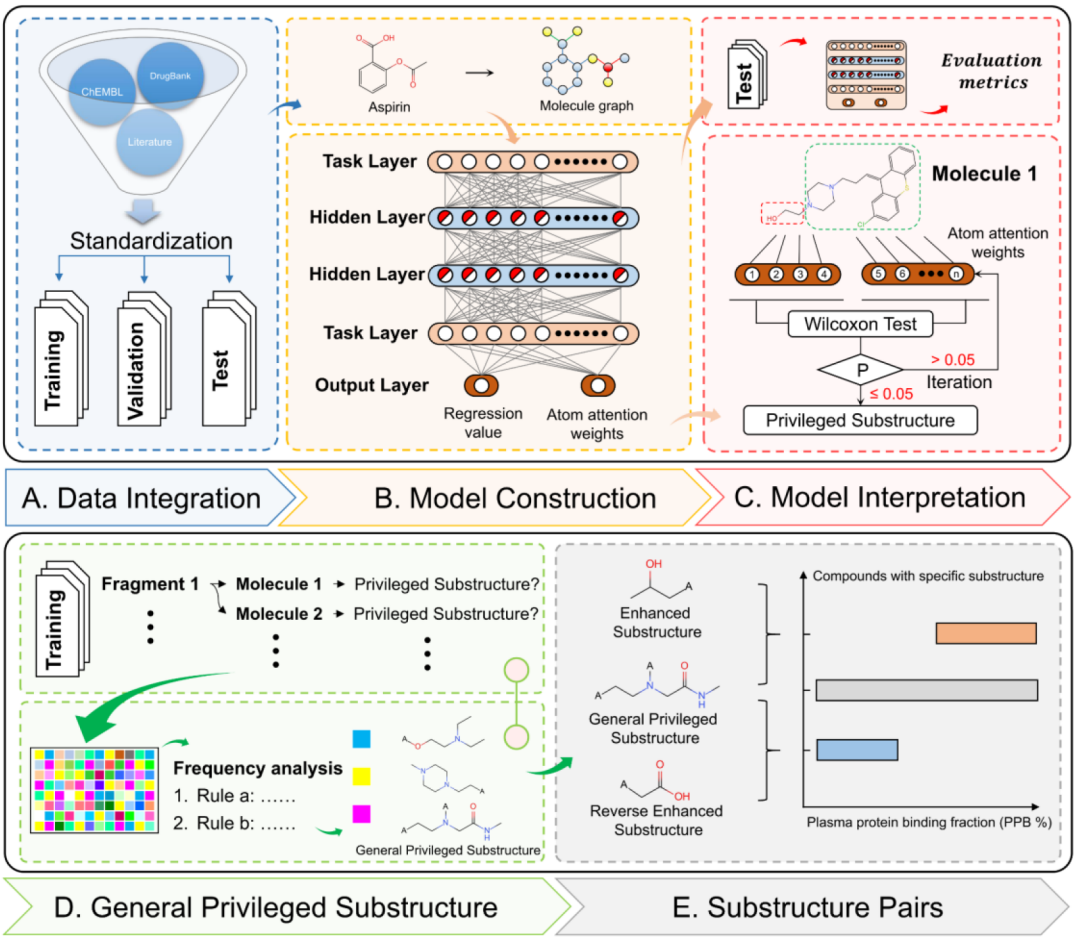


Figure 1. Whole workflow of this study, consisting of the following five steps: (A) data collection and preparation, (B) network architecture of Attentive FP, (C) model evaluation and identification of privileged substructures, (D) data mining of general privileged substructures through frequency analysis, and (E) definition of substructure pairs, namely ESub-GPSub pairs and RESub-GPSub pairs.

balance between bound and free species.[16] However, only the unbound drug can bind to the specific pharmacological target and exert its efficacy, while the bound form is stored in the plasma proteins and released slowly to prolong the duration of drug action.[17,18] The higher affinity to plasma proteins can increase the effective half-life of drugs, but higher dosing is required to achieve the effective concentration for treatment. From this perspective, the binding affinity of compounds and plasma proteins can determine the apparent volume of distribution of drugs, thus affecting drug absorption, distribution, metabolism, and excretion (ADME) properties.[19,20] In addition, drugs with high plasma protein affinity may cause drug−drug interactions due to PPB displacement, which can potentially precipitate side effects.[21] Therefore, it is extremely valuable to measure and optimize the PPB property of new drug candidates in drug discovery.

In the past decade, significant progress has been made in the field of machine learning methods, especially the proposals of many deep learning algorithms with novel neural architectures,[22−24] which greatly promote drug discovery and development.[25] The special molecular representation way and multilevel network architecture enable these algorithms to automatically extract the most relevant information regarding the properties of interest, which overcome the limitations of manual selection of molecular descriptors and fingerprints. More importantly, because of the utilization of model interpretation algorithms,[26−28] many QSAR models that are built with deep graph neural network algorithms have gradually gotten rid of the trap of the "black box model"[29] and visualized more learned knowledge to medicinal chemists.[30−32] However, most researchers focus on predictive ability while ignoring the enlightenment that machine learning knowledge can bring us. For example, interpretable deep learning models can help determine task-related atoms (or substructures) for each molecule,[33−35] which is of great significance for structural simplification to avoid "molecule obesity" in lead optimization.[36] Therefore, interpretable deep learning techniques show great superiority in predicting pharmacokinetic properties and should have more applications in the field of lead optimization.

In this study, we proposed a computational strategy named IDL-PPBopt (interpretable deep learning for plasma protein binding optimization) to predict and optimize the plasma protein binding of compounds based on an interpretable deep learning method (Figure 1). The workflow of this study consists of three essential steps: (1) to construct a well-trained interpretable deep learning model (Figure 1A,B); (2) to identify PPB-related substructures (Figure 1C); (3) to learn from the compounds containing privileged substructures and get second-level chemical rules (Figure 1E), which can be used to optimize the PPB property of compounds (Figure 1D). The proposed strategy was validated by literature, a validation set, and a test set, and obtained a satisfactory result, which

https://doi.org/10.1021/acs.jcim.2c00297
J. Chem. Inf. Model. 2022, 62, 2788−2799

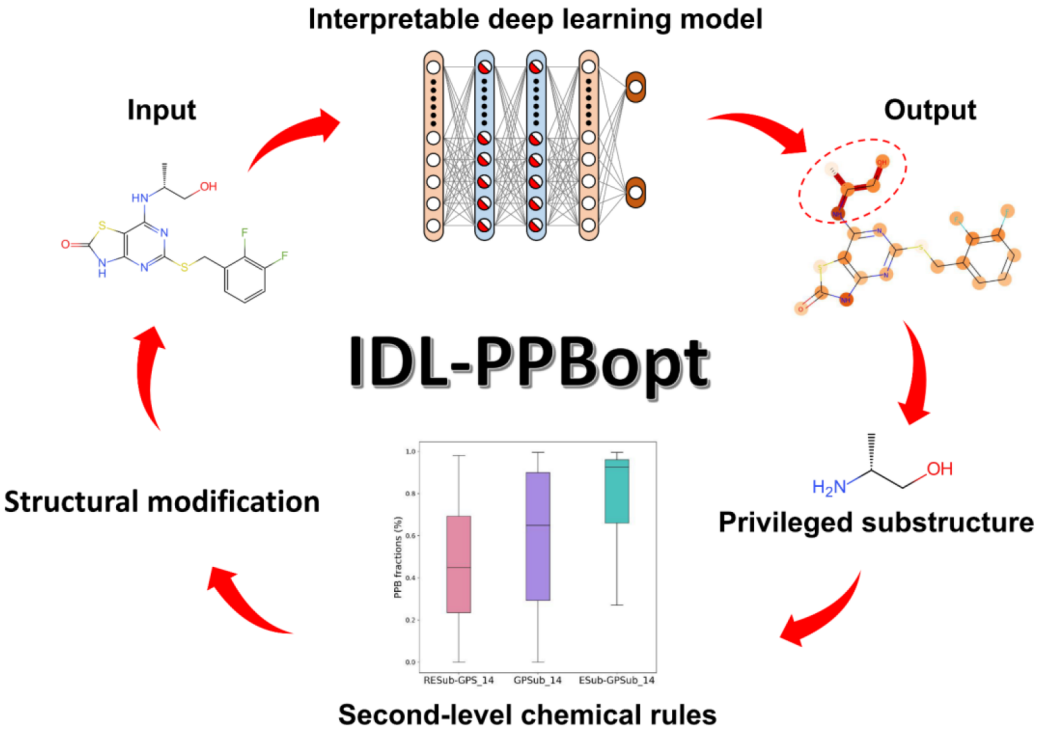


Figure 2. Diagram illustrating the workflow of the IDL-PPBopt strategy.

provided a new perspective for lead optimization in drug discovery and development.

2. MATERIALS AND METHODS

2.1. Data Collection and Preparation. The initial records of human PPB data were collected from the literature,[37,38] and two publicly accessible databases, ChEMBL[39] (version 27) and DrugBank[40] (version 5.1.6). All PPB records from Votano's data set,[37] Zhu's data set,[38] and DrugBank were obtained first. The PPB fractions in humans and fractions unbound in human plasma ($f_{u,p}$) from ChEMBL were then integrated and transformed into uniform PPB records.

The initial data set was then curated as follows. All compounds were first converted into canonical SMILES format. Then, mixtures and inorganic compounds were removed, and salts were converted into corresponding acids or bases by Pipeline Pilot Software 2017 R2 (BIOVIA, USA). When there were multiple PPB records for a compound, the standard deviation (SD) would be calculated. The average value was calculated and taken if the calculated SD was less than 0.02.[9,41] Otherwise, the most appropriate record would be manually selected or the compound would be removed. Subsequently, we assigned the PPB records randomly into the training set, validation set, and test set with the ratio of 8:1:1 and yielded three data sets that were identically distributed.

2.2. Overview of the IDL-PPBopt Strategy. The IDL-PPBopt strategy has three important components: a well-trained interpretable deep learning model that has accurate predictions and meaningful model interpretations, a detection protocol to identify substructures most relevant to the task, and a practical protocol for the derivation of second-level chemical rules. As described in Figure 2, for a given compound, the IDL-PPBopt strategy utilizes an interpretable deep learning model to make predictions and interpretations, followed by a detection protocol to capture task-related substructure. Finally, through learning from the compounds containing the substructure, the second-level chemical rules are derived as substructure pairs to guide structural modification of the given compound. The task-related substructures learned by the model are termed privileged substructures (PSubs). Substructures from second-level chemical rules are termed enhanced substructures (ESubs) and reverse enhanced substructures (RESubs). Substructure pairs are composed of a PSub and a second-level substructure, i.e., ESub or RESub.

2.3. Methods for Model Construction and Evaluation. In this study, we constructed an interpretable deep learning model with the attentive fingerprint algorithm (AFP).[26] The AFP algorithm is a special graph neural network architecture with a graph attention mechanism (Figure 1B), which has been proven to achieve excellent predictive performance on a variety of data sets.[26] More strikingly, this algorithm can extract nonlocal intramolecular interactions and visualize the model-learned knowledge.

First, a total of nine types of atomic features and four types of bond features were calculated as the node and edge features for each molecular graph. A fully connected layer was used to generate an initial vector of uniform length for each atom and its neighbors. During the next two hidden layers that contained attention mechanisms, the initial vector was updated after aggregating more neighborhood information, and a new state vector for the whole molecule was generated by assembling the state vector of each atom, where the attention weights were assigned to the neighbors based on contribution. Finally, a fully connected layer was used for task training and prediction. The whole network architecture adopted the Bayesian optimization method for hyper-parameter tuning and Adam optimizer for gradient descent optimization. Notably, the attention weights were also updated during the model iteration.

https://doi.org/10.1021/acs.jcim.2c00297
J. Chem. Inf. Model. 2022, 62, 2788−2799

In each iteration, a new model was generated and the performance was evaluated with the validation set. To avoid overfitting and determine the final model with excellent performance, we applied an early stop strategy based on the evaluation results of the training set and validation set. Thus, if the model performance was not improved in 8 epochs on the training set and 10 epochs on the validation set, the training process would be terminated early.

The final well-trained model was evaluated by the validation set and test set. Here, three statistical indexes, namely mean absolute error (MAE, eq 1 in Table S1), root-mean-square error (RMSE, eq 2 in Table S1), and determination coefficient (R^2, eq 3 in Table S1), were introduced to evaluate all models. Both MAE and RMSE could measure the errors between predicted and observed values, but the latter was more sensitive to outliers. R^2 was employed to calculate the degree of linear correlation between predicted and observed values. In general, the best-fitting model should have lower MAE and RMSE values, while R^2 is close to 1.

2.4. Identification of PPB-Related Substructure Patterns. *2.4.1. Identification of Privileged Substructure for a Single Molecule.* The well-trained interpretable PPB model (iPPB model) can output the contribution score of each atom to the PPB fractions through atom attention weights. The atom attention weights measure the contribution of each atom to the final molecular representation, and the sum of attention weights of all atoms in each molecule is 1. On this basis, these substructures composed of atoms with significantly higher atom attention weights are defined as privileged substructures (PSubs). Here, we developed a detection protocol based on the Wilcoxon test method to extract PSubs from the iPPB model (Figure 1C). This protocol contained three steps.

(1) Fragmentation. A recursive algorithm[42] was used to obtain all possible substructures (3−18 atoms) of a given compound without breaking the ring bonds. In addition, duplicate substructures were removed and virtual atoms (denoted by "*" or "A") were generated at the broken bond.

(2) Statistical tests. Each substructure was matched to the corresponding molecule, and the one-sided Wilcoxon test was performed to detect whether the given substructure had significantly higher atom attention weights than the remaining part of the molecule (i.e., $P < 0.05$). Those substructures that passed the one-sided Wilcoxon test were recorded for further analysis. The Python packages of RDKit (version 2018.09.3.0, http://www.rdkit.org) and SciPy (version 1.6.2, http://www.scipy.org) were employed for substructure matching and the Wilcoxon test.

(3) Elimination of redundancy. Redundancy was an inevitable problem in the fragmentation phase, resulting in a set of similar or partially repeated substructures. For simplicity, only the largest substructure was kept as it contained more structural information.

2.4.2. Data Mining of General Privileged Substructures from Training Set. General privileged substructures (GPSubs) are defined as the substructure patterns that are frequently labeled as PSub by the iPPB model, representing a series of model-learned important chemical rules. We first employed a recursive algorithm to get all PSubs from the training set compounds without breaking the ring bonds and then calculated their frequencies. Each GPSub should meet the following two rules: (1) to pass the one-sided Wilcoxon test in at least 50 molecules; (2) to pass the one-sided Wilcoxon test in more than half of the molecules containing this substructure. In short, the former promised the universality of GPSub in the training set, which helped eliminate contingency and systematic errors; while the latter followed the assumption that the frequency of a random event could be approximated to probability if the sample size was large enough.

2.5. Derivation of Second-Level Chemical Rules. The second-level chemical rules for each PSub are derived from compounds in the training set that contain the corresponding PSub, divided into two types of substructures, enhanced substructure (ESub) and reverse enhanced substructure (RESub). ESub is defined as a special substructure that is derived from the corresponding PSub but presents more often in compounds with relatively higher PPB values. On the contrary, RESub is defined as a special substructure that appears together with corresponding PSub but occurs more frequently in compounds with relatively lower PPB values. The presence of ESub or RESub may induce the PPB fractions of the compounds containing corresponding PSub to change in a specific direction (Figure 1E). In this study, we designed a derivation method based on a statistical test and took GPSub as examples to derive second-level chemical rules and summarize them into substructure pairs.

We first collected compounds containing specific GPSubs from the training set and integrated them into a new data set. A recursive algorithm[42] was then employed to get all possible new substructures (3−18 atoms) from the new data set. For each new substructure, according to whether it was present in the compounds or not, we divided the new data set into two classes. Subsequently, the Wilcoxon test was performed on the PPB fractions of these two classes and the substructure would be recorded if the p-value was less than 0.05. For ESub, compounds containing the substructure would have significantly higher PPB fractions, whereas compounds with RESub showed significantly lower PPB fractions. In this way, a set of substructure pairs, namely ESub-GPSub pairs and RESub-GPSub pairs, were obtained after eliminating redundant substructures.

To customize suitable structural modification schemes for each GPSub, we calculated and recorded the topological distance of each substructure pair in different molecules. Figure S1 illustrates the definition of topological distance in a molecule, which meant the shortest path between two substructures. If two substructures were partially or fully overlapped, their topological distance should be 0; whereas if two substructures were separated, the topological distance would be greater than 1 bond. Subsequently, a scoring function, namely S_{PPB} (eq 1), was defined to evaluate the impact of ESub ($S_{PPB} > 0$) and RESub ($S_{PPB} < 0$) on the corresponding GPSub through the average PPB fractions difference.

$$S_{PPB} = \bar{F}_{ER} - \bar{F}_{GP} \quad (1)$$

where $\bar{F}_{ER}$ represents the average PPB fraction of compounds containing the ESub-GPSub pairs or RESub-GPSub pairs, and $\bar{F}_{GP}$ represents the average PPB fraction of compounds only containing the corresponding GPSub.

https://doi.org/10.1021/acs.jcim.2c00297
J. Chem. Inf. Model. 2022, 62, 2788−2799

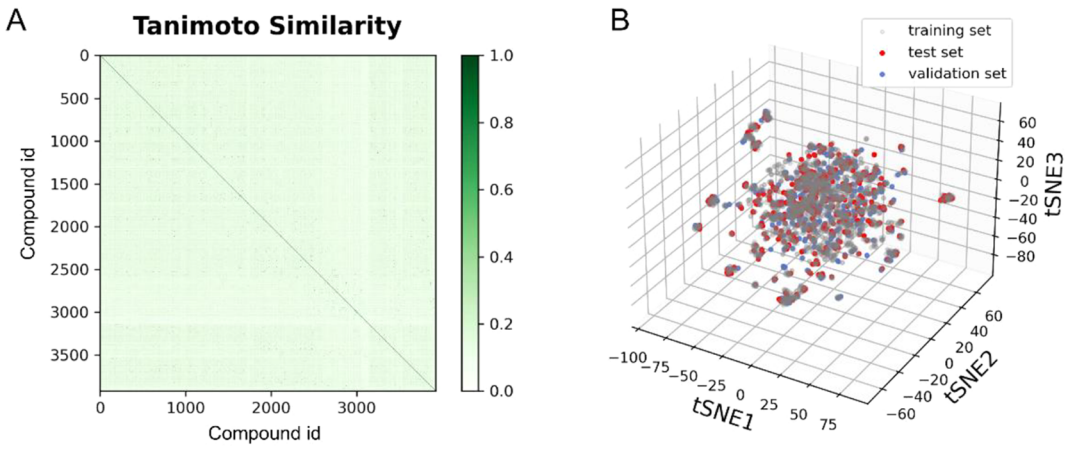


Figure 3. (A) Heat map of Tanimoto similarity of the total data set with Morgan fingerprint. (B) t-SNE of chemical diversity analysis across training set, validation set, and test set.

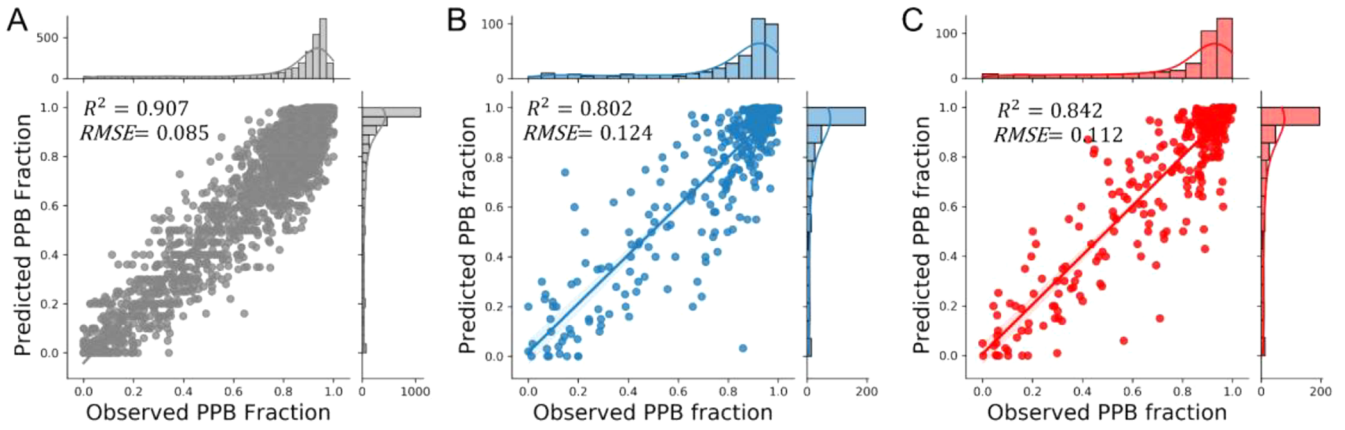


Figure 4. Scatter plots of predicted PPB fractions and observed PPB fractions of the (A) training set, (B) validation set, and (C) test set.

3. RESULTS

3.1. Construction of Interpretable PPB Model.

3.1.1. Data Collection and Analysis. From the literature and public databases, we collected 4926 PPB records. After preparation, a total of 3921 PPB fraction data with structural diversity were obtained. The comprehensive data set was then split into a training set containing 3136 compounds, a validation set of 392 compounds, and a test set of 393 compounds. These three data sets shared a similar and uneven distribution. Specifically, 2625 (nearly 65%) molecules had high plasma proteins binding affinities (PPB fraction > 80%), while only 518 (nearly 14%) molecules were at low PPB fraction levels (PPB fraction < 40%). All chemical information, including SMILES, PPB records, and data set classifications, is given in Table S2.

The t-distributed stochastic neighbor embedding (t-SNE)[43] method and Tanimoto similarity index, both using Morgan fingerprints as molecular features, were performed to visualize the chemical space of the data set (Figure 3). The overall color of the Tanimoto similarity heat map was light green with an average similarity of 0.121, indicating the high diversity of the compounds in the data set. On the other hand, according to t-SNE analysis, the central regions of the validation set and test set were highly overlapped with that of the training set, and only very few molecules in the test set fell in the area beyond the training set space, indicating that the validation set and test set were reasonable for model optimization and evaluation. These results showed the structural diversity of the data set compounds and the applicability of the validation set and test set.

3.1.2. Model Construction and Evaluation. On the basis of the curated PPB data set, we constructed an interpretable deep learning model for PPB prediction with the AFP method. All hyper-parameters are shown in Table S3. Figure S2 illustrates the learning curves of RMSE and R^2 on the training set and validation set during the whole iterative process. Overall, during the training process, the RMSE of the training set was getting lower and the R^2 was gradually close to 1, which meant that the model was continuously learning from the training set. In addition, the R^2 curve of the validation set reached a plateau after 50 epochs, indicating enough knowledge had been learned. Finally, the model at epoch 54 was considered as the best model, with excellent performance for training set (RMSE = 0.085, R^2 = 0.907) and validation set (RMSE = 0.124, R^2 = 0.802). Furthermore, 393 molecules in the test set were used to evaluate the generalizability of the model, obtaining a favorable RMSE of 0.112 and R^2 of 0.841.

We further analyzed the scatter plots of observed values and predicted values. As shown in Figure 4, compounds with high binding affinities (PPB fraction > 80%) were around the diagonal line, but a small number of compounds at moderate PPB fraction levels (40% ≤ PPB fraction ≤ 80%) and low PPB

https://doi.org/10.1021/acs.jcim.2c00297
J. Chem. Inf. Model. 2022, 62, 2788−2799

fraction levels (PPB fraction < 40%) were outside the diagonal area, indicating a better predictive ability of the iPPB model in high-binding data. Overall, the prediction results of the three data sets demonstrated that the iPPB model achieved a favorable fitting ability despite the unbalanced distribution of the PPB fraction data set.

3.1.3. Comparison with Other Models. We further listed the performance of other published PPB models (Table 1).

Table 1. Summary of the Optimal Performance of the Published Models

year	author	MAE	RMSE	R^2
2006	Votano[37]	0.141	0.186	0.700
2013	Zhu[38]	0.119	0.182	
2017	Wang[46]	0.142	0.182	0.704
2018	Tajimi[47]	0.194	0.230	
2018	Watanable[48]	0.100	0.145	0.728
2018	Sun[9]	0.114	0.145	0.817
2020	Yuan[41]	0.079	0.143	0.762
2021	Jiménez-Luna[33]		0.208	0.520

Those published models were mostly built with molecular descriptors and machine learning algorithms and evaluated with their test sets. It could be observed that the R^2 value of these models was between 0.7 and 0.8. However, because of the unavailability of those models, we could not evaluate them with the same test set. Thus, two public deep learning models, the graph convolutional neural network (GCN) model[44] and the multitask graph attention framework (MGA)[45] model, were used in further comparison with our test set. As listed in Table 2, our iPPB model and the MGA model performed

Table 2. Comparison of Different Models with the Same Test Set

model	MAE	RMSE	R^2
GCN model[44]	0.127	0.175	0.620
MGA model[45]	0.074	0.114	0.835
iPPB model	0.075	0.112	0.841

better than the GCN model across all evaluation indexes, indicating that the graph neural network with graph attention mechanism could achieve better performance. Overall, our iPPB model achieved excellent performance in the field of PPB prediction.

3.2. Evaluation of PPB-Related Substructures. On the basis of the well-trained interpretable deep learning model and Wilcoxon test, we designed a detection protocol to identify PSub through atom attention weights for each compound. In addition to PPB, we also tested the applicability of the detection protocol in other end points, including chemical Ames mutagenicity, human ether-a-go-go relate gene (hERG) blockers, HIV-1 protease inhibitors, CYP2C8 inhibitors, and Rho-associated protein kinases (ROCK) inhibitors. We developed a well-trained interpretable deep learning model for each end point, and their performances are summarized in Table S4. Then, we visualized some molecules in each end point and identified their PSubs with the detection protocol. As shown in Figure S3, the detected PSubs for Ames mutagenicity[49] and hERG blockers[50] were consistent with the structural alerts (SAs) reported in the literature. In addition, for the other end points with known PSubs, including HIV-1 protease inhibitors,[51] CYP2C8 inhibitors,[52] and ROCK inhibitors,[53] the newly detected PSubs also shared similar structural features. The results demonstrated that our detection protocol combined with interpretable deep learning models could indeed extract task-related substructures. However, current interpretable deep learning models still had limitations, such as we cannot get the information of whether the PSubs were correlated positively or negatively with activity. In this study, PSub represented the substructure within a given molecule that had great contributions to the PPB effects, and two case studies were used to prove the definition.

3.2.1. Case Study: PSub for a Single Molecule. CHEMBL372443 (PubChem CID: 10181815) was reported as a potent and selective $\alpha_v\beta_3/\alpha_v\beta_5$ antagonist with subnanomolar *in vitro* affinity (Figure 5A). However, because of its high binding affinity to serum albumin, its efficacy in animal models was far from satisfactory. According to the structure−activity analysis of human serum albumin (HSA) binding data of organic acids, it was reported that the incorporation of polar groups into a given molecule could dramatically decrease the affinity toward HSA.[54,55] Therefore, several analogs were designed and synthesized, where the introduction of *N*,*N*-dimethylaminomethyl pyridyl group greatly reduced PPB fractions, while maintaining subnanomolar activity (Figure 5B).[56] Our iPPB model successfully predicted the directionality of the PPB change of these two molecules and highlighted important atoms and substructures that might account for this change. By visualizing the atom attention weights in two molecules, the new introduced *N*,*N*-dimethylaminomethyl pyridyl group, especially the tertiary amine atom, was given a darker red than the original phenyl group (Figure 5A,B), indicating the newly introduced group had great contributions on the binding affinity of human plasma proteins. We further identified PSub for the newly designed compound and highlighted it in red. The detected PSub mainly contained a *N*,*N*-dimethylaminomethyl pyridyl group, consistent with the results of Raboisson et al.[56]

In another case of influenza cap-dependent endonuclease inhibitors, medicinal chemists disclosed the structure−activity relationships for a novel series of carbamoylpyridine bicycle compounds and found that the N-3 substituent was critical for the PPB effect.[57] Two representative compounds, CHEMBL4465872 (PubChem CID: 68096717) and CHEMBL4462899 (PubChem CID: 67471579), were selected and predicted with our iPPB model. From the prediction results (Figure 5C,D), the magnitude and directionality of PPB change were similar to the result measured in the rat experiment. Besides, the PSub detected from both compounds contained the N-3 substituents, indicating that the iPPB model had indeed learned the important structural information on the compounds related to the PPB effect.

Overall, the above two case studies demonstrated that the derivation of PSub from the iPPB model was reasonable and it could help us to get insights into the model-learned knowledge related to PPB property.

3.2.2. General Privileged Substructure Analysis. Different from the PSub for a single molecule, GPSubs are a series of important PPB-related substructure patterns extracted from the data set. Through frequency analysis of all fragments from the training set (Table S5), a total of 22 GPSubs were finally obtained from the training set.

As shown in Figure 6, nitrogen-containing groups were identified as the most important functional groups for the PPB

https://doi.org/10.1021/acs.jcim.2c00297
J. Chem. Inf. Model. 2022, 62, 2788−2799

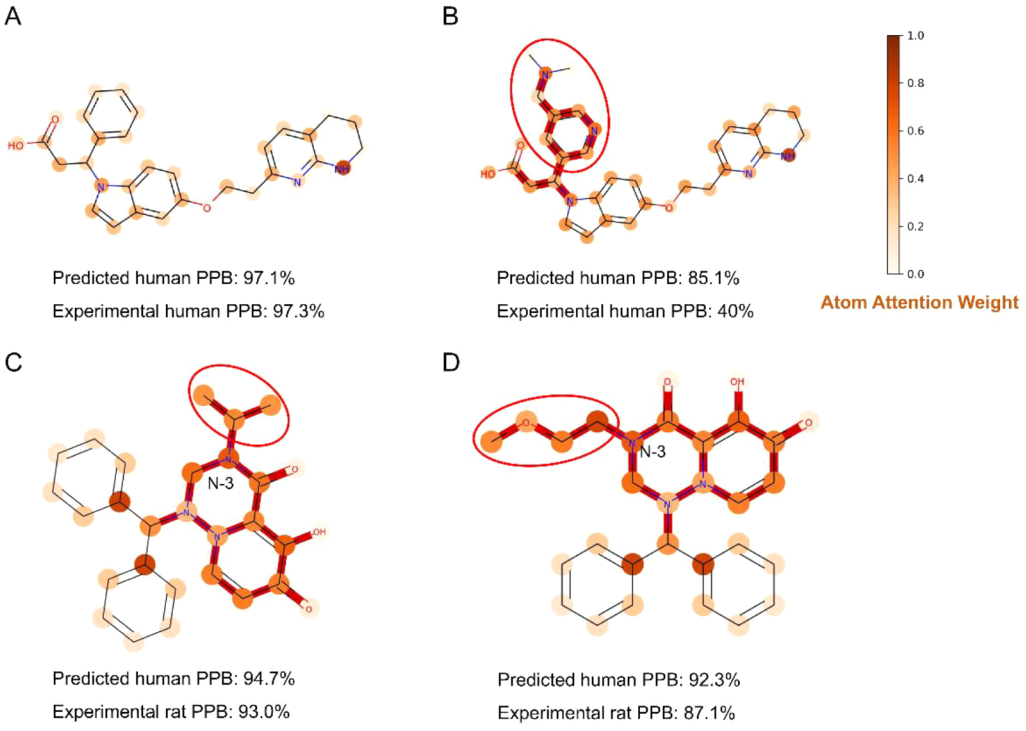


Figure 5. Structures of (A, B) two $\alpha_v\beta_3/\alpha_v\beta_5$ antagonists and (C, D) two influenza cap-dependent endonuclease inhibitors with experimental PPB fractions and predicted PPB fractions. The atom attention weights learned from the iPPB model were used to highlight the atoms. The darker the color is, the greater the attention weights are. The substructure highlighted in red represents the PSub and the red circle represents important substructures reported in the literature.

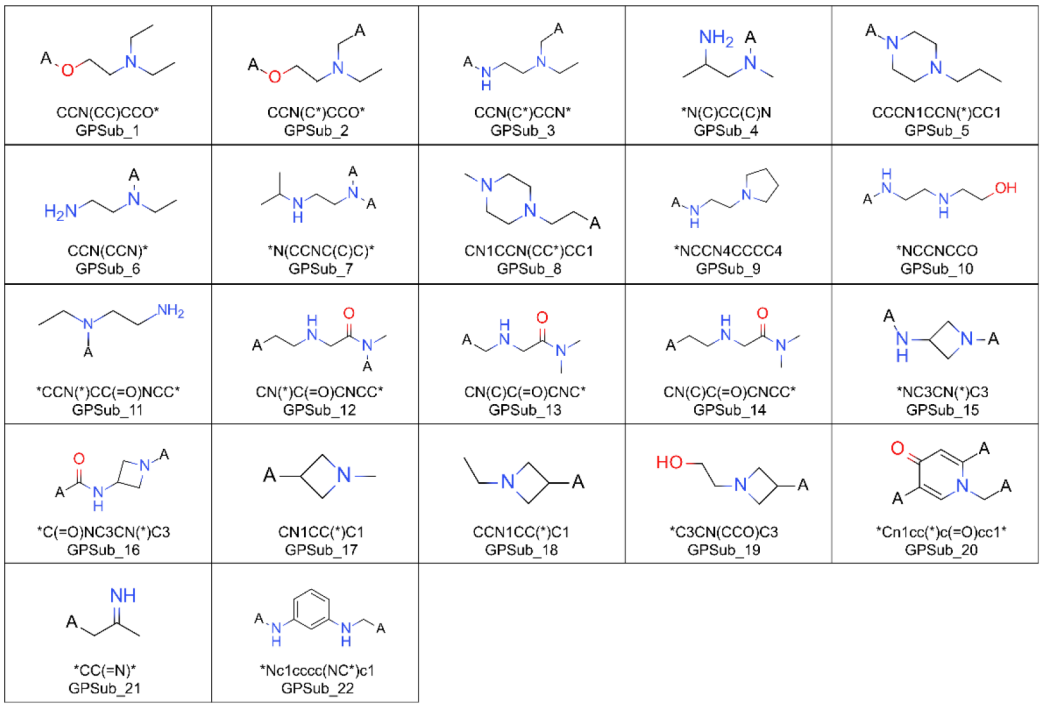


Figure 6. Structures of 22 GPSubs, where "A" represents virtual atoms that can be replaced by any non-hydrogen atoms. Note: the 22 GPSubs are in no particular order.

effect because they were present in all GPSubs, which were consistent with the conclusions reported in the previous literature.[58,59] For instance, Hajduk et al. reported a chemometric analysis of ligand binding to human serum albumin with 74 chemical fragments and found that nitrogen-containing groups were given the highest weighting coefficients.[58] In addition, Yun et al. evaluated three published QSAR models for PPB prediction with molecular descriptors, of which the number of basic functional groups was identified as one of the most critical chemical characteristics.[59] Thus, considering the

https://doi.org/10.1021/acs.jcim.2c00297
J. Chem. Inf. Model. 2022, 62, 2788−2799

interactions of drugs with proteins, we could presume that, those nitrogen-containing groups may form hydrogen-bonding or electrostatic interactions between drugs and plasma proteins, promoting the reversible binding and maintaining the concentrations of free species.

We further found that diamines with two carbon units between two amino groups were another common structural feature, where one of the amino groups could be replaced by an oxygen atom, and two amino groups could form a cyclic ring. Among these 22 GPSubs, such a structural feature of diamines could be detected from 17 GPSubs, indicating that great emphasis was placed on diamines by the iPPB model. Furthermore, azetidine derivatives (GPSub_15, GPSub_16, GPSub_17, GPSub_18, GPSub_19), 1-methyl-1,4-dihydropyridin-4-ketone (GPSub_20), ethanimine (GPSub_21), and (3-aminophenyl) amine (GPSub_22) were also highlighted as relevant substructures corresponding to the binding affinity of compounds and plasma proteins, of which the substructure of azetidine derivatives had been reported to be used in PPB optimization for a better half-life.[60]

In total, 22 GPSubs were identified from the model. Among them, we found that nitrogen-containing groups, especially diamines with two carbon units and azetidine, were important structural features.

3.3. Second-Level Chemical Rules for PPB Optimization. Given the important role of PSubs in the PPB effect, we supposed that the structural modification around PSubs could be an effective way for PPB optimization.

3.3.1. Relationships between Substructure Combinations and PPB Fractions. Taking GPSub_17 as an example, we analyzed the relationship between GPSub_17 and other surrounding substructures within the molecule. We integrated those compounds containing GPSub_17 and extracted another interesting substructure from them. We found that those compounds containing both GPSub_17 and 2-hydroxyethyl-1-ketone substructure tended to have higher PPB fractions, and the trend was more significant when coexisting with another substructure (Figure S4). The results demonstrated that the binding affinity of molecules and plasma proteins could not be determined by a single substructure but by the results of interactions between multiple substructures within the molecule. Therefore, we derived the second-level chemical rules, i.e., ESub and RESub, whose presence may affect the function of PSub and induce PPB fractions to change in a specific direction.

3.3.2. Derivation of ESubs and RESubs for 22 GPSubs. In total, 225 ESubs and 312 RESubs were derived from the training set corresponding to 22 GPSubs except for GPSub_8. Each GPSub had at least 1 and at most 65 ESubs and RESubs with different functional groups. All these substructure pairs, namely ESub-GPSub pairs and RESub-GPSub pairs, and their relationships are illustrated in Tables S6 and S7. Generally, it was easier to derive second-level chemical rules when the compounds containing specific GPSubs were of high structural diversity. Notably, ESubs and RESubs were ordinary substructures and not necessarily highlighted by the iPPB model.

Taking GPSub_14 as an example, we finally extracted 23 ESubs and 15 RESubs. As shown in Figure 7, the substructure of propan-2-ol ("*CC(C)O") was identified as an ESub corresponding to GPSub_14. The compounds with this substructure pair showed significantly higher PPB fractions ($P < 0.05$) than those only containing GPSub_14. We thought

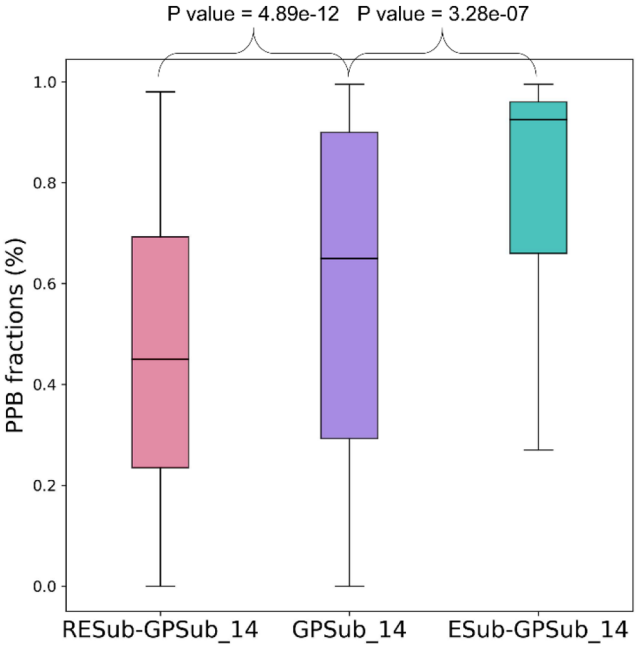


Figure 7. PPB distribution of compounds containing specific substructures. The pink box represents the compounds containing GPSub_14 and (methylamino) acetic acid (RESub). The purple box represents the compounds containing GPSub_14 and the green box represents compounds containing GPSub_14 and propan-2-ol (ESub).

that the propan-2-ol might increase the binding affinity by providing new hydrogen donors for compounds. Meanwhile, another substructure, (methylamino) acetic acid ("*C(NCC(=O)O)*"), was extracted and regarded as the RESub corresponding to GPSub_14. The box plot for two categories of compounds revealed that the presence of (methylamino) acetic acid in compounds may decrease the binding affinity to plasma proteins ($P < 0.05$) (Figure 7). It was easy to appreciate that the introduction of acetic acid could decrease the lipophilicity of compounds, thus decreasing the PPB fraction. On the basis of the results, we could presume that, given a novel chemical entity containing highlighted GPSub_14, we could improve the PPB fraction by introducing the propan-2-ol or decrease the PPB fraction through directly modifying GPSub_14 with (methylamino) acetic acid.

3.3.3. Analysis of Substructure Pairs. We calculated and recorded the frequency of topological distance for each substructure pair when they were present in compounds. It could be seen that there were certain rules in some substructure pairs, for example, "CN(C)C(=O)CNCC*" and "*C(NCC(=O)O)*" were always overlapped or adjacent when they appeared (topological distance ≤1), indicating that there may be a direct interaction between substructures.

According to the frequency of topological distance, we could establish schemes for GPSub modification: (1) to replace a portion or fully replace GPSub with ESub or RESub (topological distance ≤1 in most cases), (2) to change the local chemical environment of GPSub by adding another substructure, i.e., ESub or RESub, or remove ESub or RESub within molecules (topological distance >1 in most cases), and (3) otherwise both mentioned schemes could be considered.

Furthermore, S_{PPB} provided an intuitive digital to evaluate the different impacts of RESub/ESub on corresponding

https://doi.org/10.1021/acs.jcim.2c00297
J. Chem. Inf. Model. 2022, 62, 2788−2799

GPSub. For example, either "*CC═CC*" (S_{PPB} = −0.117) or "*CC(═O)O" (S_{PPB} = −0.239) could reduce the PPB fractions when coexisted with GPSub_14, but the effect of introducing the latter was much better as it had a lower S_{PPB}.

3.3.4. Evaluation of the Applicability for ESub and RESub. To evaluate the applicability of ESub and RESub, we analyzed the PPB change of each ESub-GPSub pair and RESub-GPSub pair with a one-sided Wilcoxon test through the mixed data set (i.e., the combination of validation set and test set). For a total set of 537 substructure pairs, there were 178 substructure pairs (nearly 33%), including 87 ESub-GPSub pairs and 91 RESub-GPSub pairs, which shared a similar magnitude and directionality of PPB change ($P < 0.05$) as in the training set. The results demonstrated that our derivation methods could indeed extract useful underlying chemical rules.

Nevertheless, there were still 359 substructure pairs that had not been verified, which could be attributed to the following two reasons: (1) the limitation of the validation set and test set. As mentioned in section 3.1.1, the number of compounds in the training set was 4-fold that of the combination of the validation set and test set. Such a large difference may result in some substructure pairs that cannot be detected in the mixed data set. (2) There were other substructures beyond the ESub and RESub, which could also have a great impact on GPSub simultaneously. In terms of ESub and RESub, we only considered the interactions between two substructures, but it was easy to speculate that new changes in binding affinity would occur when other substructures were added to the molecule.

4. DISCUSSION

In this study, we introduced a novel computational strategy, termed IDL-PPBopt, for the prediction and optimization of the PPB property. The strategy utilized interpretable deep learning techniques to develop a PPB prediction model and identify PPB-related substructure patterns (i.e., PSub). With these important substructures, we proposed the concept of substructure pairs, which were composed of a PSub and a second-level substructure (i.e., ESub or RESub). More importantly, in these substructure pairs, the second-level substructure provided a promising structural modification scheme for compounds with a corresponding PSub to obtain a favorable PPB property. Overall, the strategy identified the PPB-related substructures and then summarized a series of chemical rules into substructure pairs to guide lead optimization.

Compared with other methods such as QSAR models,[8,9,61] MMPA,[11,14,62] and scaffold hopping,[13] there are several advantages for IDL-PPBopt. The most significant advantage of IDL-PPBopt is that it could provide a series of specific structural modification schemes for lead compounds to obtain a better PPB property, whereas traditional QSAR models could not. For traditional QSAR models, one has to manually modify the structure without posterior knowledge and repeat the prediction until a favorable result was obtained.[63] By contrast, the utilization of interpretable deep learning techniques enables us to get insight into the model-learned knowledge of why makes such predictions, which helps to better understand the molecular mechanisms and rationally design the structural modification scheme.

Second, IDL-PPBopt has fewer data restrictions when compared with MMPA. Matched molecular pair refers to a pair of molecules that differ structurally at a single site, so ideally the experiment data set should be as large as possible to afford the data mining of structural transformation rules.[11,14] Different from MMPA, IDL-PPBopt explores the relationship between substructure combinations and PPB property. So even if in a small data set, this strategy still works because each substructure pair can be matched into molecules with different scaffolds. In addition, IDL-PPBopt can extract second-level chemical rules beyond matched molecular pairs.

Finally, IDL-PPBopt can identify important PPB-related atoms and substructures. As described in the case study (Figure 5C,D), both important functional groups and scaffolds can be detected, which provide clues for setting up structural modification schemes. Thus, IDL-PPBopt overcomes the limitation of scaffold hopping strategies and more specifically guides the optimization of the PPB property.

5. CONCLUSIONS

In the present study, we introduced a computational strategy named IDL-PPBopt for the prediction and optimization of the PPB property via an interpretable deep learning method. The strategy captured important substructure patterns for lead compounds and customized a series of unique structural modification schemes through deriving second-level chemical rules to obtain a favorable PPB property. Therefore, IDL-PPBopt provided an alternative for the optimization of PPB fractions of lead compounds and would be used in the optimization of other pharmacokinetic properties. Nevertheless, there is still room for improvement. For example, the present study focused on the interactions between two substructures, but given the complexity of chemical systems, the combination of more substructures should be considered in the future.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.2c00297.

Definition of the topological distance between two substructures; performance of coefficient of determination R^2 and RMSE during model training; detected privileged substructure and previously reported important structural patterns; PPB fractions of compounds containing GPSub_17, PPB fractions of compounds containing GPSub_17 and 2-hydroxyethyl-1-ketone, and PPB fractions of compounds containing GPSub_17, 2-hydroxyethyl-1-ketone, and azanethiol (PDF)

Equations of three statistical indexes; all chemical information on SMILES, PPB records, and data set classifications; all hyper-parameters of the iPPB model; model performance of five deep learning models; changes in the number of fragments during the GPSub screening process; information of all ESub-GPSub pairs; information of all RESub-GPSub pairs (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

Yun Tang − *Shanghai Frontiers Science Center of Optogenetic Techniques for Cell Metabolism, School of Pharmacy, East China University of Science and Technology, Shanghai 200237, China;* orcid.org/0000-0003-2340-1109; Email: ytang234@ecust.edu.cn

https://doi.org/10.1021/acs.jcim.2c00297
J. Chem. Inf. Model. 2022, 62, 2788−2799

Authors

Chaofeng Lou − *Shanghai Frontiers Science Center of Optogenetic Techniques for Cell Metabolism, School of Pharmacy, East China University of Science and Technology, Shanghai 200237, China*

Hongbin Yang − *Shanghai Frontiers Science Center of Optogenetic Techniques for Cell Metabolism, School of Pharmacy, East China University of Science and Technology, Shanghai 200237, China;* orcid.org/0000-0001-6740-1632

Jiye Wang − *Shanghai Frontiers Science Center of Optogenetic Techniques for Cell Metabolism, School of Pharmacy, East China University of Science and Technology, Shanghai 200237, China;* orcid.org/0000-0002-5568-1515

Mengting Huang − *Shanghai Frontiers Science Center of Optogenetic Techniques for Cell Metabolism, School of Pharmacy, East China University of Science and Technology, Shanghai 200237, China*

Weihua Li − *Shanghai Frontiers Science Center of Optogenetic Techniques for Cell Metabolism, School of Pharmacy, East China University of Science and Technology, Shanghai 200237, China;* orcid.org/0000-0001-7055-9836

Guixia Liu − *Shanghai Frontiers Science Center of Optogenetic Techniques for Cell Metabolism, School of Pharmacy, East China University of Science and Technology, Shanghai 200237, China;* orcid.org/0000-0001-9648-844X

Philip W. Lee − *Shanghai Frontiers Science Center of Optogenetic Techniques for Cell Metabolism, School of Pharmacy, East China University of Science and Technology, Shanghai 200237, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.2c00297

Notes

The authors declare no competing financial interest.

All data involved in this study are available in the Supporting Information. The commercial software platform Pipeline Pilot was purchased by the East China University of Science and Technology and licensed from BIOVIA (https://www.3ds.com/products-services/biovia/products/data-science/pipeline-pilot/). The free software including RDKit (http://www.rdkit.org), and SciPy (http://www.scipy.org) are freely available on their Web sites. The code of the IDL-PPBopt strategy is available at http://github.com/Louchaofeng/IDL-PPBopt.

■ ACKNOWLEDGMENTS

This work was supported by the National Key Research and Development Program of China (Grant 2019YFA0904800), the National Natural Science Foundation of China (Grants 81872800 and 82173746), and the 111 Project (Grant BP0719034).

■ REFERENCES

(1) Waring, M. J.; Arrowsmith, J.; Leach, A. R.; Leeson, P. D.; Mandrell, S.; Owen, R. M.; Pairaudeau, G.; Pennie, W. D.; Pickett, S. D.; Wang, J.; Wallace, O.; Weir, A. An analysis of the attrition of drug candidates from four major pharmaceutical companies. *Nat. Rev. Drug Discov* **2015**, *14*, 475−86.

(2) Smith, D. A.; van de Waterbeemd, H. Pharmacokinetics and metabolism in early drug discovery. *Curr. Opin. Chem. Biol.* **1999**, *3*, 373−378.

(3) Kola, I.; Landis, J. Can the pharmaceutical industry reduce attrition rates? *Nat. Rev. Drug Discov* **2004**, *3*, 711−5.

(4) Zhao, L.; Yin, W.; Sun, Y.; Sun, N.; Tian, L.; Zheng, Y.; Zhang, C.; Zhao, S.; Su, X.; Zhao, D.; Cheng, M. Improving the metabolic stability of antifungal compounds based on a scaffold hopping strategy: Design, synthesis, and structure-activity relationship studies of dihydrooxazole derivatives. *Eur. J. Med. Chem.* **2021**, *224*, 113715.

(5) Anderson, J. M.; Measom, N. D.; Murphy, J. A.; Poole, D. L. Bridge Functionalisation of Bicyclo[1.1.1]pentane Derivatives. *Angew. Chem., Int. Ed. Engl.* **2021**, *60*, 24754−24769.

(6) Yang, H.; Lou, C.; Li, W.; Liu, G.; Tang, Y. Computational Approaches to Identify Structural Alerts and Their Applications in Environmental Toxicology and Drug Discovery. *Chem. Res. Toxicol.* **2020**, *33*, 1312−1322.

(7) Price, E.; Kalvass, J. C.; DeGoey, D.; Hosmane, B.; Doktor, S.; Desino, K. Global Analysis of Models for Predicting Human Absorption: QSAR, In Vitro, and Preclinical Models. *J. Med. Chem.* **2021**, *64*, 9389−9403.

(8) Chen, J.; Yang, H.; Zhu, L.; Wu, Z.; Li, W.; Tang, Y.; Liu, G. In Silico Prediction of Human Renal Clearance of Compounds Using Quantitative Structure-Pharmacokinetic Relationship Models. *Chem. Res. Toxicol.* **2020**, *33*, 640−650.

(9) Sun, L.; Yang, H.; Li, J.; Wang, T.; Li, W.; Liu, G.; Tang, Y. In Silico Prediction of Compounds Binding to Human Plasma Proteins by QSAR Models. *ChemMedChem.* **2018**, *13*, 572−581.

(10) Pal, S.; Pogany, P.; Lumley, J. A. Molecule Ideation Using Matched Molecular Pairs. *Methods Mol. Biol.* **2022**, *2390*, 503−521.

(11) Fu, L.; Yang, Z. Y.; Yang, Z. J.; Yin, M. Z.; Lu, A. P.; Chen, X.; Liu, S.; Hou, T. J.; Cao, D. S. QSAR-assisted-MMPA to expand chemical transformation space for lead optimization. *Brief. Bioinform.* **2021**, *22*, bbaa374.

(12) Dalke, A.; Hert, J.; Kramer, C. mmpdb: An Open-Source Matched Molecular Pair Platform for Large Multiproperty Data Sets. *J. Chem. Inf Model* **2018**, *58*, 902−910.

(13) Yang, H.; Sun, L.; Wang, Z.; Li, W.; Liu, G.; Tang, Y. ADMETopt: A Web Server for ADMET Optimization in Drug Design via Scaffold Hopping. *J. Chem. Inf Model* **2018**, *58*, 2051−2056.

(14) Tyrchan, C.; Evertsson, E. Matched Molecular Pair Analysis in Short: Algorithms, Applications and Limitations. *Comput. Struct Biotechnol J.* **2017**, *15*, 86−90.

(15) Bohnert, T.; Gan, L. S. Plasma protein binding: from discovery to development. *J. Pharm. Sci.* **2013**, *102*, 2953−94.

(16) Otagiri, M. Study on binding of drug to serum protein. *Yakugaku zasshi: Journal of the Pharmaceutical Society of Japan* **2009**, *129*, 413−25.

(17) Smith, D. A.; Di, L.; Kerns, E. H. The effect of plasma protein binding on in vivo efficacy: misconceptions in drug discovery. *Nat. Rev. Drug Discov* **2010**, *9*, 929−39.

(18) Seyfinejad, B.; Ozkan, S. A.; Jouyban, A. Recent advances in the determination of unbound concentration and plasma protein binding of drugs: Analytical methods. *Talanta* **2021**, *225*, 122052.

(19) Levy, G. Effect of plasma protein binding on renal clearance of drugs. *J. Pharm. Sci.* **1980**, *69*, 482−3.

(20) Basant, N.; Gupta, S.; Singh, K. P. Predicting binding affinities of diverse pharmaceutical chemicals to human serum plasma proteins using QSPR modelling approaches. *SAR QSAR Environ. Res.* **2016**, *27*, 67−85.

(21) Schmidt, S.; Röck, K.; Sahre, M.; Burkhardt, O.; Brunner, M.; Lobmeyer, M. T.; Derendorf, H. Effect of protein binding on the pharmacological activity of highly bound antibiotics. *Antimicrob. Agents Chemother.* **2008**, *52*, 3994−4000.

(22) Wang, J.; Cao, D.; Tang, C.; Xu, L.; He, Q.; Yang, B.; Chen, X.; Sun, H.; Hou, T. DeepAtomicCharge: a new graph convolutional network-based architecture for accurate prediction of atomic charges. *Brief. Bioinform.* **2021**, *22*, bbaa183.

(23) Deng, D.; Chen, X.; Zhang, R.; Lei, Z.; Wang, X.; Zhou, F. XGraphBoost: Extracting Graph Neural Network-Based Features for a

Better Prediction of Molecular Properties. *J. Chem. Inf Model* **2021**, *61*, 2697−2705.
(24) Chen, J.; Cheong, H. H.; Siu, S. W. I. xDeep-AcPEP: Deep Learning Method for Anticancer Peptide Activity Prediction Based on Convolutional Neural Network and Multitask Learning. *J. Chem. Inf Model* **2021**, *61*, 3789−3803.
(25) Zhavoronkov, A.; Ivanenkov, Y. A.; Aliper, A.; Veselov, M. S.; Aladinskiy, V. A.; Aladinskaya, A. V.; Terentiev, V. A.; Polykovskiy, D. A.; Kuznetsov, M. D.; Asadulaev, A.; Volkov, Y.; Zholus, A.; Shayakhmetov, R. R.; Zhebrak, A.; Minaeva, L. I.; Zagribelnyy, B. A.; Lee, L. H.; Soll, R.; Madge, D.; Xing, L.; Guo, T.; Aspuru-Guzik, A. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nature biotechnology* **2019**, *37*, 1038−1040.
(26) Xiong, Z.; Wang, D.; Liu, X.; Zhong, F.; Wan, X.; Li, X.; Li, Z.; Luo, X.; Chen, K.; Jiang, H.; Zheng, M. Pushing the Boundaries of Molecular Representation for Drug Discovery with the Graph Attention Mechanism. *J. Med. Chem.* **2020**, *63*, 8749−8760.
(27) Mukherjee, A.; Su, A.; Rajan, K. Deep Learning Model for Identifying Critical Structural Motifs in Potential Endocrine Disruptors. *J. Chem. Inf Model* **2021**, *61*, 2187−2197.
(28) Rodriguez-Perez, R.; Bajorath, J. Interpretation of Compound Activity Predictions from Complex Machine Learning Models Using Local Approximations and Shapley Values. *J. Med. Chem.* **2020**, *63*, 8761−8777.
(29) Raunio, H. In silico toxicology - non-testing methods. *Front. Pharmacol.* **2011**, *2*, 33.
(30) Wu, Z.; Jiang, D.; Wang, J.; Hsieh, C. Y.; Cao, D.; Hou, T. Mining Toxicity Information from Large Amounts of Toxicity Data. *J. Med. Chem.* **2021**, *64*, 6924−6936.
(31) Gini, G.; Zanoli, F.; Gamba, A.; Raitano, G.; Benfenati, E. Could deep learning in neural networks improve the QSAR models? *SAR QSAR Environ. Res.* **2019**, *30*, 617−642.
(32) Lim, H.; Jung, Y. Delfos: deep learning model for prediction of solvation free energies in generic organic solvents. *Chem. Sci.* **2019**, *10*, 8306−8315.
(33) Jimenez-Luna, J.; Skalic, M.; Weskamp, N.; Schneider, G. Coloring Molecules with Explainable Artificial Intelligence for Preclinical Relevance Assessment. *J. Chem. Inf Model* **2021**, *61*, 1083−1094.
(34) Hung, C.; Gini, G. QSAR modeling without descriptors using graph convolutional neural networks: the case of mutagenicity prediction. *Mol. Divers* **2021**, *25*, 1283−1299.
(35) Kim, H.; Nam, H. hERG-Att: Self-attention-based deep neural network for predicting hERG blockers. *Comput. Biol. Chem.* **2020**, *87*, 107286.
(36) Wang, S.; Dong, G.; Sheng, C. Structural simplification: an efficient strategy in lead optimization. *Acta Pharm. Sin B* **2019**, *9*, 880−901.
(37) Votano, J. R.; Parham, M.; Hall, L. M.; Hall, L. H.; Kier, L. B.; Oloff, S.; Tropsha, A. QSAR modeling of human serum protein binding with several modeling techniques utilizing structure-information representation. *J. Med. Chem.* **2006**, *49*, 7169−81.
(38) Zhu, X. W.; Sedykh, A.; Zhu, H.; Liu, S. S.; Tropsha, A. The use of pseudo-equilibrium constant affords improved QSAR models of human plasma protein binding. *Pharm. Res.* **2013**, *30*, 1790−8.
(39) Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magariños, M. P.; Overington, J. P.; Papadatos, G.; Smit, I.; Leach, A. R. The ChEMBL database in 2017. *Nucleic Acids Res.* **2017**, *45*, D945−d954.
(40) Wishart, D. S.; Feunang, Y. D.; Guo, A. C.; Lo, E. J.; Marcu, A.; Grant, J. R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; Assempour, N.; Iynkkaran, I.; Liu, Y.; Maciejewski, A.; Gale, N.; Wilson, A.; Chin, L.; Cummings, R.; Le, D.; Pon, A.; Knox, C.; Wilson, M. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* **2018**, *46*, D1074−d1082.
(41) Yuan, Y.; Chang, S.; Zhang, Z.; Li, Z.; Li, S.; Xie, P.; Yau, W.-P.; Lin, H.; Cai, W.; Zhang, Y.; Xiang, X. A novel strategy for prediction of human plasma protein binding using machine learning techniques. *Chemometrics and Intelligent Laboratory Systems* **2020**, *199*, 103962.
(42) Ferrari, T.; Cattaneo, D.; Gini, G.; Golbamaki Bakhtyari, N.; Manganaro, A.; Benfenati, E. Automatic knowledge extraction from chemical structures: the case of mutagenicity prediction. *SAR QSAR Environ. Res.* **2013**, *24*, 365−83.
(43) Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579−2605.
(44) Yang, H.; Lou, C.; Sun, L.; Li, J.; Cai, Y.; Wang, Z.; Li, W.; Liu, G.; Tang, Y. admetSAR 2.0: web-service for prediction and optimization of chemical ADMET properties. *Bioinformatics* **2019**, *35*, 1067−1069.
(45) Xiong, G.; Wu, Z.; Yi, J.; Fu, L.; Yang, Z.; Hsieh, C.; Yin, M.; Zeng, X.; Wu, C.; Lu, A.; Chen, X.; Hou, T.; Cao, D. ADMETlab 2.0: an integrated online platform for accurate and comprehensive predictions of ADMET properties. *Nucleic Acids Res.* **2021**, *49*, W5−W14.
(46) Wang, N.-N.; Deng, Z.-K.; Huang, C.; Dong, J.; Zhu, M.-F.; Yao, Z.-J.; Chen, A. F.; Lu, A.-P.; Mi, Q.; Cao, D.-S. ADME properties evaluation in drug discovery: Prediction of plasma protein binding using NSGA-II combining PLS and consensus modeling. *Chemometrics and Intelligent Laboratory Systems* **2017**, *170*, 84−95.
(47) Tajimi, T.; Wakui, N.; Yanagisawa, K.; Yoshikawa, Y.; Ohue, M.; Akiyama, Y. Computational prediction of plasma protein binding of cyclic peptides from small molecule experimental data using sparse modeling techniques. *BMC Bioinformatics* **2018**, *19*, 527.
(48) Watanabe, R.; Esaki, T.; Kawashima, H.; Natsume-Kitatani, Y.; Nagao, C.; Ohashi, R.; Mizuguchi, K. Predicting Fraction Unbound in Human Plasma from Chemical Structure: Improved Accuracy in the Low Value Ranges. *Mol. Pharmaceutics* **2018**, *15*, 5302−5311.
(49) Sushko, I.; Salmina, E.; Potemkin, V. A.; Poda, G.; Tetko, I. V. ToxAlerts: a Web server of structural alerts for toxic chemicals and compounds with potential adverse reactions. *J. Chem. Inf Model* **2012**, *52*, 2310−6.
(50) Bains, W.; Basman, A.; White, C. HERG binding specificity and binding site structure: evidence from a fragment-based evolutionary computing SAR study. *Prog. Biophys. Mol. Biol.* **2004**, *86*, 205−33.
(51) Vilums, M.; Heuberger, J.; Heitman, L. H.; IJzerman, A. P. Indanes–Properties, Preparation, and Presence in Ligands for G Protein Coupled Receptors. *Med. Res. Rev.* **2015**, *35*, 1097−126.
(52) Zhang, X.; Zhao, P.; Wang, Z.; Xu, X.; Liu, G.; Tang, Y.; Li, W. In Silico Prediction of CYP2C8 Inhibition with Machine-Learning Methods. *Chem. Res. Toxicol.* **2021**, *34*, 1850−1859.
(53) Green, J.; Cao, J.; Bandarage, U. K.; Gao, H.; Court, J.; Marhefka, C.; Jacobs, M.; Taslimi, P.; Newsome, D.; Nakayama, T.; Shah, S.; Rodems, S. Design, Synthesis, and Structure-Activity Relationships of Pyridine-Based Rho Kinase (ROCK) Inhibitors. *J. Med. Chem.* **2015**, *58*, 5028−37.
(54) Mao, H.; Hajduk, P. J.; Craig, R.; Bell, R.; Borre, T.; Fesik, S. W. Rational design of diflunisal analogues with reduced affinity for human serum albumin. *J. Am. Chem. Soc.* **2001**, *123*, 10429−35.
(55) Izumi, T.; Kitagawa, T. Protein binding of quinolonecarboxylic acids. I. Cinoxacin, nalidixic acid and pipemidic acid. *Chemical & pharmaceutical bulletin* **1989**, *37*, 742−5.
(56) Raboisson, P.; Manthey, C. L.; Chaikin, M.; Lattanze, J.; Crysler, C.; Leonard, K.; Pan, W.; Tomczuk, B. E.; Marugan, J. J. Novel potent and selective alphavbeta3/alphavbeta5 integrin dual antagonists with reduced binding affinity for human serum albumin. *Eur. J. Med. Chem.* **2006**, *41*, 847−61.
(57) Miyagawa, M.; Akiyama, T.; Taoda, Y.; Takaya, K.; Takahashi-Kageyama, C.; Tomita, K.; Yasuo, K.; Hattori, K.; Shano, S.; Yoshida, R.; Shishido, T.; Yoshinaga, T.; Sato, A.; Kawai, M. Synthesis and SAR Study of Carbamoyl Pyridone Bicycle Derivatives as Potent Inhibitors of Influenza Cap-dependent Endonuclease. *J. Med. Chem.* **2019**, *62*, 8101−8114.
(58) Hajduk, P. J.; Mendoza, R.; Petros, A. M.; Huth, J. R.; Bures, M.; Fesik, S. W.; Martin, Y. C. Ligand binding to domain-3 of human serum albumin: a chemometric analysis. *J. Comput. Aided Mol. Des* **2003**, *17*, 93−102.

https://doi.org/10.1021/acs.jcim.2c00297
J. Chem. Inf. Model. 2022, 62, 2788−2799

(59) Yun, Y. E.; Tornero-Velez, R.; Purucker, S. T.; Chang, D. T.; Edginton, A. N. Evaluation of Quantitative Structure Property Relationship Algorithms for Predicting Plasma Protein Binding in Humans. *Comput. Toxicol* **2021**, *17*, 100142.
(60) Gardiner, P.; Cox, R. J.; Grime, K. Plasma Protein Binding as an Optimizable Parameter for Acidic Drugs. *Drug Metab. Dispos.* **2019**, *47*, 865−873.
(61) Liu, L.; Zhang, L.; Feng, H.; Li, S.; Liu, M.; Zhao, J.; Liu, H. Prediction of the Blood-Brain Barrier (BBB) Permeability of Chemicals Based on Machine-Learning and Ensemble Methods. *Chem. Res. Toxicol.* **2021**, *34*, 1456−1467.
(62) Dossetter, A. G.; Griffen, E. J.; Leach, A. G. Matched molecular pair analysis in drug discovery. *Drug Discov Today* **2013**, *18*, 724−31.
(63) Wang, T.; Wu, M. B.; Lin, J. P.; Yang, L. R. Quantitative structure-activity relationship: promising advances in drug discovery platforms. *Expert Opin Drug Discov* **2015**, *10*, 1283−300.

https://doi.org/10.1021/acs.jcim.2c00297
J. Chem. Inf. Model. 2022, 62, 2788−2799