THE JOURNAL OF PHYSICAL CHEMISTRY B
A JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/JPCB

Article

3D-RISM-AI: A Machine Learning Approach to Predict Protein−Ligand Binding Affinity Using 3D-RISM

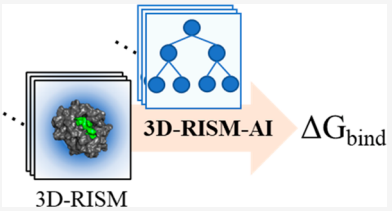
Kazu Osaki, Toru Ekimoto, Tsutomu Yamane, and Mitsunori Ikeguchi*

Cite This: *J. Phys. Chem. B* 2022, 126, 6148−6158

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Hydration free energy (HFE) is a key factor in improving protein−ligand binding free energy (BFE) prediction accuracy. The HFE itself can be calculated using the three-dimensional reference interaction model (3D-RISM); however, the BFE predictions solely evaluated using 3D-RISM are not correlated to the experimental BFE for abundant protein−ligand pairs. In this study, to predict the BFE for multiple sets of protein−ligand pairs, we propose a machine learning approach incorporating the HFEs obtained using 3D-RISM, termed 3D-RISM-AI. In the learning process, structural metrics, intra-/intermolecular energies, and HFEs obtained via 3D-RISM of ∼4000 complexes in the PDBbind database (ver. 2018) were used. The BFEs predicted using 3D-RISM-AI were well correlated to the experimental data (Pearson's correlation coefficient of 0.80 and root-mean-square error of 1.91 kcal/mol). As important factors for the prediction, the difference in the solvent accessible surface area between the bound and unbound structures and the hydration properties of the ligands were detected during the learning process.



INTRODUCTION

Developing accurate and efficient methods for predicting the binding affinity of ligands to target proteins is required in computer-aided drug discovery.[1] The binding affinity experimentally evaluated using the half-maximal inhibitory concentration (IC50) or the dissociation constant (K_d) can be converted to binding free energy (BFE).[2,3] Currently, BFE calculations based on atomic structures are widely performed in pharmaceutical processes daily. For example, in *in silico* screening processes, to rank many ligands in terms of affinity, BFEs of the ligands are calculated quickly. Thus, empirical scores of docking simulations (e.g., Glide[4]) are frequently used. After the screening process, ligand optimization processes are conducted to increase the ligand activity. Therefore, the BFEs of a few tens of ligands must be calculated accurately to examine the structure−activity relationships. Thus, free energy calculations based on molecular dynamics (MD) simulations are useful.[1] To improve the accuracy of BFE calculations, the accurate treatment of the hydration effects is a key factor because many water molecules are involved in the ligand-binding process. Upon ligand binding, water molecules in the binding site are replaced by the ligand and are forced to rearrange in the bound state.[5] The hydration water molecules around the ligand in the isolated state are dispelled upon ligand binding. In MD simulations using an explicit solvent, the hydration effects can be considered explicitly. However, MD simulations have high computational demands, leading to a loss in computational efficiency. Therefore, precisely handling the hydration effects is a trade-off between accuracy and efficiency.

Various computational methods for BFE calculation based on physicochemical approaches[1,4,6−22] using the thermodynamic cycle (Figure 1) or data-driven approaches using machine learning[23−31] have been proposed to date. The required computational burdens of physicochemical ap-

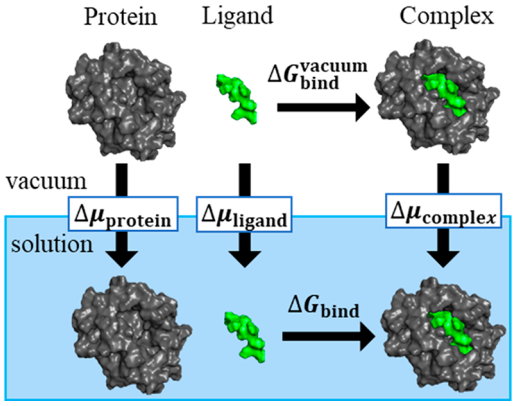
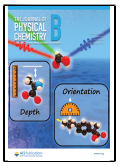


Figure 1. Thermodynamic cycle of the binding free energy. The protein (gray) and ligand (green) are shown under the upper and lower conditions representing vacuum and solution, respectively. The binding free energy (ΔG_{bind}) is obtained from the binding free energy in vacuum (ΔG^{vacuum}_{bind}) and the hydration free energies for the protein ($\Delta\mu_{protein}$), ligand ($\Delta\mu_{ligand}$), and complex ($\Delta\mu_{complex}$).

Received: May 16, 2022
Revised: July 27, 2022
Published: August 15, 2022



https://doi.org/10.1021/acs.jpcb.2c03384
J. Phys. Chem. B 2022, 126, 6148−6158

proaches differ depending on the incorporation methods of the hydration effects. In most physicochemical approaches, the interaction energy between a protein and a ligand can be calculated using force fields, whereas the hydration effects are handled differently. For example, in docking simulations with abundant protein−ligand pairs, the receptor structure is often fixed, and the hydration effect is implicitly and approximately evaluated using empirical score functions (e.g., Glide score[4]) to quickly evaluate BFEs. In molecular mechanics (MM) and Poisson−Boltzmann surface area (PB/SA) or MM and generalized Born surface area (GB/SA) methods, the interaction energy between a protein and a ligand and the configurational entropies are calculated using MM, where the hydration free energies (HFEs) of the protein, ligand, and complex are approximately calculated using a continuous dielectric model (i.e., implicit solvent model).[9,10] In continuous dielectric models, the detailed molecular features of water molecules such as hydrogen bonding, the hydrophobic effect, and the rearrangement of the water molecules upon solute insertion are missing. Recently, instead of continuum models, a solution statistical-mechanics theory-based method, called the three-dimensional reference interaction model (3D-RISM), was combined with MM.[11−16] The 3D-RISM method can evaluate HFEs, preserving the molecular features.[34−37] The 3D-RISM method provides a thermally averaged distribution of solvent molecules around solute molecules, and HFE can be calculated through an integral equation using the distribution function. The MM/3D-RISM method can successfully predict the BFE of a group of ligands with different activities for the same protein. The 3D-RISM method has also been applied to the statistical analysis of hydration water based on a large number of 3D-RISM calculations,[38] development of an efficient method to calculate HFEs,[39] or binding-site searches by extending the theory to the distribution of atoms constituting the ligand.[40] The 3D-RISM method, combined with MM or extended protocols described above, improves the prediction of HFEs.[11−16,38−40]

As a more rigorous approach, free energy calculations using all-atom MD simulations have been used to obtain the BFE between the bound and unbound states, such as the alchemical approach,[17] the potential of mean force (PMF)-based approach,[18] free energy perturbation (FEP)+,[19] generalized replica exchange with solute tempering (gREST)+FEP,[20] pmx,[21] and massively parallel computation of absolute BFE with well-equilibrated states (MP-CAFEE).[22] The free energy calculations have been applied to the relative or absolute BFE calculations, and they exhibit a good correlation with experimental BFEs. However, because of the high computational demands of MD simulations, free energy calculations for abundant protein−ligand pairs are practically difficult.

In contrast to the physicochemical approach described above, machine learning approaches predict BFEs by learning the correlation between experimental BFE data and input features, such as structural metrics and scoring functions.[23−31] Recently, a large amount of experimental BFE data and crystal structures of the protein−ligand complex have been stored in databases (e.g., the PDBbind database[32,33]), and machine learning models using the data for BFE prediction, for example, K_{DEEP},[24] XGB-Score,[25] and SFCscore^{RF}[26], have been proposed. As the input features, the descriptors given by structural metrics, such as atomic coordinates, distances between atoms, and amino acid sequences, or energetic metrics, such as scoring functions, were employed,[23] in which the relationships between the descriptors and the experimental BFE data were learned and a regression model was built. The selection of the appropriate features describing the experimental BFE data is one of the key points for accuracy.[23] Therefore, as descriptors, incorporating well-defined physicochemical quantities related to BFE, as well as conventional structural metrics, has the potential for accurate predictions of abundant protein−ligand pairs. Thus, incorporating hydration effects as an input feature may improve the accuracy of BFE predictions.

Herein, we propose a machine learning approach combined with the 3D-RISM method for BFE predictions. First, we calculated the BFEs for 3993 protein−ligand pairs in the PDBbind database using MM/3D-RISM (Figure 1). However, the BFEs evaluated using the 3D-RISM method exhibited a poor correlation with the experimental BFEs. Although the MM/3D-RISM strategy predicted the BFEs of a few ligands with a similar scaffold for the same protein, the strategy failed with multiple types of protein−ligand pairs. We also attempted to apply the improved version of 3D-RISM to BFE calculations. According to Palmer et al., the HFEs for 185 neutral small molecules calculated using the 3D-RISM method deviated from the experimental HFEs, and the difference was proportional to the partial molar volume.[41] Therefore, they proposed a universal correction, in which the contribution of the partial molar volume was simply subtracted from the HFE obtained using the 3D-RISM method. However, in our calculation of the BFEs, their correction did not improve the correlation between the calculated and experimental BFEs. Therefore, we developed a machine learning approach using thermodynamic quantities obtained from the 3D-RISM method as principal descriptors, termed 3D-RISM-AI. By introducing a machine learning algorithm for regression, we aimed to predict BFEs for abundant protein−ligand pairs, which cannot be expressed simply by energy addition and subtraction operations in the thermodynamic cycle. Regression models were constructed from the structural features and thermodynamic quantities calculated using 3D-RISM for the 3993 protein−ligand pairs in the PDBbind database. Four machine learning algorithms were used for the regression, and their performance was verified. The best-performing learning model exhibited a high correlation between the predicted and experimental BFEs: Pearson's correlation coefficient (R) of 0.80, Spearman's rank correlation coefficient (ρ) of 0.77, and root-mean- square error (RMSE) of 1.91 kcal/mol. Although the performance of 3D-RISM-AI is comparable to that of other machine learning models (R = 0.753−0.806, ρ = 0.647−0.796, and RMSE = 1.80−2.22 kcal/mol), the advantage of 3D-RISM-AI is that the feature importance analysis allows us to determine thermodynamic quantities that are effective in the BFE predictions.

THEORETICAL BACKGROUND

The BFE (ΔG_{bind}) between a protein and a ligand is defined as

$$\Delta G_{bind} = G_{complex} - (G_{protein} + G_{ligand}) \quad (1)$$

where $G_{complex}$, $G_{protein}$, and G_{ligand} are the free energies of the complex, protein, and ligand, respectively. The free energy is decomposed into the following three terms: the internal energy of the solute molecule (E), configurational entropy of the solute molecule (S), and HFE ($\Delta\mu$), as follows:

$$G_X = E_X - TS_X + \Delta\mu_X \quad (2)$$

https://doi.org/10.1021/acs.jpcb.2c03384
J. Phys. Chem. B 2022, 126, 6148−6158

where T is the temperature and X is one of the complexes, proteins, or ligands. According to the thermodynamic cycle of BFE (Figure 1), ΔG_{bind} can be obtained by the sum of the BFE in vacuum ($\Delta G_{\mathrm{bind}}^{\mathrm{vacuum}}$) and the difference in the HFE ($\Delta\Delta\mu$) as follows:

$$\Delta G_{\mathrm{bind}} = \Delta G_{\mathrm{bind}}^{\mathrm{vacuum}} + \Delta\Delta\mu \tag{3}$$

Here, we focus on one conformation of the solute complex (e.g., the crystal structure) and suppose that the conformations of both protein and ligand do not change from those in the complex, and the contributions of the entropy term can be neglected. Under this assumption, the terms on the right-hand side of eq 3 are

$$\Delta G_{\mathrm{bind}}^{\mathrm{vacuum}} \approx E_{\mathrm{complex}} - E_{\mathrm{protein}} - E_{\mathrm{ligand}} \tag{4}$$

$$\Delta\Delta\mu = \Delta\mu_{\mathrm{complex}} - \Delta\mu_{\mathrm{protein}} - \Delta\mu_{\mathrm{ligand}} \tag{5}$$

The internal energies in eq 4 can be calculated using a force field.

The HFE (eq 5) is obtained using the 3D-RISM method.[34−37] On the basis of the statistical solution theory, thermodynamic quantities are obtained through the pair distribution function, $g(r)$, which can be obtained by solving the Ornstein−Zernike (OZ) integral equation as a function of the total correlation function, $h(r) = g(r) - 1$, together with incorporating closure approximations. In the case of molecular liquids, the OZ equation includes degrees of freedom for both configuration and rotation, and it is difficult to solve such high-dimensional equations for complicated molecules. In contrast, using the RISM approximation, the molecules are described as a set of atom sites corresponding to atom types, and the molecular OZ equation can be approximately rewritten as the equation of site−site distance, which is called 1D-RISM.[42,43] Because the 1D-RISM approach uses the spherically symmetric site−site correlation function, the three-dimensional distribution of the solvent molecules around the solute molecule cannot be described. Therefore, an extension of RISM to three dimensions, called 3D-RISM, was introduced. In 3D-RISM, the total correlation functions of the solute and solvent sites are obtained using the 3D-RISM equation:

$$h_\alpha(\mathbf{r}) = \sum_\xi \int c_\xi(\mathbf{r}')\chi_{\xi\alpha}(|\mathbf{r} - \mathbf{r}'|)\, \mathrm{d}\mathbf{r}' \tag{6}$$

$$\chi_{\xi\alpha}(r) = \omega_{\xi\alpha}(r) + \rho_\alpha h_{\xi\alpha}(r) \tag{7}$$

where $h_\alpha(\mathbf{r})$ is the total correlation function of the solute site and the solvent site of the atom type for α, $c_\xi(\mathbf{r}')$ is the direct correlation function of the solvent atom type for ξ, and $\chi_{\xi\alpha}(|\mathbf{r}-\mathbf{r}'|)$ is the susceptibility function of the solvent sites in the bulk solvent given by eq 7. In eq 7, $\omega_{\xi\alpha}(r)$ is the intramolecular correlation function of the solvent molecule, ρ_α is the site-number density for atom type α in the bulk solvent, and $h_{\xi\alpha}(r)$ is the total correlation function of the intramolecular sites calculated from the 1D-RISM. To solve eq 6, the Kovalenko−Hirata (KH) closure, which is an approximate relationship between the total and direct correlation functions, is introduced.[37] The KH closure is given by

$$h_\alpha(\mathbf{r}) + 1 = \begin{cases} \exp(d_\alpha(\mathbf{r})) \text{ when } d_\alpha(\mathbf{r}) \le 0 \\ 1 + d_\alpha(\mathbf{r}) \text{ when } d_\alpha(\mathbf{r}) > 0 \end{cases} \tag{8}$$

$$d_\alpha(\mathbf{r}) = -\beta u_\alpha(\mathbf{r}) + h_\alpha(\mathbf{r}) - c_\alpha(\mathbf{r}) \tag{9}$$

where β is $1/k_{\mathrm{B}}T$, k_{B} is the Boltzmann constant, and $u_\alpha(\mathbf{r})$ is the solute−solvent site potential calculated from the force field. Using eqs 6−9, HFE is calculated as follows:

$$\Delta\mu = k_{\mathrm{B}}T \sum_\alpha \rho_\alpha \int \left[\frac{1}{2} h_\alpha{}^2(\mathbf{r})\Theta(-h_\alpha(\mathbf{r})) - c_\alpha(\mathbf{r}) - \frac{1}{2} h_\alpha(\mathbf{r}) c_\alpha(\mathbf{r}) \right] \mathrm{d}\mathbf{r} \tag{10}$$

where Θ is the Heaviside step function.

In the supervised machine learning, the relationship between objective and explanatory variables is expressed as a regression model.[23] Using the training data set comprising the experimental BFE values for the n protein−ligand pairs as the objective variable and the thermodynamic quantities and the structural indices calculated from each pair as the explanatory valuables, a regression model is given by

$$\Delta G_{\mathrm{bind\_exp}}^{(n)} = f(\mathbf{x}^{(n)}, \mathbf{y}^{(n)}, ..., \mathbf{z}^{(n)}; \theta) \tag{11}$$

where $\Delta G_{\mathrm{bind\_exp}}^{(n)}$ is the experimental BFE for the nth protein−ligand pair, $\{\mathbf{x}^{(n)}, \mathbf{y}^{(n)}, ..., \mathbf{z}^{(n)}\}$ represents the input feature vector comprising multiple descriptors calculated from the nth pair, θ is the hyperparameter, and the function f is determined by optimizing the objective function L as

$$\min_\theta L\{\left\| f(\mathbf{x}^{(n)}, \mathbf{y}^{(n)}, ..., \mathbf{z}^{(n)}; \theta) - \Delta G_{\mathrm{bind\_exp}}^{(n)} \right\|\} \tag{12}$$

Using the training regression model f, the BFE for the new $(n+1)$th pair, which is not included in the training data set, can be predicted as

$$\Delta G_{\mathrm{bind\_predict}}^{(n+1)} = f(\mathbf{x}^{(n+1)}, \mathbf{y}^{(n+1)}, ..., \mathbf{z}^{(n+1)}; \theta) \tag{13}$$

The settings of function f, the hyperparameters, and the objective function to be optimized depend on the regression algorithm. In this study, four types of regression algorithms were employed: ridge regression (RR),[44] support vector regression (SVR),[45] random forest regression (RFR),[46] and extreme gradient boosting regression (XGBR).[47] RR and SVR use a linear function with regularization term or the support vector and kernel trick, and RFR and XGBR use the decision tree for the regression process. The machine learning approach is expected to provide functions of BFEs that cannot be expressed by a simple linear operation of energies (eqs 1−5 and 10), as defined by the thermodynamic cycle, because the function given by the regression algorithm is represented by a nonlinear optimized function depending on the data set and explanatory variables used.

METHODS

Data Sets. The experimental BFE values and the crystal structures for the protein−ligand complex in the PDBbind database (ver. 2018) were used.[32,33] In the PDBbind database, from 16151, the refined set (ver. 2018, the number of complexes: n = 4463) was defined according to the quality of the crystal structures.[48] In addition, the core set (ver. 2016, n = 285) was selected from the refined set, 57 clusters were determined using amino acid sequence similarities, and five representative proteins were selected from each cluster. Crystal structures in the refined set containing unnatural amino acids or atomic collisions that could not be treated in the computational preprocess were excluded from the calculations. Finally, we used 3933 complexes in the refined set, including

https://doi.org/10.1021/acs.jpcb.2c03384
J. Phys. Chem. B 2022, 126, 6148−6158

217 complexes as the core set. In the machine learning process, 217 complexes for the core set were used as test data and 3716 complexes (the refined set except for the core set) were used as training data. The complexes used as training data did not contain complexes in the test data.

3D-RISM Calculations. Before the 3D-RISM calculations, structural modeling was performed on 3993 crystal structures. All crystal water molecules were removed. With multiple protein complexes, a protein located within 10 Å of the ligand was used. For missing residues, the acetyl cap and *N*-methyl cap were added to the missing N- and C-termini, respectively. The protonation states of histidine were determined using the ProtAssign module implemented in Maestro.[49] Energy minimizations for the modeling structures were performed using the GB solvent model implemented in Amber18 and AmberTools19.[50] The force fields for proteins and ligands were ff14SB and the generalized amber force field (GAFF), respectively.[51,52] The partial charge of the ligand was determined using the AM1-BCC method.[53,54] Two-step minimizations were performed: a 250-step steepest descent method and a 250-step conjugate gradient method. In the energy minimizations, position constraints with a force constant of 5 kcal/mol were added to the heavy atoms. The cutoff distance of the long-range interactions was set to 12 Å.

3D-RISM calculations were performed using the rism3d.snglpnt command in AmberTools19.[50] The complex structure after the energy minimizations was used as the input structure. For closure, the KH closure was used.[37] For 1D-RISM, the dielectrically consistent reference interaction site model (DRISM) theory was used.[36] The water model was the SPC/E model,[55] the water density was set to 0.999 g/cm^3, and the temperature was 298 K. The buffer distance between the solute molecule and the boundary in the calculation box was set to 20 Å, and the grid spacing on the three-dimensional grid was set to 0.5 × 0.5 × 0.5 Å^3. The other parameters were set to default values.

Descriptors Used in 3D-RISM-AI. Thirteen thermodynamic quantities described below were calculated from complex structures. Using the Amber force field and 3D-RISM method, the internal energy (E, eq 4) and HFE ($\Delta\mu$, eqs 5 and 10) were calculated for the complex, protein, and ligand, as well as BFE (eq 3). The protein and ligand structures were extracted from the complex structures. From the 3D-RISM calculations, thermodynamic quantities related to the HFE, such as the partial molar volume (V), enthalpy term (ε), and entropy term ($-TS$), were calculated. In addition, the $\Delta\mu$, V, ε, and TS terms were decomposed into polar and apolar contributions. Consequently, 13 thermodynamic quantities (E, $\Delta\mu$, $\Delta\mu_{apolar}$, $\Delta\mu_{polar}$, V, V_{apolar}, V_{polar}, $-TS$, $-TS_{apolar}$, $-TS_{polar}$, ε, ε_{apolar}, and ε_{polar}) were obtained.

The structural indices were also calculated. Because the BFEs are correlated to the solvent accessible surface area (SASA),[1] the SASA values for the complex, protein, and ligand were calculated using CppTraj in AmberTools19.[50] To incorporate the conformational entropy of the ligand, the number of rotatable bonds was calculated using Rdkit,[56] because the entropy is correlated to the rotatable bonds.[57] In addition, the difference in each quantity between the complex and the isolated protein or ligand was calculated as $X_{Bind} = X_{complex} - X_{protein} - X_{ligand}$. Hereinafter, the difference was denoted as "Bind" and the Bind descriptors were calculated for the 13 thermodynamic quantities and SASA.

In summary, 58 descriptors were used as input features in the 3D-RISM-AI. The descriptors comprised four types (i.e., complex, protein, ligand, and Bind) of the 13 thermodynamic quantities and SASA, the number of rotatable bonds of the ligand, and BFE. The overall procedure and summary of the descriptors are shown in Figure 2. Each descriptor value was standardized to a mean of zero and a variance of one. All training data are available on GitHub (https://github.com/IkeguchiLab/3D-RISM-AI).

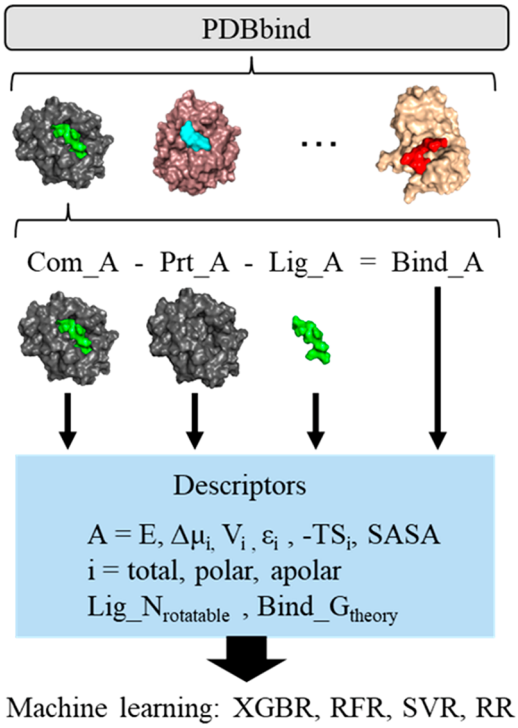


Figure 2. Overall procedure of 3D-RISM-AI. Quantities of the complex, protein, ligand, and binding are denoted as Com_A, Prt_A, Lig_A, and Bind_A, respectively. The descriptors denoted as A are the internal energy (E), HFE ($\Delta\mu$), partial molar volume (V), enthalpy term of the HFE (ε), entropy term of the HFE ($-TS$), and solvent accessible surface area (SASA). The index i represents the total value and the polar and apolar decomposed values. The number of rotatable bonds of the ligand is $Lig\_N_{rotatable}$. The BFE is $Bind\_G_{theory}$. The regression algorithms used are the extreme gradient boosting regression (XGBR), random forest regression (RFR), support vector regression (SVR), and ridge regression (RR).

Machine Learning. Four regression algorithms were used in supervised machine learning: RR, SVR, RFR, and XGBR. The models for RR, SVR, and RFR were built using scikit-learn,[58] and the model for XGBR was built using xgboost.[47] A Gaussian kernel is used in SVR. The hyperparameters and feature selection were optimized by 5-fold cross-validation using the training data set. A grid search is used in the hyperparameter search. The objective variables—the experimental BFE data—were calculated from K_d as $\Delta G = k_B T \ln K_d$. The optimization was evaluated using the averaged root-mean-squared error (RMSE) from the 5-fold cross-validation. After the optimization, a regression model with all training data (n = 3716) was built using a combination of the hyperparameters and the set of descriptors that exhibited the lowest mean RMSE.

https://doi.org/10.1021/acs.jpcb.2c03384
J. Phys. Chem. B 2022, 126, 6148−6158

Using the training regression model, BFEs were predicted for the test data set (n = 217). Compared to the experimental and predicted BFEs, the accuracy of the learning model was evaluated using Pearson's correlation coefficient (R), Spearman's rank correlation coefficient (ρ), and the RMSE. In the evaluation using RMSE, data with an absolute value of the difference between the experimental and predicted values within 2 kcal/mol were classified as a low absolute error (LAE), and the data with an error larger than 2 kcal/mol were classified as a high absolute error (HAE). In the decision tree-based algorithms (RFR and XGBR), the feature importance was evaluated using the information gain, a measure of improvement in the objective function when creating a branch in the decision tree, and the default measure was used in scikit-lean (RFR) and xgboost (XGBR).[47,58] All test data together with learning and prediction codes are available on GitHub (https://github.com/IkeguchiLab/3D-RISM-AI).

RESULTS AND DISCUSSION

Binding Free Energies Calculated Using 3D-RISM and the Thermodynamic Cycle. First, we calculated the BFEs for 3993 protein−ligand pairs using the 3D-RISM method and the thermodynamic cycle (Figure 1). The calculated BFEs exhibited a poor correlation with the experimental BFEs (Figure 3), with a Pearson's correlation coefficient (R) of

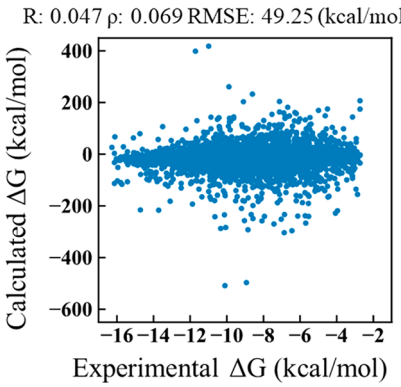
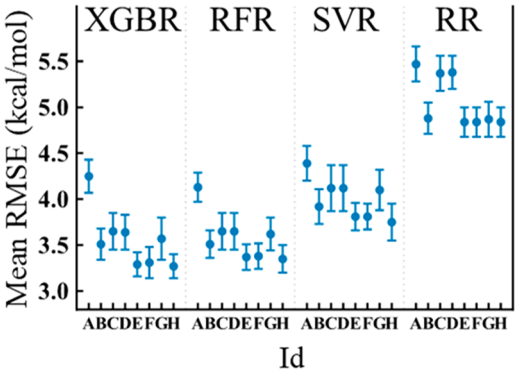


Figure 3. Correlation between the experimental and calculated BFEs using the 3D-RISM method and thermodynamic cycle for 3993 protein−ligand pairs. The Pearson's correlation coefficient (R), Spearman's rank correlation coefficient (ρ), and root-mean-squared error (RMSE) are 0.047, 0.069, and 49.25 kcal/mol, respectively.

0.047, Spearman's rank correlation coefficient (ρ) of 0.069, and RMSE of 49.25 kcal/mol. The calculated BFEs were one digit larger than the experimental BFEs. We also attempted an improved version of 3D-RISM in which the contribution of the partial molar volume was subtracted from the HFE obtained using the original 3D-RISM method.[41] However, the correlation between the calculated and experimental BFEs did not improve (Figure S1), with R = 0.049, ρ = 0.124, and RMSE = 187.68 kcal/mol.

Although the MM/3D-RISM strategy predicted the BFEs of a few ligands with a similar scaffold for the same protein, the strategy failed with multiple types of protein−ligand pairs. Therefore, we developed a machine learning method, termed 3D-RISM-AI, using thermodynamic quantities obtained by the 3D-RISM method as principal descriptors. Because the machine learning method can handle nonlinear relationships between an objective variable (BFEs) and explanatory variables (thermodynamic and structural features), it is possible to improve the BFE prediction using 3D-RISM.

Training Process. To build an optimal regression model, feature selection was performed using the training data set (n = 3716). The best combination of descriptor types was searched for in each regression algorithm (Figure 4). In this feature



Id	A	B	C	D	E	F	G	H
Bind	✓	✓	✓	✓	✓	✓	✓	✓
Ligand		✓			✓	✓		✓
Protein			✓		✓		✓	✓
Complex				✓		✓	✓	✓

Figure 4. Performance depends on features included in the regression algorithms. The feature combination patterns denoted as A to H are represented in the table.

selection, eight combinations of the descriptor types involved in Bind and the three remaining types (i.e., complex, protein, and ligand) were examined, and the mean RMSE was evaluated with 5-fold cross-validation (Table S1). For all algorithms, combining the descriptor types of the ligand, protein, and complex, besides the Bind, resulted in a lower mean RMSE, suggesting that the Bind-type descriptors alone were not effective for the BFE predictions.

Consequently, we selected the best combination exhibiting the lowest mean RMSE and employed all descriptor types (denoted as H in Figure 4) for XGBR, RFR, and SVR and the Bind, ligand, and complex descriptors (denoted as F in Figure 4) for the RR. Before examining the performance of the test data set, as shown in the next section, using all training data sets, a learning model was constructed for each algorithm with the optimized hyperparameters and the best combination of the feature types.

Performance of the 3D-RISM-AI. Using the test data set (n = 217), BFEs were predicted using each learning model (Figure 5). In contrast to the BFE prediction based solely on the 3D-RISM method (Figure 3), the BFEs predicted from all regression models in 3D-RISM-AI were well correlated to the experimental BFEs. In a comparison of the four algorithms, XGBR exhibited the best performance for all indicators: R = 0.80, ρ = 0.77, RMSE = 1.91 kcal/mol, and 154 data in LAE (Figure 5a). Here, the data in LAE represents the number of predicted BFEs within a 2 kcal/mol deviation from the experimental BFE. RFR, which is a decision tree-based method similar to XGBR, also showed a better performance (R = 0.78, ρ = 0.75, RMSE = 2.02 kcal/mol, and 141 data in LAE) (Figure 5b). In contrast, the performance of RR (R = 0.65, ρ = 0.66, RMSE = 2.36 kcal/mol, and 132 data points in LAE) and SVR (R = 0.68, ρ = 0.64, RMSE = 2.21 kcal/mol, and 150 data

https://doi.org/10.1021/acs.jpcb.2c03384
J. Phys. Chem. B 2022, 126, 6148−6158

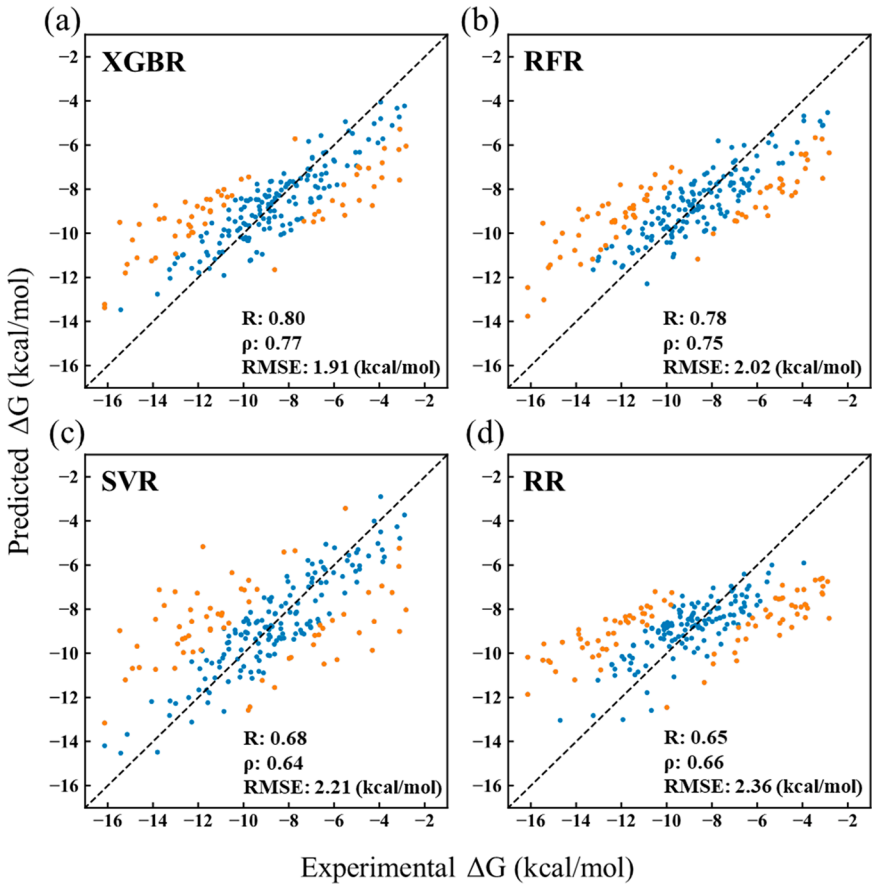


Figure 5. Comparison of the predicted BFEs using four learning models. Blue and orange dots represent the low absolute error (LAE) data and the high absolute error (HAE) data, respectively. LAE means that the difference between the predicted and experimental BFE is smaller than 2 kcal/mol. The difference for HAE is larger than 2 kcal/mol.

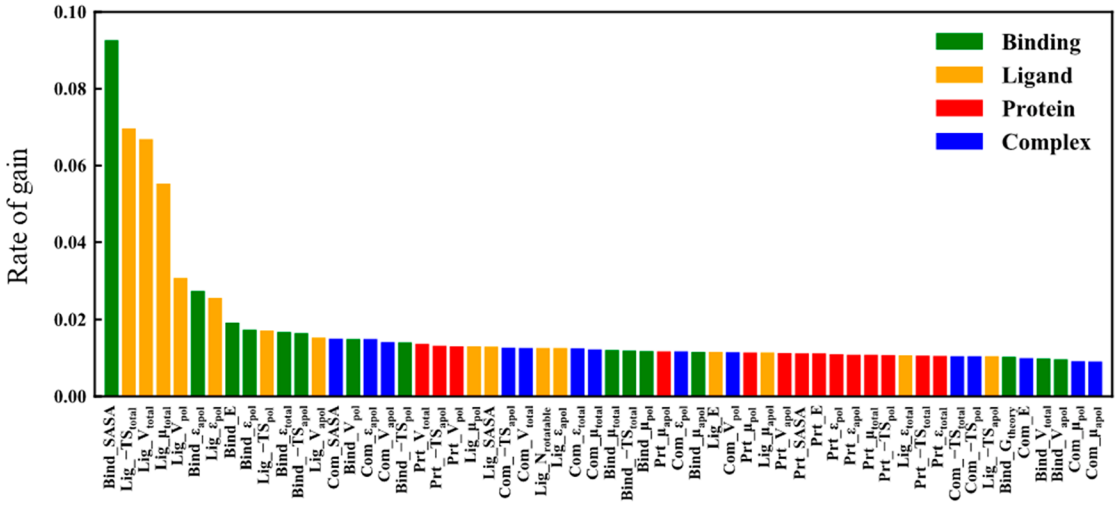


Figure 6. Rate of information gain in the training process of XGBR. The top descriptor is the difference in SASA upon ligand binding (Bind_SASA). The second to fifth descriptors are the HFE components of ligands: the entropy term (Lig_$-TS_{total}$), partial molar volume (Lig_V_{total}), HFE (Lig_μ_{total}), and polar component of the partial molar volume (Lig_V_{pol}). The sixth and seventh descriptors are the apolar component of the enthalpy (Lig_ε_{apol}) and the apolar component of the enthalpy difference in HFE upon ligand binding (Bind_ε_{apol}), respectively.

points in LAE) were relatively poor in this examination (Figure 5c,d).

Decision tree-type regressions, such as XGBR and RFR, worked well for predicting BFE. Because RR uses a linear function for the regression model, its capability to express the relationship between the experimental BFE and the input features may be poor. As for SVR, the mean RMSE in the training process for only Bind descriptors (denoted as A in Figure 4) showed similar values of XGBR or RFR; however, the mean RMSE of SVR did not decrease when other descriptors were included, which was a different behavior from the decision tree-type algorithms.

https://doi.org/10.1021/acs.jpcb.2c03384
J. Phys. Chem. B 2022, 126, 6148−6158

The performance of XGBR and RFR in 3D-RISM-AI was compared with that of other machine learning approaches. For example, the performance of the K_{DEEP} based on the three-dimensional convolutional neural network was R = 0.82, ρ = 0.82, and RMSE = 1.73 kcal/mol.[24] As for the decision tree type learning model with structure and interaction-energy features, such as RF-Score, ID-Score, SFCscore^{RF}, XGB-Score, and Δ_{vina}XGB, their performances were R = 0.753−0.806, ρ = 0.647−0.796, and RMSE = 1.80−2.22 kcal/mol for the best model for each approach.[25−29] The performance of the 3D-RISM-AI was comparable to that of previous machine learning models using different descriptors and algorithms. However, because 3D-RISM-AI was based on well-defined thermodynamic quantities, the factor analysis of the decision tree method allowed us to analyze descriptors that were important for predicting the BFE, as shown in the next section.

Feature Importance in the BFE Prediction. To understand descriptors that were effective for BFE prediction, the feature importance was analyzed using information gain in the XGBR learning model. The contribution to reducing the loss function was evaluated by determining the information gain of each descriptor. According to the ratio of the information gain of each descriptor to the total gain, the difference in SASA upon ligand binding (Bind_SASA) and the descriptors related to the HFE components of ligands (Lig_$-TS_{total}$, Lig_V_{total}, Lig_μ_{total}, Lig_V_{pol}, and Lig_ε_{apol}) were important in the learning process (Figure 6). Among the top seven descriptors with ratios above 0.02, besides Bind_SASA, only Bind_ε_{apol} was the Bind-type descriptor. In RFR, the most important descriptors were almost the same as the top important descriptors in XGBR. In particular, Bind_SASA and the HFE components of the ligands were assigned as important descriptors (Figure S2). These results suggest that, except for Bind_SASA, the descriptors related to the HFE components of the ligands calculated from 3D-RISM notably contributed to the BFE prediction.

To further understand the important physicochemical descriptors for prediction, the correlation among the descriptors was analyzed. The top seven descriptors of XGBR described above were classified into two groups exhibiting high correlations among the descriptors, which were hydrophobic and hydrophilic features (Figure 7). Although the descriptors related to hydrophobicity (Bind_SASA, Lig_$-TS_{total}$, and Lig_V_{total}) were the top three (Figure 6), the remaining important descriptors were hydrophilic features. Because both hydrophobic and hydrophilic features are important for BFE prediction, the physicochemical features of hydrophobicity and hydrophilicity should be learned in a well-balanced manner.

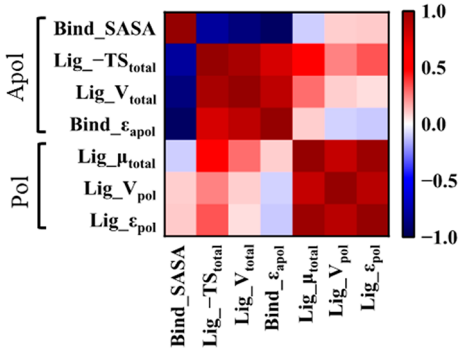


Figure 7. Correlation matrix of the top seven descriptors. The polar and apolar components, that is, hydrophilic and hydrophobic features, are denoted by Pol and Apol, respectively.

The correlations between the descriptors and experimental BFEs were compared with the machine learning results (Figure 8). The largest correlation coefficient of the individual descriptors with the experimental BFEs was less than 0.5. Considering that the correlation coefficients of the machine learning results were 0.6−0.8 (Figure 5), the machine learning algorithms could develop regression models showing higher correlations with the experimental BFEs. The top descriptors correlated to the experimental BFEs were similar to those that were important in the learning process of the decision tree (Figure 7). The top descriptors were Bind_SASA, Bind_ε_{apol}, and the ligand HFE-related descriptors. The top seven descriptors were hydrophobic features that correlated well with each other in the correlation matrix (Figure S3). Interestingly, the correlations between the hydrophilic features and the experimental BFEs were relatively small (no hydrophilic descriptor in the top seven), although their feature importance was not small (three in the top seven). These results indicate that in the machine learning algorithms the weights of the hydrophobic and hydrophilic descriptors were learned in a well-balanced manner.

Trends in Predicted BFEs by XGBR and Data Bias. To investigate data bias, the distributions of the predicted BFEs using XGBR and the experimental BFEs were compared (Figure 9). In a comparison of the distributions across all test data (Figure 9a), the predicted BFEs were biased toward the mean of BFEs about −9 kcal/mol. In the LAE data, the distributions were similar to each other (Figure 9b). By contrast, in the HAE data, the experimental BFEs were broadly distributed and the predicted BFEs were biased toward the mean (Figure 9c). This inaccuracy could be attributed to the experimental BFE data used in the training process were not uniformly distributed and biased toward the mean (Figure S4). Therefore, the amount of data largely deviating from the mean was small, and the construction of the regression model failed to predict the data deviating from the mean.

Finally, we should discuss future approaches to address the limitations in the above analyses and to improve accuracy by incorporating new features not considered in the 3D-RISM-AI. First, because the training data set in the PDBbind database was biased around the mean, more training data, including uniformly distributed BFEs over a wide energy range, are necessary. The BFE data for multiple types of ligands frequently used in the structure−activity relationship analyses would be effective. Second, the performance of 3D-RISM-AI depended on the performance of the 3D-RISM method and the force field used. Because the important features were the ligand HFEs and their components are given by the 3D-RISM method, solving the challenges inherent to the theoretical framework and numerical calculations, such as the closure and force field, would directly improve 3D-RISM-AI. In addition, the interactions of ligand molecules include weak interactions, such as cation−π, CH−π, and interactions related to halogen atoms, such as halogen bonding and sigma holes, which cannot be represented by conventional force fields but captured by quantum chemical (QM) calculations. Therefore, it would be a good strategy to improve the accuracy of the ligand force field or to add the interactions given by QM calculations as new descriptors. Third, the 3D-RISM-AI does not incorporate dynamic features. Both structural fluctuations in the solution

https://doi.org/10.1021/acs.jpcb.2c03384
J. Phys. Chem. B 2022, 126, 6148−6158

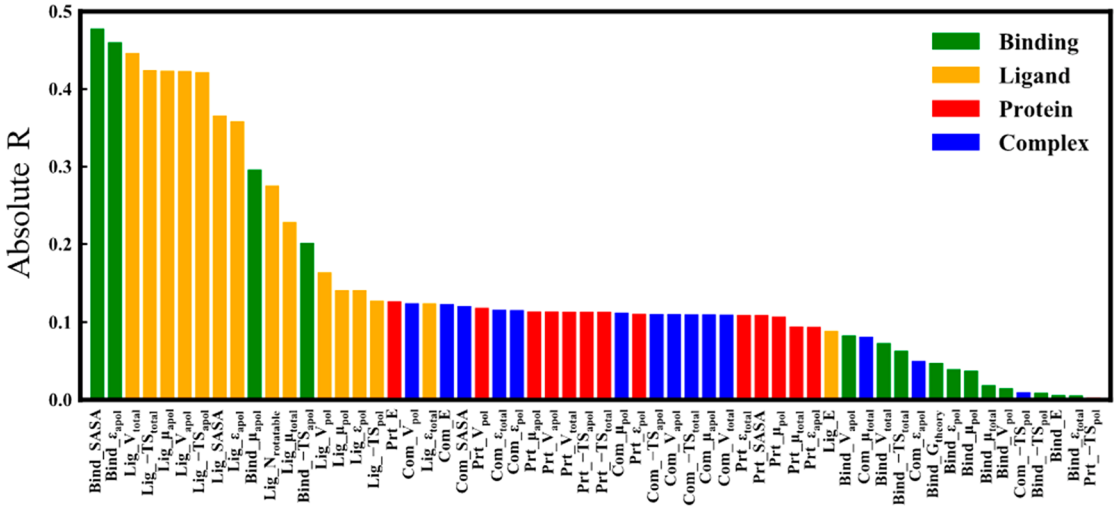


Figure 8. Correlations between the descriptors and experimental BFEs. The correlation is evaluated using the absolute value of Pearson's correlation coefficient (R).

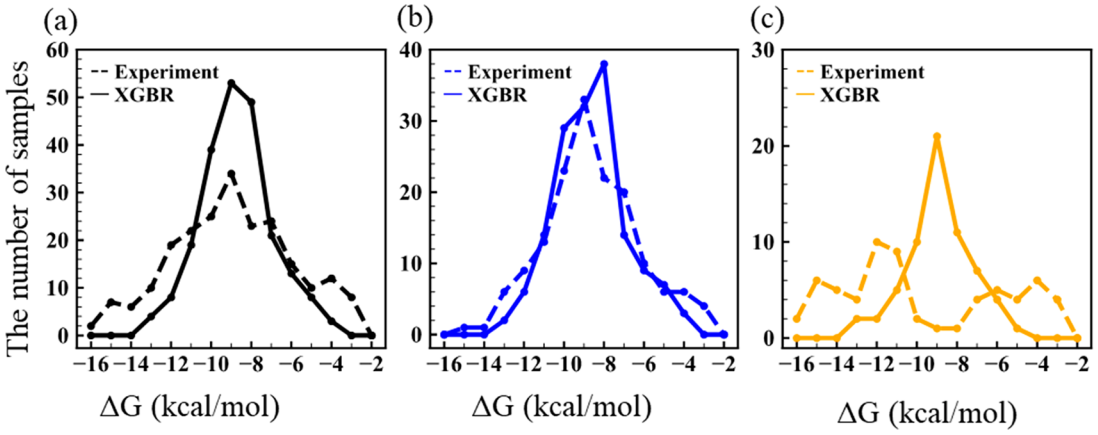


Figure 9. Histograms of BFEs. The distributions are (a) all test data sets, (b) data for the low absolute error (LAE) with $|\Delta G_{XGBR} - \Delta G_{exp})| \leq 2$ kcal/mol, and (c) data for the high absolute error (HAE) with $|\Delta G_{XGBR} - \Delta G_{exp})| > 2$ kcal/mol. The solid and dashed lines represent the predicted BFEs by XGBR and the experimental BFEs, respectively.

and the structural diversity of the unbound states leading to configurational entropy changes could be considered combined with MD simulations. Fourth, 3D-RISM-AI is complementary to other machine learning approaches, such as K_{DEEP}, which employs structural features as descriptors. The 3D-RISM-AI uses thermodynamic quantities based on HFE as descriptors and can incorporate structural information with different properties in a complementary manner.

Here, we discuss several issues to be resolved in the future before applying 3D-RISM-AI to *in silico* pharmaceutical processes, such as high-throughput virtual screening (HTVS) and structure−activity relationship analyses for abundant protein−ligand pairs. First, in these processes, complex models are obtained using docking simulations; however, the models usually deviate from experimental ones. As only experimental structures were used in the current 3D-RISM-AI, it is not clear how sensitive the machine learning model is to the quality of the structures of the protein−ligand complexes. Next, the computational time required to calculate the descriptors is an issue in HTVS. In this study, it required approximately 6 h using one central processing unit (CPU) core to conduct the 3D-RISM calculation for a protein or a complex, depending on the size of the protein: approximately 2 and 42 h were the minimum and maximum, respectively, and 6 h was the median. To perform the 3D-RISM calculations for abundant pairs in HTVS, which typically involves at least 100000s of molecules and often millions, a large number of CPU cores such as supercomputers are required. Alternatively, a graphics processing unit (GPU) version of the 3D-RISM method could be used for speeding up the 3D-RISM calculation.[59] Furthermore, the recently developed deep learning model for the 3D-RISM results,[60] which is capable of quickly predicting water distributions around proteins, could potentially work with 3D-RISM-AI. Finally, it is required to examine whether 3D-RISM-AI is a better approach to find hit compounds because it is not clear whether the machine learning model is significantly more accurate than the docking scores. Solving these issues and applying 3D-RISM-AI to pharmaceutical processes are future challenges.

CONCLUSIONS

Using the HFE based on the 3D-RISM method as the principal input feature, we proposed a machine learning approach to predict the BFE for abundant protein−ligand pairs, termed 3D-RISM-AI. Whereas the BFEs solely evaluated using the 3D-RISM method through the thermodynamic cycle were not

https://doi.org/10.1021/acs.jpcb.2c03384
J. Phys. Chem. B 2022, 126, 6148−6158

correlated to the experimental data, the BFEs predicted using 3D-RISM-AI showed a good correlation with the experimental BFEs: R = 0.80, ρ = 0.77, and RMSE = 1.91 kcal/mol. Although the performance was comparable to that of other machine learning approaches using other input features, the important factor analysis allowed us to understand the important features for predicting BFEs, such as the difference in the SASA between the unbound and bound structures and the ligand HFE-related descriptors. The physicochemical features described by the most important descriptors were hydrophobic, but the rest were hydrophilic, indicating that the balance between them is important for the predictions. Although the importance of both hydrophobic and hydrophilic interactions is well-known, the fine balance between them could be automatically detected using the machine learning approach of the physicochemical-based 3D-RISM-AI. In addition, the machine learning framework of 3D-RISM-AI can incorporate both structural diversity sampled from MD simulations and other structural or energetic input features.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpcb.2c03384.

Figure S1: correlation between the experimental and calculated BFEs using the 3D-RISM method with universal correction; Figure S2: rate of information gain in the training process of RFR; Figure S3: correlation matrix of the top seven descriptors; Figure S4: histogram of experimental BFEs in the training data set; Table S1: mean RMSE on cross-validation (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Mitsunori Ikeguchi − *Graduate School of Medical Life Science, Yokohama City University, Yokohama 230-0045, Japan; Center for Computational Science, RIKEN, Yokohama 230-0045, Japan;* orcid.org/0000-0003-3199-6931; Email: ike@tsurumi.yokohama-cu.ac.jp

Authors

Kazu Osaki − *Graduate School of Medical Life Science, Yokohama City University, Yokohama 230-0045, Japan*

Toru Ekimoto − *Graduate School of Medical Life Science, Yokohama City University, Yokohama 230-0045, Japan*

Tsutomu Yamane − *Center for Computational Science, RIKEN, Yokohama 230-0045, Japan*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpcb.2c03384

Author Contributions

M.I. designed the study. K.O., T.E., and T.Y. performed the calculations, K.O. built machine learning models, and all authors analyzed and discussed the data. T.E., K.O., and M.I. wrote the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Dr. Kei Terayama and Dr. Shoichi Ishida from Yokohama City University for their helpful discussions and comments. This work was financially supported by "Priority Issue on Post-K computer" (Building Innovative Drug Discovery Infrastructure Through Functional Control of Biomolecular Systems) (Project ID: hp150269, hp160223, hp170255, and hp180191) from the Ministry of Education, Culture, Sports, Science and Technology (MEXT); "Program for Promoting Researches on the Supercomputer Fugaku" (MD-driven Precision Medicine) (Project ID: hp200129 and hp210172) from MEXT; the Basis for Supporting Innovative Drug Discovery and Life Science Research (BINDS) (Project ID: JP21am0101109) and Research Support Project for Life Science and Drug Discovery (Project ID: JP22ama121023 to M.I.) from the Japan Agency for Medical Research and Development (AMED); a Grant-in-Aid for Scientific Research on Innovative Areas "Molecular Engine" (Grant No.: 18H05426 to M.I.) from MEXT; the RIKEN Dynamic Structural Biology Project; and the grant for 2021−2023 Strategic Research Promotion (No. SK202202 to M.I.) of Yokohama City University.

■ ABBREVIATIONS

BFE, binding free energy; HFE, hydration free energy; 3D-RISM, three-dimensional reference interaction model; MD, molecular dynamics; FEP, free energy perturbation; RR, ridge regression; SVR, support vector regression; RFR, random forest regression; XGBR, extreme gradient boosting regression; R, Pearson's correlation coefficient; ρ, Spearman's rank correlation coefficient; RMSE, root mean-squared error.

■ REFERENCES

(1) King, E.; Aitchison, E.; Li, H.; Luo, R. Recent Developments in Free Energy Calculations for Drug Discovery. *Front. Mol. Biosci.* **2021**, *8*, 712085.

(2) Testa, B.; Jenner, P.; Kilpatrick, G. J.; El Tayar, N.; Van de Waterbeemd, H.; Marsden, C. D. Do Thermodynamic Studies Provide Information on both the Binding to and the Activation of Dopoaminergic and Other Receptors? *Biochem. Pharmacol.* **1987**, *36* (23), 4041−4046.

(3) Raffa, R. B.; Porreca, F. Thermodynamic Analysis of the Drug-Receptor Interaction. *Life Sci.* **1989**, *44*, 245−258.

(4) Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T. A.; Sanschagrin, P. C.; Mainz, D. T. Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein-Ligand Complexes. *J. Med. Chem.* **2006**, *49*, 6177−6196.

(5) Ladbury, J. E. Just add water! The Effect of Water on the Specificity of Protein-Ligand Binding Sites and its Potential Application to Drug Design. *Chem. Biol.* **1996**, *3*, 973−980.

(6) Wang, L.; Wu, Y.; Deng, Y.; Kim, B.; Pierce, L.; Krilov, G.; Lupyan, D.; Robinson, S.; Dahlgren, M. K.; Greenwood, J.; et al. Accurate and Reliable Prediction of Relative Ligand Binding Potency in Prospective Drug Discovery by Way of a Modern Free-Energy Calculation Protocol and Force Field. *J. Am. Chem. Soc.* **2015**, *137*, 2695−2703.

(7) Gilson, M. K.; Given, J. A.; Bush, B. L.; McCammon, J. A. The Statistical-Thermodynamic Basis for Computation of Binding Affinities: A Critical Review. *Biophys. J.* **1997**, *72*, 1047−1069.

(8) Wang, L.; Berne, B. J.; Friesner, R. A. On Achieving High Accuracy and Reliability in the Calculation of Relative Protein-Ligand Binding Affinities. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 1937−1942.

(9) Rastelli, G.; Del Rio, A.; Degliesposti, G.; Sgobba, M. Fast and Accurate Predictions of Binding Free Energies Using MM-PBSA and MM-GBSA. *J. Comput. Chem.* **2010**, *31*, 797−810.

(10) Schaefer, M.; Karplus, M. A Comprehensive Analytical Treatment of Continuum Electrostatics. *J. Phys. Chem.* **1996**, *100*, 1578−1599.

(11) Gohda, K. An Attempt to Incorporate Effect of Direct Interaction between a Ligand and Explicit Water Molecules into MM/3D-RISM. *J. Chem. Biol. Drug Des.* **2018**, *92*, 1788−1800.
(12) Hasegawa, T.; Sugita, M.; Kikuchi, T.; Hirata, F. A Systematic Analysis of the Binding Affinity between the Pim-1 Kinase and Its Inhibitors Based on the MM/3D-RISM/KH Method. *J. Chem. Inf. Model.* **2017**, *57*, 2789−2798.
(13) Sugita, M.; Hirata, F. Predicting the Binding Free Energy of the Inclusion Process of 2-hydroxypropyl-β-cyclodextrin and Small Molecules by Means of the MM/3D-RISM method. *J. Phys.: Condens. Matter* **2016**, *28*, 384002−384012.
(14) Phanich, J.; Rungrotmongkol, T.; Sindhikara, D.; Phongphanphanee, S.; Yoshida, N.; Hirata, F.; Kungwan, N.; Hannongbua, S. A 3D-RISM/RISM Study of the Oseltamivir Binding Efficiency with the Wild-type and Resistance-associated Mutant Forms of the Viral Influenza B Neuraminidase. *Protein Sci.* **2016**, *25*, 147−158.
(15) Yesudas, J. P.; Blinov, N.; Dew, S. K.; Kovalenko, A. Calculation of Binding Free Energy of Short Double Stranded Oligonucleotides using MM/3D-RISM-KH Approach. *J. Mol. Liq.* **2015**, *201*, 68−76.
(16) Genheden, S.; Luchko, T.; Gusarov, S.; Kovalenko, A.; Ryde, U. An MM/3D-RISM Approach for Ligand Binding Affinities. *J. Phys. Chem. B* **2010**, *114*, 8505−8516.
(17) Deng, Y.; Roux, B. Computations of Standard Binding Free Energies with Molecular Dynamics Simulations. *J. Phys. Chem. B* **2009**, *113*, 2234−2246.
(18) Gumbart, J. C.; Roux, B.; Chipot, C. Standard Binding Free Energies from Computer Simlations: What Is the Best Strategy? *J. Chem. Theory Comput.* **2013**, *9*, 794−802.
(19) Abel, R.; Wang, L.; Harder, E. D.; Berne, B. J.; Friesner, R. A. Advancing Drug Discovery through Enhanced Free Energy Calculations. *Acc. Chem. Res.* **2017**, *50*, 1625−1632.
(20) Oshima, H.; Re, S.; Sugita, Y. Prediction of Protein-Ligand Binding Pose and Affinity Using the gREST+FEP Method. *J. Chem. Inf. Model.* **2020**, *60*, 5382−5394.
(21) Gapsys, V.; Michielssens, S.; Seeliger, D.; de Groot, B. L. pmx: Automated Protein Structure and Topology Generation for Alchemical Perturbations. *J. Comput. Chem.* **2015**, *36*, 348−354.
(22) Fujitani, H.; Tanida, Y.; Matsuura, A. Massively Parallel Computation of Absolute Binding Free Energy with Well-Equilibrated States. *Phys. Rev. E* **2009**, *79*, 021914.
(23) Wang, D. D.; Zhu, M.; Yan, H. Computationally Predicting Binding Affinity in Protein-Ligand Complexes: Free Energy-based Simulations and Machine Learning-based Scoring Functions. *Brief. Bioinform.* **2021**, *22*, bbaa107.
(24) Jimenez, J.; Skalic, M.; Martinez-Rosell, G.; De Fabritiis, G. K_{DEEP}: Protein-Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks. *J. Chem. Inf. Model.* **2018**, *58*, 287−296.
(25) Li, H.; Peng, J.; Sidorov, P.; Leung, Y.; Leung, K.-S.; Wong, M.-H.; Lu, G.; Ballester, P. J Classical Scoring Functions for Docking are Unable to Exploit Large Volumes of Structural and Interaction Data. *Bioinformatics* **2019**, *35*, 3989−3995.
(26) Zilian, D.; Sotriffer, C. A. $SFCscore^{RF}$: A Random Forest-Based Scoring Function for Improved Affinity Prediction of Protein-Ligand Complexes. *J. Chem. Inf. Model.* **2013**, *53*, 1923−1933.
(27) Lu, J.; Hou, X.; Wang, C.; Zhang, Y. Incorporating Explicit Water Molecules and Ligand Conformation Stability in Machine-Learning Scoring Functions. *J. Chem. Inf. Model.* **2019**, *59*, 4540−4549.
(28) Ballester, P. J.; Mitchell, J. B. A Machine Learning Approach to Predicting Protein-Ligand Binding Affinity with Applications to Molecular Docking. *Bioinformatics.* **2010**, *26*, 1169−1175.
(29) Li, G. B.; Yang, L. L.; Wang, W. J.; Li, L. L.; Yang, S. Y. ID-Score: A New Empirical Scoring Function Based on a Comprehensive Set of Descriptors Related to Protein-Ligand Interactions. *J. Chem. Inf. Model.* **2013**, *53*, 592−600.
(30) Francoeur, P. G.; Masuda, T.; Sunseri, J.; Jia, A.; Iovanisci, R. B.; Snyder, I.; Koes, D. R. 3D Convolutional Neural Networks and a Cross-Docked Dataset for Structure-Based Drug Design. *J. Chem. Inf. Model.* **2020**, *60*, 4200−4215.
(31) Tsubaki, M.; Tomii, K.; Sese, J. Compound-Protein Interaction Prediction with End-to-end Learning of Neural Networks for Graphs and Sequences. *Bioinformatics.* **2019**, *35*, 309−318.
(32) Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind Database: Collection of Binding Affinities for Protein-Ligand Complexes with Known Three-dimensional Structures. *J. Med. Chem.* **2004**, *47*, 2977−2980.
(33) Wang, R.; Fang, X.; Lu, Y.; Yang, C. Y.; Wang, S. The PDBbind Database: Methodologies and Updates. *J. Med. Chem.* **2005**, *48*, 4111−4119.
(34) Beglov, D.; Roux, B. Solvation of Complex Molecules in A Polar Liquid: An Integral Equation Theory. *J. Chem. Phys.* **1996**, *104*, 8678−8689.
(35) Beglov, D.; Roux, B. An Integral Equation to Describe the Solvation of Polar Molecules in Liquid Water. *J. Phys. Chem. B* **1997**, *101*, 7821−7826.
(36) Ratkova, E. L.; Palmer, D. S.; Fedorov, M. V. Solvation Thermodynamics of Organic Molecules by the Molecular Integral Equation Theory: Approaching Chemical Accuracy. *Chem. Rev.* **2015**, *115*, 6312−6356.
(37) Kovalenko, A.; Hirata, F. Self-Consistent Description of a Metal-Water Interface by the Kohn-Sham Density Functional Theory and the Three-Dimensional Reference Interaction Site Model. *J. Chem. Phys.* **1999**, *110*, 10095−10112.
(38) Yoshidome, T.; Ikeguchi, M.; Ohta, M. Comprehensive 3D-RISM Analysis of the Hydration of Small Molecule Binding Sites in Ligand-free Protein Structures. *J. Comput. Chem.* **2020**, *41*, 2406−2419.
(39) Hikiri, S.; Hayashi, T.; Inoue, M.; Ekimoto, T.; Ikeguchi, M.; Kinoshita, M. An Accurate and Rapid Method for Calculating Hydration Free Energies of A Variety of Solutes Including Proteins. *J. Chem. Phys.* **2019**, *150*, 175101−175113.
(40) Sugita, M.; Hamano, M.; Kasahara, K.; Kikuchi, T.; Hirata, F. New Protocol for Predicting the Ligand-Binding Site and Mode Based on the 3D-RISM/KH Theory. *J. Chem. Theory Comput.* **2020**, *16*, 2864−2876.
(41) Palmer, D. S.; Frolov, A. I.; Ratkova, E. L.; Fedorov, M. V. Towards A Universal Method for Calculating Hydration Free Energies: A 3D Reference Interaction Site Model with Partial Molar Volume Correction. *J. Phys.: Condens. Matter* **2010**, *22*, 492101−492110.
(42) Hirata, F.; Pettitt, B. M.; Rossky, P. J. Application of an Extended RISM Equation to Dipolar and Quadrupolar Fluids. *J. Chem. Phys.* **1982**, *77* (1), 509−520.
(43) Perkyns, J. S.; Pettitt, B. M. A Dielectrically Consistent Interaction Site Theory for Solvent-Electrolyte Mixtures. *Chem. Phys. Lett.* **1992**, *190*, 626−630.
(44) Hoerl, A.; Kennard, R. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics.* **2000**, *42*, 80−86.
(45) Cortes, C.; Vapnik, V. Support-Vector Networks. *Machine Learning* **1995**, *20*, 273−297.
(46) Breiman, L. Random Forests. *Machine Learning.* **2001**, *45*, 5−32.
(47) Chen, T.; Guestrin, C. Xgboost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, 2016; pp 785−794.
(48) Su, M.; Yang, Q.; Du, Y.; Feng, G.; Liu, Z.; Li, Y.; Wang, R. Comparative Assessment of Scoring Functions: The CASF-2016 Update. *J. Chem. Inf. Model.* **2019**, *59*, 895−913.
(49) Schrödinger Release 2019-1: *Protein Preparation Wizard; Epik*; Schrödinger, LLC: New York, 2019.
(50) Case, D. A.; Ben-Shalom, I. Y.; Brozell, S. R.; Cerutti, D. S.; Cheatham, III, T. E.; Cruzeiro, V. W.D.; Darden, T. A.; Duke, R. E.; Ghoreishi, D.; Giambasu, G.; et al. *AMBER 2019*; University of California: San Francisco, 2019.

https://doi.org/10.1021/acs.jpcb.2c03384
J. Phys. Chem. B 2022, 126, 6148−6158

(51) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **2015**, *11*, 3696−3713.
(52) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of A General Amber Force Field. *J. Comput. Chem.* **2004**, *25*, 1157−1174.
(53) Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I. Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: I. Method. *J. Comput. Chem.* **2000**, *21*, 132−146.
(54) Jakalian, A.; Jack, D. B.; Bayly, C. I. Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: II. Parameterization and Validation. *J. Comput. Chem.* **2002**, *23*, 1623−1641.
(55) Berendsen, H.; Grigera, J. R.; Straatsma, T. P. The Missing Term in Effective Pair Potentials. *J. Phys. Chem.* **1987**, *91*, 6269−6271.
(56) RDKit: Open-source cheminformatics software, ver. 2019.03.3; http://www.rdkit.org.
(57) Fukunishi, Y.; Nakamura, H. Improved Estimation of Protein-Ligand Binding Free Energy by Using the Ligand-Entropy and Mobility of Water Molecules. *Pharmaceuticals (Basel).* **2013**, *6*, 604−622.
(58) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *JMLR.* **2011**, *12*, 2825−2830.
(59) Maruyama, Y.; Hirata, F. Modified Anderson Method for Accelerating 3D-RISM Calculations using Graphics Processing Unit. *J. Chem. Theory Comput.* **2012**, *8*, 3015−3021.
(60) Kawama, K.; Fukushima, Y.; Ikeguchi, M.; Ohta, M.; Yoshidome, T. gr Predictor: a Deep-Learning Model for Predicting the Hydration Structures around Proteins. *bioRxiv* **2022**, 488616.

https://doi.org/10.1021/acs.jpcb.2c03384
J. Phys. Chem. B 2022, 126, 6148−6158