gr Predictor: a Deep-Learning Model for Predicting the Hydration Structures around Proteins

Kosuke Kawama[†], *Yusaku Fukushima*[†], *Mitsunori Ikeguchi*[‡,§], *Masateru Ohta*[‡], *and Takashi Yoshidome*[*†]

[†]*Department of Applied Physics, Graduate School of Engineering, Tohoku University, Sendai 980-8579, Japan*

[‡]*AI-driven Drug Discovery Collaborative Unit, HPC- and AI-driven Drug Development Platform Division, Center for Computational Science, RIKEN, 1-7-29, Suehiro-cho, Tsurumi-ku, Yokohama 230-0045, Japan*

[§]*Graduate School of Medical Life Science, Yokohama City University, 1-7-29, Suehiro-cho, Tsurumi-ku, Yokohama 230-0045, Japan*

ABSTRACT. Among the factors affecting biological processes such as protein folding and ligand binding, hydration, which is represented by a three-dimensional water-site-distribution-function around the protein, is crucial. The typical methods for computing the distribution functions, including molecular dynamics simulations and the three-dimensional reference interaction site model (3D-RISM) theory, require a long computation time from hours to tens of

hours. Here, we propose a deep-learning model rapidly estimating the distribution functions around proteins obtained by the 3D-RISM theory from the protein 3D structure. The distribution functions predicted using our deep-learning model are in good agreement with those obtained by the 3D-RISM theory. Particularly, the coefficient of determination between the distribution function obtained by the deep-learning model and that obtained using the 3D-RISM theory is approximately 0.98. Furthermore, using a graphics processing unit (GPU), the calculation by the deep learning model is completed in less than one minute, more than 2 orders of magnitude faster than the calculation time of 3D-RISM theory. Therefore, our deep learning model provides a practical and efficient way to calculate the three-dimensional water-site-distribution-functions. The program called "gr Predictor" is available under the GNU General Public License from https://github.com/YoshidomeGroup-Hydration/gr-predictor.

1. INTRODUCTION

Protein hydration is one of the factors governing the biophysical processes involving the protein, including folding and ligand binding[1,2]. Particularly, hydration strongly affects the stability of native structure and denaturation of proteins, whereas the ligand-protein complex is often stabilized by water-mediated interactions between the ligand and protein. Thus, elucidating the protein hydration properties is crucial to understand the biophysical processes and perform structure-based drug design in consideration of water molecules.

Hydration of protein is characterized by the three-dimensional water-site-distribution-functions around the protein. The distribution functions can be obtained using molecular dynamics (MD)

simulations and the three-dimensional reference-interaction site model (3D-RISM) theory[3]. MD simulations exactly compute the three-dimensional water site distribution functions, whereas the 3D-RISM theory, which is a statistical mechanical theory of solvation, approximately computes the distribution functions with the force fields employed by the MD simulations. The usefulness of the distribution functions has been demonstrated. As an example, the performance of the deep-learning (DL) model for the pose prediction improved upon the incorporation of the three-dimensional water-site-distribution-functions, obtained with MD simulations, inside the ligand-bind pocket[4]. The distribution function obtained using the 3D-RISM theory has also been widely discussed. The position of the crystal waters inside the cavity of hen egg-white lysozyme is difficult to compute via MD simulation, because the movement of a water molecule from outside the protein towards inside is hardly attained in a reasonable simulation time; conversely, the three-dimensional water-site-distribution-functions obtained with the 3D-RISM theory successfully reproduced the crystal-water positions[5]. Furthermore, the partial molar volume (PMV) that can be computed with the Kirkwood-Buff solution theory combined with the three-dimensional water-site-distribution-functions[6] exhibited a perfect agreement with the experimental data of several proteins using the distribution functions obtained with the 3D-RISM theory[7,8].

The advantages of the 3D-RISM theory over the MD simulations include the shorter computation time for obtaining the three-dimensional water-site-distribution-functions. Exploiting the power of a supercomputer and the advantage of the shorter computation time of the 3D-RISM theory, we have statistically analyzed the hydration states of 3,706 static crystallographic structures of a protein[9]. However, the investigation of amino acid mutations, ligand binding processes, and protein conformational changes require more than hundred

thousand of protein structures. To this aim, the 3D-RISM theory is inadequate, because the calculation of the hydration state of a protein requires a few hours with a single central processing unit (CPU). Thus, a new method drastically reducing the computation time of the 3D-RISM theory should be developed.

In the present paper, we propose a DL model for predicting the three-dimensional water-site-distribution-functions around the proteins obtained with the 3D-RISM theory. The data used for training the DL model, network architecture, prediction accuracy, and computation time of the DL model are described. Finally, a comparison of our DL model with other methodologies for obtaining the hydration states around proteins is discussed. Because our DL model accurately reproduced the distribution functions obtained with the 3D-RISM theory with a computation time of less than one minute using a single graphics processing unit (GPU), our DL model enables us to investigate amino acid mutations, ligand binding processes, and protein conformational changes.

2. MATERIALS AND METHODS

Proteins used for the computation. Twenty-seven proteins were selected from the proteins used in our previous study[9] considering 3,706 proteins taken from the protein-ligand complexes deposited in the PDBbind refined set (v. 2017)[11-16]. The selection for this study was performed as follows. From the initial 3,706 proteins, only the proteins without ions were considered, to reduce the number of atom types in the deep-learning model. This led to the selection of 2,718 proteins summarized in "Data2718-SI-Forsubmit.xlsx". Afterwards, the twenty-seven proteins shown in Table 1 were randomly selected. Among the twenty-seven proteins, twenty-two

proteins were used for training, and the remaining five proteins were used for the test. The following preprocesses were conducted to the 3,706 proteins in the previous study[9]: the ligand and the crystal waters were removed, and the chain closest to the ligand was employed for the protein structure with multiple chains.

The similarities of the sequences between the twenty-seven proteins are discussed in text S1 (Supporting Information). None of the proteins used in the test had 90% sequence similarity to the twenty-two proteins used in the training. The effects of the random selection of the twenty-seven proteins are discussed in the subsection "Discussion of the selection of twenty-seven proteins".

3D-RISM theory. In our previous study[9], the 3D-RISM theory was applied to obtain the distribution function at the position $\boldsymbol{r}$ for the water site α = H (hydrogen) or O (oxygen), denoted by $g_\alpha(\boldsymbol{r})$ hereinafter. In the present study, $g_\alpha(\boldsymbol{r})$ was used as a target variable for the construction of the DL model. In the following, the force fields and parameters employed in the previous study[9] are described. The distribution functions were obtained using the Amber ff99SB force-fields[17] for the proteins, whereas the coincident SPC/E model[18] was employed for the water molecule. The values of the dielectric constant, bulk density, and temperature were 78.497, 0.03332 Å^{-3}, and 310 K, respectively. In the computation using the 3D-RISM theory, a water box surrounding the protein was prepared so that the minimum distance between the protein and the edge of the box was 14 Å. The linear grid spacing of 0.5 Å was set for the x, y, and z coordinates.

Input and output formats for the deep-leaning model. To input a protein structure into our DL model, the protein structure was converted into the voxel format schematically illustrated in Fig. 1. First, the protein was decomposed into five atom types composed of carbon, nitrogen, oxygen,

sulfur, and hydrogen. Afterwards, the box surrounding the protein with the same size as that of the water box used for the 3D-RISM theory was prepared. The grid size of the voxel and the position of the protein were also the same as those used for the computation using the 3D-RISM theory. Then, the contribution of the *i*-th atom of atom type *j* to *k*-th voxel, $n(k,i,j)$, was computed in accordance with Eq. (1)[19]:

$$n(k,i,j) = 1 - \exp\left[-\left(\frac{\sigma_{\mathrm{vdw},j}}{r_{ik}}\right)^{12}\right] \quad (1)$$

where $\sigma_{\mathrm{vdw},j}$ is the van-der-Waals-radius of the atom type *j*, and r_{ik} is the distance between the *i*-th atom and the position of the *k*-th voxel. The value of $\sigma_{\mathrm{vdw},j}$ for each atom type is summarized in Table S1 in the Supporting Information. Finally, the contribution of the atom type *j* to *k*-th voxel, $n(k,i,j)$, $N(k,j)$, was computed according to Eq. (2):

$$N(k,j) = \sum_{i=1}^{N_j} n(k,i,j), \quad (2)$$

where N_j is the number of atoms for the atom type *j* in the protein. Through this procedure, the protein structure was decomposed into five boxes according to the atom type, with each box composed of the voxels with the value of $N(k,j)$.

By decomposing each box into small boxes of 48^3 voxels (Fig. 1), we made our DL model applicable to proteins with arbitrary sizes. Hereinafter, the box is referred to as the "partial protein box" and the set of the boxes of five atom types at the same position in the protein is referred to as the "set of partial protein box". To exclude the effect of the boundary of the partial protein box on the training and test by conducting them with the central 16^3 voxels, the decomposition was performed by translating the partial protein box by 16 voxels (Fig. 1).

The output format of our DL model was the same as that of the water box of $g_\alpha(\boldsymbol{r})$ (α = H or O) obtained using the 3D-RISM theory. For the training of our DL model, the water box was also decomposed into the boxes of 48^3 voxels as previously described. The resulting box is referred to as the "partial water box". With a set of partial protein box as input, our DL model outputs the corresponding partial water box. By summing the central 16^3 voxels of each partial water box, $g_{\mathrm{H}}(\boldsymbol{r})$ or $g_{\mathrm{O}}(\boldsymbol{r})$ were obtained.

Deep-learning model. For the network architecture for our DL model that predicts the protein hydration structure, we employed the U-net[20]. As schematically shown in Fig. 2, the U-net is an encoder-decoder type architecture[21]. The deep-learning model was constructed for predicting a distribution function $g_\alpha(\boldsymbol{r})$ (α = H or O).

The architecture we employed was essentially the same as that used in the original U-net model used for biomedical image segmentation[10]. As shown in Fig. 2, both encoder and decoder consisted of four layers, referred to as the "encoder-decoder layers", and a 5th layer was prepared between the 4th layers of encoder and decoder. Each encoder-decoder layer consisted of two convolutional layers, each followed by the activation using the ReLU function. A $2\times2\times2$ max-pooling layer was added after the second convolutional layer of the encoder-decoder layer in the encoder. The max-pooling layer was replaced by an upsampling layer[20] in the encoder-decoder layer in the decoders. The number of filters in the first convolutional layers in an encoder-decoder layer was doubled from the previous layer in the encoder, and a half from the previous layer in the decoder, respectively. Skip connection was added according to the original U-net architecture.

Our DL model differed from the original U-net model in four aspects. First, while the original U-net model was implemented for two-dimensional images, our model was implemented for three-dimensional data of a partial protein box and a partial water box. Furthermore, in our model the zero-padding was added in the convolutional layers. Moreover, when training our DL model, the dropout layer was also added after the ReLU activation in the first convolutional layer, to reduce the overfitting. Finally, the convolution followed by the up-sampling in the original U-net architecture was removed in our architecture. The effect of convolution in the upsampling layer was small (text S2 in the Supporting Information).

As shown in Table 2, our DL model had four hyperparameters, collectively referred to as "hyperparameter set", one of which was the filter size (i.e., number of voxels in the filters) in the convolutional layers. Three sizes were prepared, namely 3^3, 4^3, and 5^3. The number of filters ($N_{\mathrm{Firstfilter}}$) for an atom type at the first encoder-decoder layer in the encoder was another hyperparameter. Because the number of atom types was five, the total number of filters was $5 \cdot N_{\mathrm{Firstfilter}}$. The third hyperparameter was the number of voxels whose value was set at zero in the dropout layer, N_{D}. The ratio of N_{D} and the total number of voxels in the dropout layer, referred to as "dropout ratio", was set to 0.3 or 0.5. Finally, the dropout was applied to the 5^{th} layer, 4^{th}–5^{th} layers, 3^{rd}–5^{th} layers, 2^{nd}–5^{th} layers, or all layers for each of (i) only the encoder, (ii) only the decoder, or (iii) both the encoder and the decoder. The case in which no dropout was applied to both the encoder and the decoder was also considered.

Our DL model was implemented using the TensorFlow library (2.1.0) for the models predicting $g_{\alpha}(\boldsymbol{r})$ ($\alpha = \mathrm{H}$ or O). The training was performed using Adam optimizer with default parameters.

Hyperparameter optimization. The hyperparameters were optimized through the two-fold cross validation. The twenty-two proteins used for the training were split into two sets to homogeneously adjust the total number of partial protein boxes of two sets: ten proteins and the remaining twelve proteins (Table 1). The number of partial protein boxes (N_{TData}) was 6,858 and 7,101 from the ten and twelve proteins, respectively. The number of partial water boxes was also N_{TData}. In the cross validation, when the ten proteins were used for the training of our DL model, the remaining twelve proteins were used for the validation of the model and vice versa. Hereinafter, the data for the training and validation are denoted by "training data" and "validation data", respectively. Each data is composed of the partial protein boxes and the corresponding partial water boxes.

For a hyperparameter set of our DL model, the number of epochs was set to 200 and the mean-square error in Eq. (3) was employed for the loss function, E:

$$E = \frac{1}{N_{\mathrm{TData}} N_{\mathrm{Voxel}}} \sum_{j=1}^{N_{\mathrm{TData}}} \sum_{i=1}^{N_{\mathrm{Voxel}}} \left(g_{\alpha,j}^{\mathrm{Model}}(\boldsymbol{r}_i) - g_{\alpha,j}^{\mathrm{3D-RISM}}(\boldsymbol{r}_i) \right)^2 \qquad (3)$$

where N_{Voxel} is the number of voxels in the box, equal to 16^3; $g_{\alpha,j}^{\mathrm{X}}(\boldsymbol{r}_i)$ (X="Model" or "3D-RISM", and α="O" or "H") is the $g_\alpha(\boldsymbol{r})$ value at the voxel position of $\boldsymbol{r}_i$ for the j-th box; the superscripts "Model" and "3D-RISM" indicate the $g_\alpha(\boldsymbol{r})$ obtained from our DL model and the 3D-RISM theory, respectively. Hereinafter, the E value at i epoch is denoted by $E_{\mathrm{Train}}(i)$.

For each epoch (i) in the training, we also computed the loss function of Eq. (3) using the validation data ($E_{\mathrm{Validation}}(i)$) to check whether the overfitting did not occur during the training. In the computation, all dropout layers used in the training were not used. The possible overfitting was checked by the comparability of the $E_{\mathrm{Validation}}(200)$ and $E_{\mathrm{Training}}(200)$ values.

Each of the four hyperparameters was set to a value within the range shown in Table 3, leading to 162 hyperparameter sets. For each hyperparameter set, the following computations were performed: (1) Training was performed with the ten proteins and the corresponding $g_{\mathrm{H}}(\boldsymbol{r})$ or $g_{\mathrm{O}}(\boldsymbol{r})$ as the training data; (2) the $E_{\mathrm{Validation}}(200)$ value was saved; (3) the procedures (1) and (2) were repeated with the twelve proteins for the training, and (4) the average of the two $E_{\mathrm{Validation}}(200)$ values, namely $\bar{E}_{\mathrm{Validation}}(200)$, was computed. After the computations for all the 162 hyperparameter sets, the hyperparameter set with the smallest $\bar{E}_{\mathrm{Validation}}(200)$ value, denoted by "optimized hyperparameter set", was selected.

Tests. After the training using the optimized hyperparameter set and the twenty-two proteins, $g_{\alpha}(\boldsymbol{r})$ (α="O" or "H") for the five proteins described in Table 1 was computed as a test. The training procedure was the same as that described in the previous subsection. In the test, each protein was first converted into the voxel format described in "Input and output formats for the DL model" subsection. Afterwards, using a partial protein box, $g_{\alpha}(\boldsymbol{r})$ of the corresponding partial water box was predicted using our DL model. Finally, $g_{\alpha}(\boldsymbol{r})$ of the whole protein was obtained by summing the central 16^3 voxels of the predicted water boxes.

To quantitatively compare the peak positions of $g_{\mathrm{O}}(\boldsymbol{r})$ of our DL model with those of the 3D-RISM theory, the following analysis was conducted. First, water oxygen atoms were placed using the program Placevent[22], in which water oxygen atoms were placed using the $g_{\mathrm{O}}(\boldsymbol{r})$ values. In the program, the placement of water oxygen atoms was performed in three steps: (i) A water oxygen atom was placed at the position of the voxel with the largest $g_{\mathrm{O}}(\boldsymbol{r})$ value (denoted by $\boldsymbol{r}_{\mathrm{Max}}$); (ii) The region $\boldsymbol{\delta}$ satisfying Eq. (4) was identified:

$$\int_{\boldsymbol{r}_{\mathrm{Max}}}^{\boldsymbol{r}_{\mathrm{Max}}+\boldsymbol{\delta}} \rho_0 \, g_{\mathrm{O}}(\boldsymbol{r}) d\boldsymbol{r} = 1 \tag{4}$$

(iii) The $g_{\mathrm{O}}(\boldsymbol{r})$ values at the voxels within the region $\boldsymbol{\delta}$ were set at zero (the obtained $g_{\mathrm{O}}(\boldsymbol{r})$ is referred to as "new $g_{\mathrm{O}}(\boldsymbol{r})$"); (iv) Steps (i), (ii), and (iii) were repeated using the new $g_{\mathrm{O}}(\boldsymbol{r})$ until $g_{\mathrm{O}}(\boldsymbol{r}_{\mathrm{Max}}) < 1.5$, which is the default value in the program Placevent. The placement of water oxygen atoms was performed using the $g_{\mathrm{O}}(\boldsymbol{r})$ values obtained with the 3D-RISM theory and those obtained with our DL model. The positions of the *i*-th water oxygen atom obtained using the 3D-RISM theory and that using our DL model are denoted by $\boldsymbol{r}_i^{\mathrm{RISM}}$ and $\boldsymbol{r}_i^{\mathrm{Model}}$, respectively. The number of placed water oxygen atoms is denoted by N_{RISM} and N_{Model}, and typically $N_{\mathrm{RISM}} \neq N_{\mathrm{Model}}$ because $g_{\mathrm{O}}(\boldsymbol{r})$ obtained using our DL model was slightly different from that obtained using the 3D-RISM theory.

Afterwards, for each water oxygen atom placed using $g_{\mathrm{O}}(\boldsymbol{r})$ obtained with the 3D-RISM theory, the distance D_i defined in Eq. (5) was computed:

$$D_i \equiv \min_j \left| \boldsymbol{r}_i^{\mathrm{RISM}} - \boldsymbol{r}_j^{\mathrm{Model}} \right| \tag{5}$$

The average of D_i among the water oxygen atoms placed using the 3D-RISM theory and its standard deviation were obtained to analyze the results and histogram of D_i.

To investigate the prediction performance at the ligand-binding pocket from the viewpoint of D_i, the following analysis was further performed. The water oxygen atoms at the ligand-binding pocket were defined by the placed water oxygen atoms within 5 Å from the heavy atoms of the ligand. For each water oxygen atom placed using the $g_{\mathrm{O}}(\boldsymbol{r})$ of 3D-RISM theory, the D_i value was computed.

Finally, the prediction performance from the viewpoint of the positions of crystal waters was studied via the analysis of the crystal waters within 5 Å from the heavy atoms in the protein. For each crystal water, D_i was computed.

3. RESULTS AND DISCUSSION

In this section, the results obtained for $g_{\mathrm{O}}(\boldsymbol{r})$ are presented, whereas those for $g_{\mathrm{H}}(\boldsymbol{r})$, which were analogous to those for $g_{\mathrm{O}}(\boldsymbol{r})$, are discussed in the subsection "Prediction of the distribution function of water hydrogen site".

Cross validations. To determine the optimized hyperparameter set, a two-fold cross validation was conducted as described in "Optimization of the hyperparameters". The number of partial protein boxes was 6,858 and 7,101 from the ten and twelve protein sets, respectively. The error values at 200 epoch defined by Eq. (3) were computed using the ten and twelve proteins, and their $E_{\mathrm{Validation}}(200)$ and $\bar{E}_{\mathrm{Validation}}(200)$ values are summarized in Table S1.

The optimized hyperparameter with the smallest $\bar{E}_{\mathrm{Validation}}(200)$ value was the hyperparameter set number 44 (Table S2). The hyperparameter values and their statistics are summarized in Table 3. The $\bar{E}_{\mathrm{Validation}}(200)$ value, namely the difference between the $g_{\mathrm{O}}(\boldsymbol{r})$ values of 3D-RISM and of our DL model, was sufficiently small (0.0042). The corresponding average deviation of $g_{\mathrm{O}}(\boldsymbol{r})$ obtained by our DL model from $g_O(\boldsymbol{r})$ obtained by the 3D-RISM theory was 0.06. As shown in Fig. S1, $\bar{E}_{\mathrm{Validation}}(200)$ was analogous to $\bar{E}_{\mathrm{Train}}$ (200), indicating the absence of overfitting.

The results reported in the next paragraphs were performed using the DL model trained with the optimized hyperparameter set. The results for the other hyperparameters are discussed in text S3 (Supporting Information).

Prediction tests. The correlation between the $g_O(\boldsymbol{r})$ values predicted by our DL model and those calculated by the 3D-RISM theory, coefficient of determination R^2 score values, and root mean square error (RMSE) of five proteins for the test are shown in Fig. 3. For the five proteins tested, the R^2 values were high and most of the points resided close to the line representing $y=x$. Moreover, the RMSE values indicated the accuracy of the $g_O(\boldsymbol{r})$ prediction of our DL model. The encouraging result on the accuracy was accompanied by a drastic decrease of computation time of two orders of magnitude: the computation was completed within a minute with our DL model and a single GPU.

Furthermore, the comparison was performed with the voxels in the ligand-binding pocket defined by those within 5 Å from the heavy atoms in the ligand (Fig. 4). The prediction performance of our DL model was high also in the ligand-binding pocket: most of the points resided close to the line $y=x$, with high R^2 values. Therefore, our DL model can successfully predict the hydration structure in the ligand-binding pocket. High prediction-accuracy of $g_O(\boldsymbol{r})$ in the ligand-binding site is important for the structure-based design of new molecules using the information of the hydration in the binding site.

For all the proteins, the RMSE value of the ligand-binding site was larger than that of all points, reasonably due to the value of $g_O(\boldsymbol{r})$ at the bulk region. It was found from slice 5 of Fig. 5 that the prediction performance at the bulk region is good: $g_O(\boldsymbol{r})$ at the bulk region was one for both of

the 3D-RISM theory and our DL model. This result explains the RMSE value of all points, including most of the bulk points, smaller than that of the ligand-binding site.

Finally, the results for shank3 PDZ domain (Protein Data Bank (PDB) code: 3o5n) are shown in Fig. 5 to discuss how our DL model reproduced the $g_O(\boldsymbol{r})$ values in detail. The agreement between the $g_O(\boldsymbol{r})$ values of the 3D-RISM theory and those of our DL model was good, as both the peak heights and the peak positions were well predicted by our DL model. Additionally, our DL model reproduced the $g_O(\boldsymbol{r})$ values inside the protein (slice 6 in Fig. 5) and those at a bulk region (slice 5 in Fig. 5). Moreover, a high R^2-score value (0.985) indicated the good correlation between $g_O(\boldsymbol{r})$ values of our DL model and those of 3D-RISM theory (Fig. 3). However, for few points, the $g_O(\boldsymbol{r})$ values of our DL model deviated from those of the 3D-RISM theory. Particularly, the points with large deviation corresponded to the areas in the cavities with a size comparable to that of the water molecule (Fig. S2). The current training data did not contain sufficient data for such cavities. Adding such data would therefore improve the performance of our DL model.

Placement of water oxygen atoms. To discuss how our DL model successfully predicted the peak positions of $g_O(\boldsymbol{r})$, water oxygen atoms were placed at the $g_O(\boldsymbol{r})$ peaks using the program Placevent. For the placement of water oxygen atoms, the values of $g_O(\boldsymbol{r})$ obtained using the 3D-RISM theory and our DL model were used. The histograms of D_i for the five proteins are shown in Fig. 6, whereas the average of D_i, N_{RISM}, and N_{Model} for each protein are summarized in Table 4. The histograms related to the water molecule placed at the point $g_O(\boldsymbol{r})$>1.5 (probability 1.5 times higher than that of a bulk water) in Fig. 6 (a), (d), (g), (j), and (m) indicate that approximately 60% of the water oxygen atoms of our DL model was placed within 0.5 Å from the water oxygen atoms of the 3D-RISM theory. The average of D_i was 0.6–0.7 Å for all five

proteins (Table 4 and Fig. 7). The calculated value was 1/4–1/5 of the Lennard-Jones sigma value, associated to the radius of the atom, of the water oxygen atom for the coincident SPC/E model (3.17 Å). Therefore, the $g_{\mathrm{O}}(\boldsymbol{r})$ peak positions obtained using our DL model were close to those obtained using the 3D-RISM theory. Essentially the same results were obtained for the $g_{\mathrm{O}}(\boldsymbol{r})$ values at the ligand-binding pocket (Fig. 6(b), (e), (h), (k), (n), and Table 5).

Nevertheless, the N_{Model} values were different from the N_{RISM} values (Table 4), reasonably because the peak height and peak position of $g_{\mathrm{O}}(\boldsymbol{r})$ were slightly different in the two methods. Particularly, the N_{Model} values were smaller than the corresponding N_{RISM} value for all the proteins because our DL model predicted smaller peak values of $g_{\mathrm{O}}(\boldsymbol{r})$ than those obtained using the 3D-RISM theory.

The water placement results of our DL model were afterwards compared with the positions of the crystal waters. To this end, D_i was computed for the crystal waters within 5 Å from the heavy atoms of the protein. As shown in Table 4 and Fig. 6 and 7, the average of D_i (1.0–1.6 Å for all five proteins) was 1/2–1/3 of the Lennard-Jones sigma value of the water oxygen atom for the coincident SPC/E model, indicating that the positions of water oxygen atoms obtained using our DL model were close to those of crystal waters.

Prediction of the distribution function of water hydrogen sites. A DL model for predicting $g_{\mathrm{H}}(\boldsymbol{r})$ was constructed using the same U-net architecture as that used in the deep-learning model for $g_{\mathrm{O}}(\boldsymbol{r})$. The optimized hyperparameter set (44 in Table S1) was selected considering that the prediction results were not sensitive to this factor (text S3 in Supporting Information). After training the DL model for $g_{\mathrm{H}}(\boldsymbol{r})$ using the optimized hyperparameter set and the twenty-two proteins described in Table 1, the DL model for $g_{\mathrm{H}}(\boldsymbol{r})$ was applied to the five proteins described

in Table 1 to predict $g_{\mathrm{H}}(\boldsymbol{r})$. The correlation between the predicted $g_{\mathrm{H}}(\boldsymbol{r})$ values and the $g_{\mathrm{H}}(\boldsymbol{r})$ values obtained using the 3D-RISM theory is reported in Fig. 8, together with the R^2 score and RMSE values. The DL model for predicting $g_{\mathrm{H}}(\boldsymbol{r})$ exhibited an analogous performance as that for predicting $g_{\mathrm{O}}(\boldsymbol{r})$. Additionally, the RMSE values of $g_{\mathrm{H}}(\boldsymbol{r})$ were smaller than those of $g_{\mathrm{O}}(\boldsymbol{r})$.

Selection of twenty-seven proteins. To investigate the possible effects of the selection of the proteins on the performance of our DL model, two analyses were conducted.

First, our DL model was applied to the prediction of $g_{\mathrm{O}}(\boldsymbol{r})$ for the 2,691 proteins that were not involved in the twenty-seven proteins in Table 1. The PDB codes of the 2,691 proteins, their R^2 score values, and their classes are summarized in "Data2718-SI-Forsubmit.xlsx". The R^2 score values for the 2691 proteins and the five test proteins were larger than 0.98 for all proteins (Fig. 9). Therefore, our DL model can successfully be applied to various proteins.

In the second analysis, a different pool of twenty-seven proteins was randomly selected for the training and test (Table S4). Twenty-two proteins were used for the training of the DL model for predicting $g_{\mathrm{O}}(\boldsymbol{r})$, and the remaining five proteins were used for the test, with set 44 in Table 3 adopted as hyperparameter set. As shown in Fig. S3, the prediction performance was comparable to that shown in Fig. 3, indicating that the negligible effects of the selection of the twenty-seven proteins on the performance of our DL model.

Therefore, the selection of the twenty-seven proteins shown in Table 1 did not affect the performance of our deep-learning model.

Comparison of our deep-learning model with other related methodologies. Our DL model was compared with three related methodologies. Two of the three are the method for obtaining

the hydration structures around proteins within a short computation time[23,24]. The other is the hybrid method of a DL and the 3D-RISM theory[25].

First, our DL model is compared with the hybrid method of a DL and the 3D-RISM theory proposed by Sosnin *et al*[25]. Contrarily to our DL model directly predicting $g_{\mathrm{O}}(\boldsymbol{r})$ and $g_{\mathrm{H}}(\boldsymbol{r})$, Sosnin *et al*. proposed a DL model for predicting the bioconcentration-factor values of organic molecules with the input of $g_{\mathrm{O}}(\boldsymbol{r})$ and $g_{\mathrm{H}}(\boldsymbol{r})$ obtained with the 3D-RISM theory. The employed DL model was also different: Sosnin *et al*. employed a three-dimensional convolutional neural network.

Ghanbarpour *et al*.[23] proposed a DL model for predicting the hydration structure around the proteins. In their study, the hydration structure was characterized by the water occupancy, namely the probability that a water molecule is found at a given grid position. From the definition of the water occupancy, it is closely related to $g_{\mathrm{O}}(\boldsymbol{r})$. Although Ghanbarpour *et al*. attempted to predict the water occupancies using the model based on the U-net architecture, the prediction performance was unsatisfactory. Therefore, they proposed another regression model to predict the water occupancies. However, their model required a preliminary classification using the model predicting the grid points into those high and low water occupancies. Such classification was not required in our DL model.

Maruyama and Hirata[24] have proposed a fast algorithm to accelerate the 3D-RISM calculation using GPU. The computation of the 3D-RISM calculation for a single protein was finished within a few minutes with a Tesla-K40 GPU[26]. Compared with the algorithm proposed by Maruyama and Hirata, our DL model had two advantages. First, even with a single CPU, the computation was rapidly completed (a few minutes). Furthermore, our DL model enabled to

compute $g_{\mathrm{O}}(\boldsymbol{r})$ at a focused region in the protein, such as the ligand-binding pocket or another region of interest, because the protein was decomposed into small boxes of 48^3 voxels. Such computation is unfeasible for the 3D-RISM theory.

4. CONCLUSIONS

In the present study, we proposed a DL model for predicting the hydration structure around the protein based on the U-net architecture. The output was the distribution function of water oxygen $g_{\mathrm{O}}(\boldsymbol{r})$ and hydrogen $g_{\mathrm{H}}(\boldsymbol{r})$ solely with the input of the protein 3D structure.

Our DL model successfully reproduced $g_{\mathrm{O}}(\boldsymbol{r})$ and $g_{\mathrm{H}}(\boldsymbol{r})$ obtained using the 3D-RISM theory of five proteins not included in the training set. The coefficient of determination, R^2-score values were approximately 0.98 for the five proteins, indicating the good performance of our DL model. Moreover, the model accurately predicted the peak positions of $g_{\mathrm{O}}(\boldsymbol{r})$ from the comparison of the positions of the water oxygen atoms, using Placevent, between our DL model and the 3D-RISM theory. The average of D_i (0.6–0.7 Å), which is the distance of water molecules between that placed by the 3D-RISM theory and the one predicted by our DL model, was small compared to the size of the water oxygen atom, 3 Å. Our DL model also successfully predicted $g_{\mathrm{H}}(\boldsymbol{r})$. In summary, our DL model exhibited a good prediction performance for $g_{\mathrm{O}}(\boldsymbol{r})$ and $g_{\mathrm{H}}(\boldsymbol{r})$.

For the whole protein, our DL model predicted $g_{\mathrm{O}}(\boldsymbol{r})$ within a minute using a single GPU on average. Moreover, $g_{\mathrm{O}}(\boldsymbol{r})$ was predicted for only a focused region of interest, such as the ligand binding domain.

One of the limitations of our DL model is the restricted atom types that can be included, namely carbon, nitrogen, oxygen, sulfur, and hydrogen. Therefore, the application of the current DL model to protein systems involving other atoms (e.g., metals, phosphorus of phosphorylated amino acids, selenium of selenomethione, ions, halogens of ligands, and co-factors) is unfeasible. To extend the applicability of our DL model, the number of atom types should be increased. The data including these atom types and training of our DL model are the object of our future publication.

DATA AND SOFTWARE AVAILABILITY

Our program, named "gr Predictor", is available under the GNU General Public License from https://github.com/YoshidomeGroup-Hydration/gr-predictor. Usage of the program is described in the web page described above. All the data used in the present study have been exhaustively presented in the manuscript.

ASSOCIATED CONTENT

Supporting Information.

The following files are available free of charge.

brief description (file type, i.e., PDF)

AUTHOR INFORMATION

Corresponding Author

Takashi Yoshidome - Department of Applied Physics, Graduate School of Engineering, Tohoku University, Sendai 980-8579, Japan; ORCID ID: 0000-0001-7407-1942; Email: takashi.yoshidome.b1@tohoku.ac.jp.

Authors

Kousuke Kawama - Department of Applied Physics, Graduate School of Engineering, Tohoku University, Sendai 980-8579, Japan.

Yusaku Fukushima - Department of Applied Physics, Graduate School of Engineering, Tohoku University, Sendai 980-8579, Japan.

Mitsunori Ikeguchi - Graduate School of Medical Life Science, Yokohama City University, 1-7-29, Suehiro-cho, Tsurumi-ku, Yokohama 230-0045, Japan, and AI-driven Drug Discovery Collaborative Unit, HPC- and AI-driven Drug Development Platform Division Center for Computational Science, RIKEN, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa, 230-0045, Japan; ORCID ID: 0000-0003-3199-6931.

Masateru Ohta - AI-driven Drug Development Platform Division Center for Computational Science, RIKEN, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa, 230-0045, Japan; ORCID ID: 0000-0002-6580-7185.

Author Contribution

T.Y., M.O., and M.I. designed the study. K.K. and Y.F. performed the computations, and K.K., Y.F., and T.Y. analyzed the data. T.Y. and M.O. wrote the article.

Funding

This work was financially supported by JSPS KAKENHI, Grant Number 21K06107, and by a Grant-in-Aid for Scientific Research on Innovative Areas "Molecular Engine" (JSPS KAKENHI Grant Number: 21H00381).

ACKNOWLEDGMENT

Part of the computation was carried out using the computer resource offered under the category of HPCI System Research Project (Project ID: hp210081) by Research Institute for Information Technology, Kyushu University. T.Y. thanks to Mr. Dan Ohashi for the computation of the prediction of $g_O(\boldsymbol{r})$ for 2,696 proteins.

REFERENCES

(1) 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*, 973-80.

(2) Kinoshita, M. A New Theoretical Approach to Biological Self-Assembly, *Biophys. Rev.* **2013**, *5*, 283-293.

(3) Hirata F. Ed., Molecular Theory of Solvation, Springer Science & Business Media, Dordrecht: 2003.

(4) Mahmoud, A.H.; Masters, M.R.; Yang, Y.; Lill, M.A.; Elucidating the Multiple Roles of Hydration for Accurate Protein-ligand Binding Prediction via Deep Learning, *Commun. Chem.* **2020**, *3*, 19(1-13).

(5) Imai, T.; Hiraoka, R.; Kovalenko, A.; Hirata, F. Water Molecules in a Protein Cavity Detected by a Statistical-Mechanical Theory, *J. Am. Chem. Soc.* **2005**, *127*, 44, 15334-15335.

(6) Kirkwood, J.G.; Buff, F.P. The Statistical Mechanical Theory of Solutions. I, *J. Chem. Phys.* **1951**, *19*, 774-777.

(7) Imai T.; Kovalenko A.; Hirata F. Solvation Thermodynamics of Protein Studied by the 3D-RISM Theory, *Chem. Phys. Lett.* **2004**, *395*, 1-6.

(8) Imai T.; Kovalenko A.; Hirata F. Partial Molar Volume of Proteins Studied by the Three-Dimensional Reference Interaction Site Model Theory, *J. Phys. Chem. B* **2005**, *109*, 6658-6665.

(9) 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.

(10) Begnini, F.; Geschwindner, S.; Johansson, P.; Wissler, L.; Lewis, R. J.; Danelius, E.; Luttens, A.; Matricon, P.; Carlsson, J.; Lenders, S.; König, B.; Friedel, A.; Sjö, P.; Schiesser, Stefan.; Kihlberg, J.; Importance of Binding Site Hydration and Flexibility Revealed When Optimizing a Macrocyclic Inhibitor of the Keap1–Nrf2 Protein–Protein Interaction, *J. Med. Chem.* **2022**, 65, 3473-3517.

(11) 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.

(12) Wang R.; Fang X.; Lu Y.; Wang S. The PDBbind Database: Methodologies and Updates, *J. Med. Chem.* **2005**, *48*, 4111-4119.

(13) Cheng T.; Li X.; Li Y.; Liu Z.; Wang R. Comparative Assessment of Scoring Functions on a Diverse Test Set, *J. Chem. Inf. Model.* **2009**, *49*, 1079-1093.

(14) Li Y.; Liu Z.; Li J.; Han L.; Liu J.; Zhao Z.; Wang R. Comparative Assessment of Scoring Functions on an Updated Benchmark: 1. Compilation of the Test Set, *J. Chem. Inf. Model.* **2014**, *54*, 1700-1716.

(15) Liu Z.; Li Y.; Han L.; Li J.; Liu J.; Zhao Z.; Nie W.; Liu Y.; Wang R. PDB-Wide Collection of Binding Data: Current Status of the PDBbind Database, *Bioinformatics* **2015**, *31*, 405-412.

(16) Liu Z.; Su M.; Han L.; Liu J.; Yang Q.; Li Y.; Wang R. Forging the Basis for Developing Protein-Ligand Interaction Scoring Functions, *Acc. Chem. Res.* **2017**, *50*, 302-309.

(17) Luchko, T.; Gusarov, S.; Roe, D.R.; Simmerling, C.; Case, D.A.; Tuszynski, J.; Kovalenko, A. Three-Dimensional Molecular Theory of Solvation Coupled with Molecular Dynamics in Amber, *J. Chem. Theory Comput.* **2010**, *6*, 607−624.

(18) Maier, J.A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K-E.; Simmerling K.E. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB, *J. Chem. Theory Comput.* **2015**, *11*, 3696−3713.

(19) Jiménez, J.; Škalič, M.; Martínez-Rosell, G.; De Fabritiis, G. K_{DEEP}: Protein-Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks, *J. Chem. Inf. Model.* **2018**, *26*, 287-296.

(20) Ronneberger, O.; Fischer, P.; Brox, T.; U-Net: Convolutional Networks for Biomedical Image Segmentation, *Lecture Notes in Computer Science* **2015**, *9341*, 234-241.

(21) Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-to-Image Translation with Conditional Adversarial Networks, arXiv:1611.07004

(22) Sindhikara, D.J.; Yoshida, N.; Hirata, F. Placevent: An Algorithm for Prediction of Explicit Solvent Atom Distribution-Application to HIV-1 Protease and F-ATP Synthase, *J. Comput. Chem.* **2012**, *33*, 1536-1543.

(23) Ghanbarpour, A.; Mahmoud, A.H.; Lill, M.A. Instantaneous Generation of Protein Hydration Properties from Static Structures, *Commun. Chem.* **2020**, *3*, 188(1-19).

(24) Maruyama, Y.; Hirata, F. Modified Anderson Method for Accelerating 3D-RISM Calculations Using Graphics Processing Unit, *J. Chem. Theory Comput.* **2012**, *8*, 3015-3021.

(25) Sosnin, S.S.; Maksim Misin, M.; David S Palmer, D.S; Fedorov, M.V; 3D Matters! 3D-RISM and 3D Convolutional Neural Network for Accurate Bioaccumulation Prediction, *J. Phys. Condens. Matter* **2018**, *30*, 32LT03(1-7).

(26) Yoshida, N. Role of Solvation in Drug Design as Revealed by the Statistical Mechanics Integral Equation Theory of Liquids, *J. Chem. Inf. Model.* **2017**, *57*, 2646-2656.

Table 1. Proteins used for developing and evaluating the deep-learning model. "Train: 10" and "Train: 12" are equal to "Train: Ten proteins" and "Train: Twelve proteins", respectively.

PDB ID	Structure Title	High Reso. Limit (Å)	Dataset
1A30	HIV-1 protease complexed with a tripeptide inhibitor	2.00	Train: Ten proteins
1FCH	PTS1 complexed to the TPR region of human PEX5	2.20	Train: 10
1PZ5	Antibody in complex with octapeptide	1.80	Train: 10
2CE9	A peptide bound to the Groucho-TLE WD40 domain.	2.12	Train: 10
2HKF	The complex Fab M75- Peptide	2.01	Train: 10
2PV1	SurA complexed with peptide WEYIPNV	1.30	Train: 10
2QBW	PDZ-Fibronectin fusion protein	1.80	Train: 10
3BZF	Major histocompatibility in complex with HLA-E	2.50	Train: 10

3DRF	OppA complexed with an endogenous peptide	1.30	Train: 10
3DRI	OppA co-crystallized with an octamer peptide	1.80	Train: 10
3ERY	H-2 class I histocompatibility antigen in complex with peptide	1.95	Train: Twelve proteins
3G19	ClpS protease adaptor protein in complex with peptide	1.85	Train: 12
3IFL	Amyloid beta peptide:antibody complex	1.50	Train: 12
3P9M	H2-Kb in complex with epitope OVA-G4	2.00	Train: 12
3T6B	human DPPIII in complex with Tynorphin	2.40	Train: 12
3TCG	E. coli OppA complexed with the tripeptide KGE	2.00	Train: 12
3UPV	pHsp70-complex of yeast Sti1	1.60	Train: 12
4EZR	E.coli DnaK in complex with drosocin	1.90	Train: 12
4EZZ	E.coli DnaK in complex with peptide ELPLVKI	2.05	Train: 12

4YNL	HetR in complex with the hexapeptide ERGSGR	2.10	Train: 12
5E6O	C. elegans LGG-2 bound to an AIM/LIR motif	1.80	Train: 12
5LSO	SPF45 UHM domain with cyclic peptide inhibitor	2.22	Train: 12
2HA2	Acetylcholinesterase complexed with succinylcholine	2.05	Test
2O4L	HIV-1 Protease in Complex with Tipranavir	1.33	Test
3JVR	Chk1 complexed with allosteric inhibitor	1.76	Test
3O5N	Shank PDZ domain complexed with small molecule	1.83	Test
4KAO	Focal adhesion kinase in complex with inhibitor	2.39	Test

Table 2. Hyperparameters and optimization ranges of our deep-learning model.

Hyperparameter	Range of the parameters for optimization
A. The size of filter for convolution	$[3^3]$, $[4^3]$, or $[5^3]$
B. The number of filters at the first layer in the encoder	16 or 32
C. The dropout ratio	0.3 or 0.5
D. The layers to which the dropout is applied	✓ The dropout was applied to (i) only the encoder, (ii) only the decoder, or (iii) both the encoder and the decoder. ✓ For each of (i), (ii), and (iii), the dropout was applied to 5^{th} layer, 4^{th}-5^{th} layers, 3^{rd}-5^{th} layers, 2^{nd}-5^{th} layers, or all layers. ✓ The case in which no dropout was applied to both the encoder and the decoder was also considered.

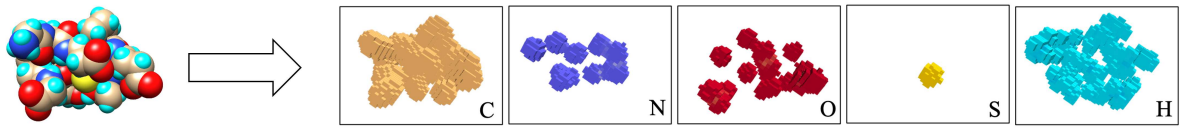
Table 3. The optimized hyperparameter and statics of our deep-learning model. The $E_{\text{Validation}}(200)$ values obtained using the ten and twelve proteins are denoted as E^1_{Valid} and E^2_{Valid}, respectively.

Hyperparameter	A. The size of filter for convolution	$[3^3]$
	B. The number of filters at the first layer in the encoder	32
	C. The dropout ratio	0.3
	D. The layers to which the dropout is applied	✓ both encoder and decoder ✓ 2^{nd}-5^{th} layers
Statistics	E^1_{Valid}	0.0045
	E^2_{Valid}	0.0039
	Average of E^1_{Valid} and E^2_{Valid}	0.0042

Table 4. Results using the program Placevent for the five proteins.

PDB code	Placevent $g_O(\boldsymbol{r}) > 1.5$			Placevent Ligand-binding pocket			Placevent Crystal waters	
	Average and SD of D_i (Å)	N_{RISM}	N_{Model}	Average and SD of D_i (Å)	N_{RISM}	N_{Model}	Average and SD of D_i (Å)	N_{CW}
2ha2	0.625±0.888	2326	2204	0.601±0.968	31	29	1.089±0.331	436
2o4l	0.704±0.924	753	701	0.558±0.746	64	62	1.259±0.452	217
3jvr	0.607±0.876	1530	1458	0.509±0.839	41	41	1.141±0.451	203
3o5n	0.705±0.935	682	630	1.066±1.000	33	35	1.674±1.206	28
4kao	0.626±0.860	1480	1400	0.542±0.671	39	37	0.994±0.427	30

1. Conversion of the protein structure into the voxel format in accordance with Eq. (2)



2. Decomposition into $(48)^3$ voxels by translating 16 voxels

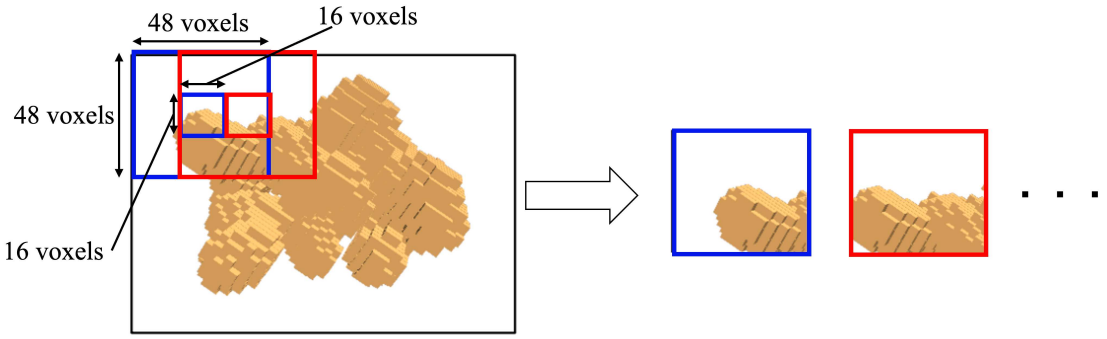


Fig. 1. Schematic of the conversion of a protein structure into the voxel format.

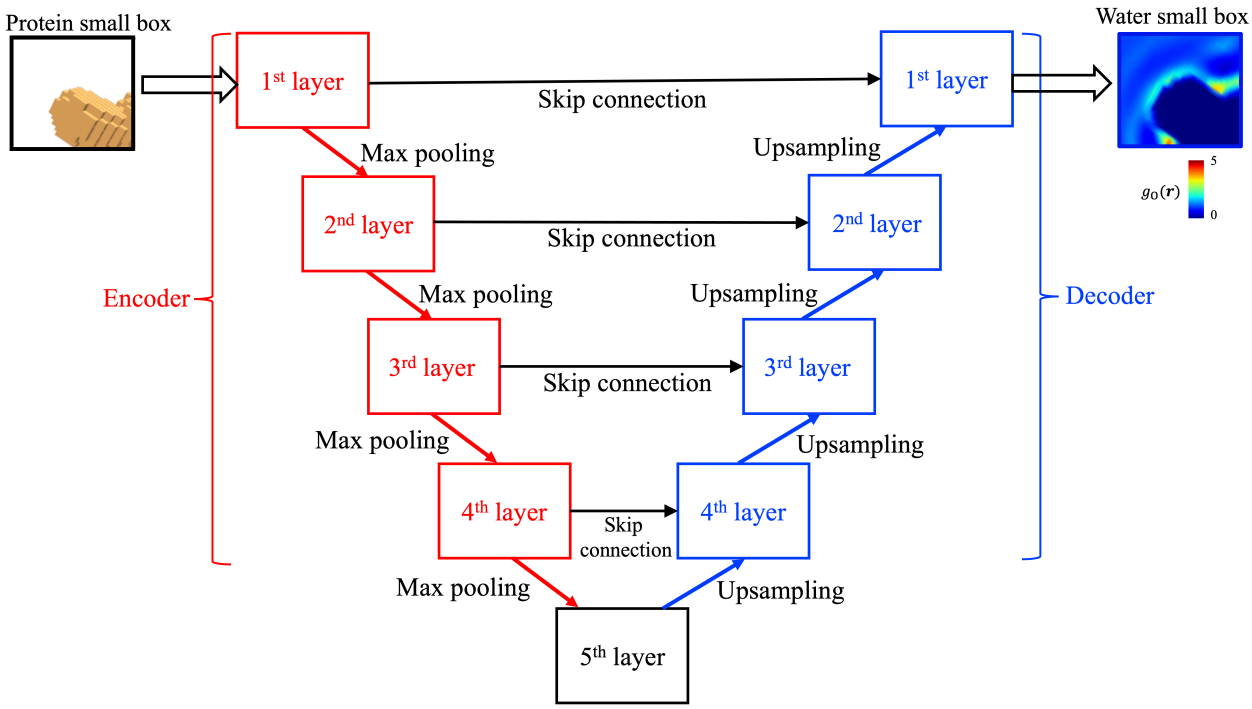


Fig. 2. Schematic of the U-net architecture.

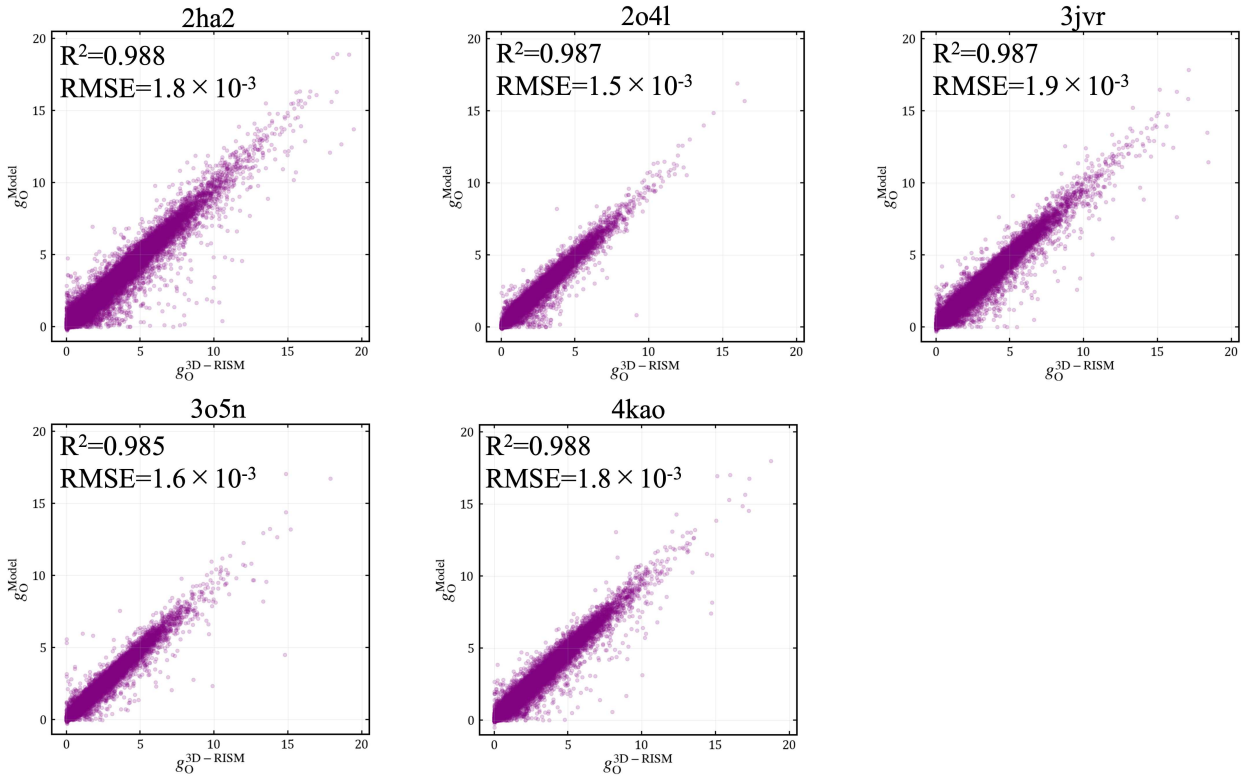


Fig. 3. Correlation between the $g_O(\boldsymbol{r})$ values predicted by our deep-learning model and those calculated by the 3D-RISM theory. The coefficient of determination R^2-score values and root mean square error (RMSE) are indicated.

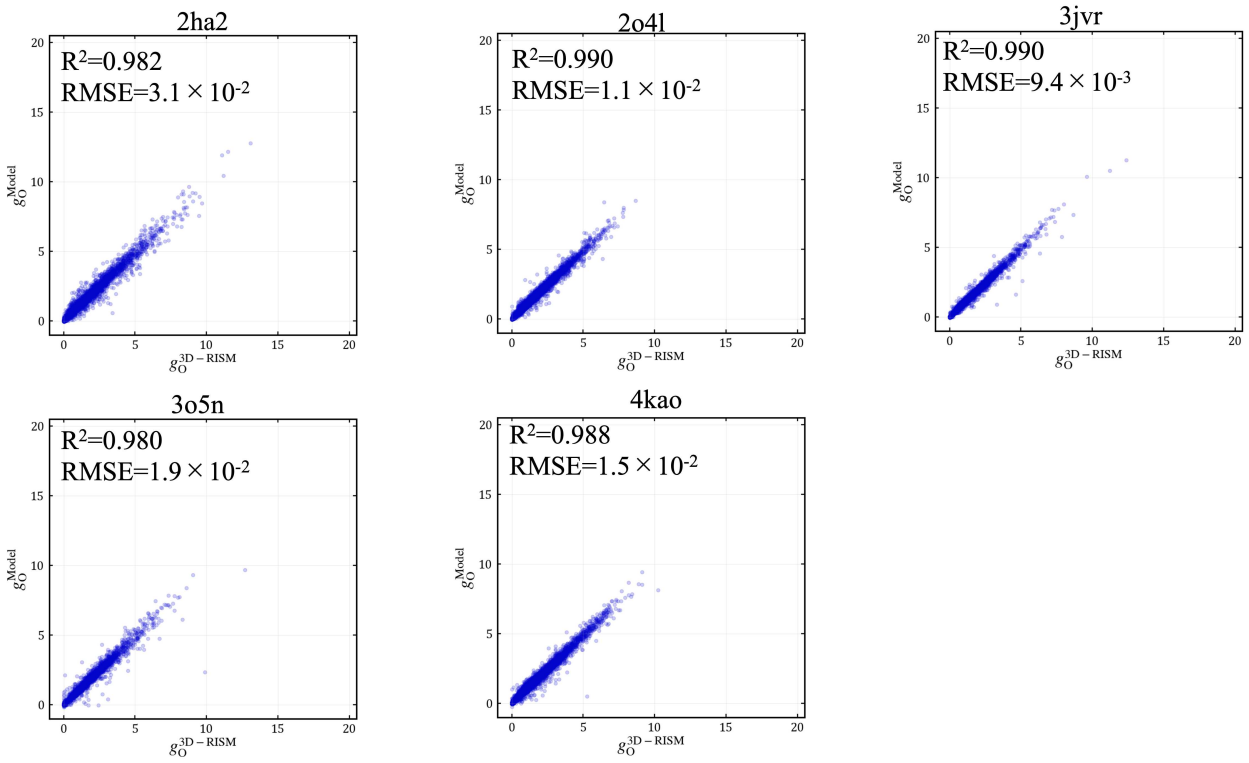


Fig. 4. Results at the ligand-binding site for the five test proteins. Correlation between the $g_O(\boldsymbol{r})$ values predicted by our deep-learning model and those calculated by 3D-RISM theory. The coefficient of determination R^2-score values and root mean square error (RMSE) are indicated.

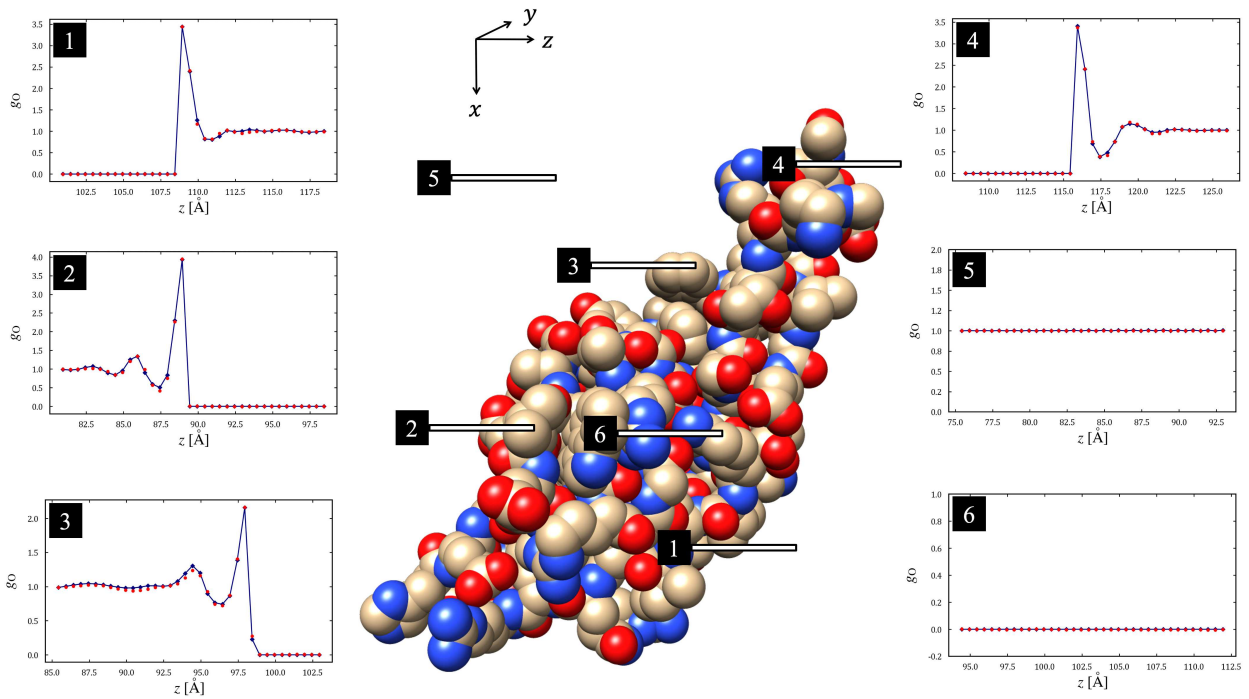


Fig. 5. Results of $g_O(\boldsymbol{r})$ for shank3 PDZ domain (PDB code: 3o5n). The $g_O(\boldsymbol{r})$ values at the six-line regions illustrated in the protein are shown. The blue lines and blue points represent the $g_O(\boldsymbol{r})$ values obtained using the 3D-RISM theory, whereas the red points represent the $g_O(\boldsymbol{r})$ values obtained with our deep-learning model.

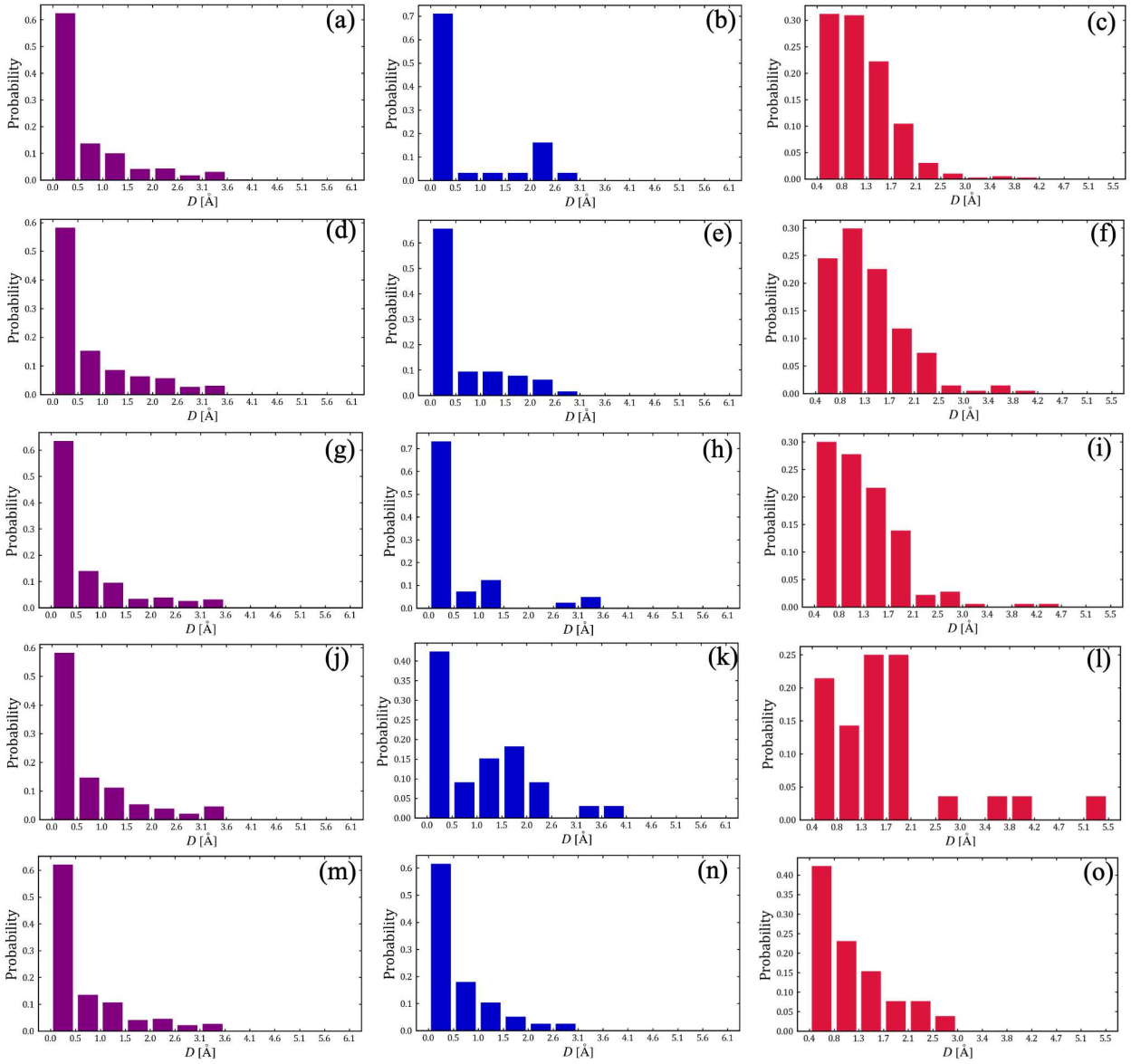


Fig. 6. Histograms of the D_i values (Eq. (5)) for five proteins, namely (a), (b), and (c) 2ha2; (d), (e), and (f) 2o4l; (g), (h), and (i) 3jvr; (j), (k), and (l) 3o5n; (m), (n), and (o) 4kao. For each protein, D_i was computed for the oxygen atoms at the positions with $g_O(\boldsymbol{r}) > 1.5$ [(a), (d), (g), (j), and (m)] and those in the ligand-binding pocket [(b), (e), (h), (k), and (n)], and for the crystal waters [(c), (f), (i), (l), and (o)].

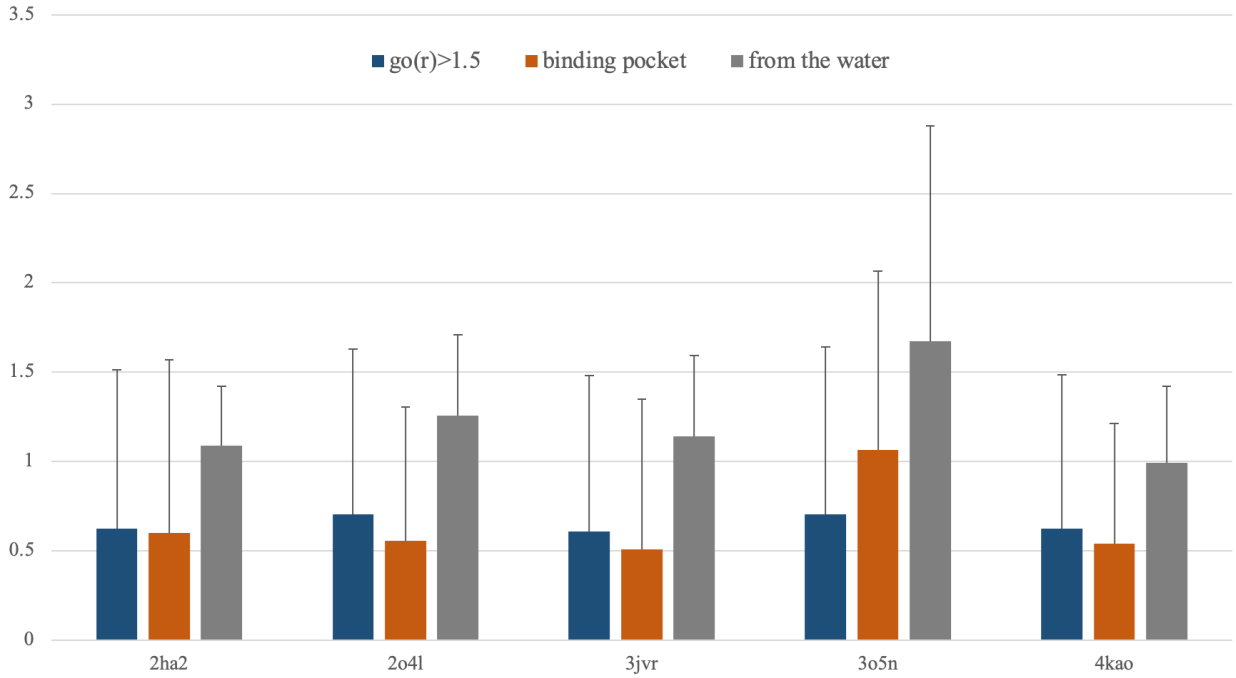


Fig. 7. Average and standard deviation of the D_i values for the five proteins.

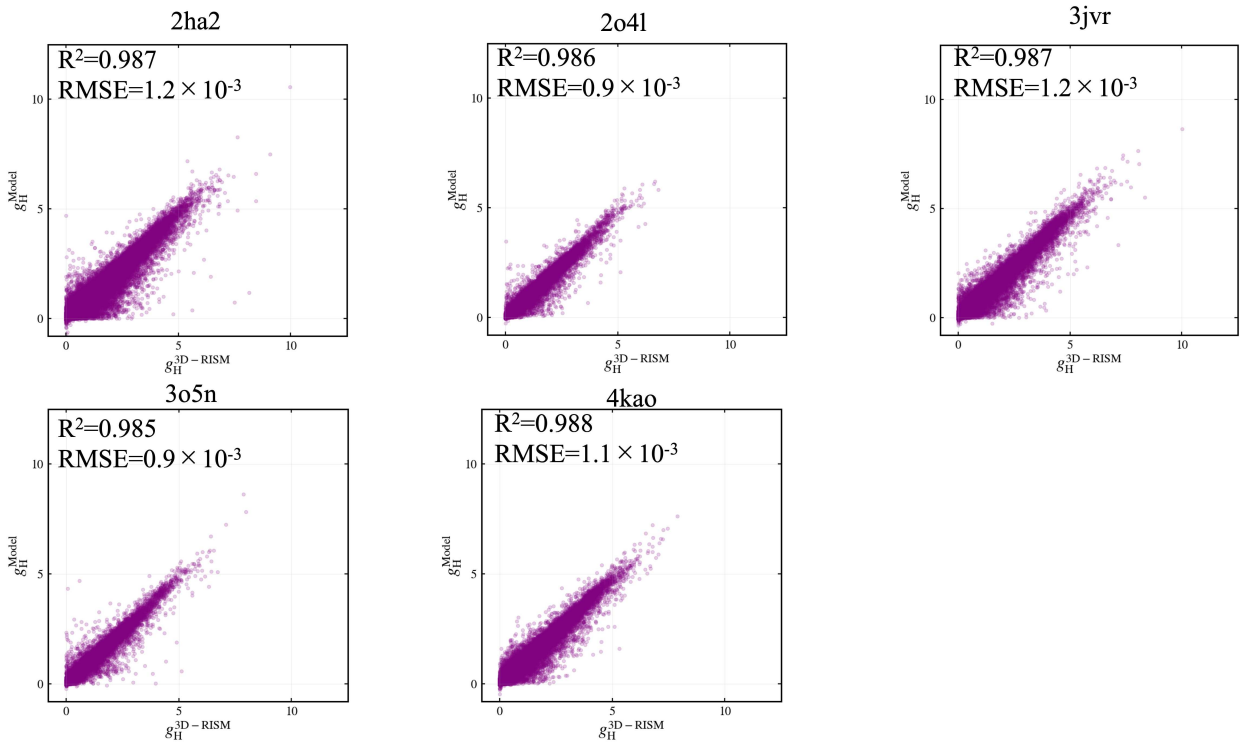


Fig. 8. Comparison of the $g_H(\boldsymbol{r})$ values obtained using the 3D-RISM theory and those obtained with our deep-learning model for (a) mouse acetylcholinesterase (PDB code: 2ha2), (b) HIV-1 Protease (PDB code: 2o4l), (c) Checkpoint kinase 1 (PDB code: 3jvr), (d) shank3 PDZ domain (PDB code: 3o5n), and (e) focal adhesion kinase (PDB code: 4kao). The coefficient of determination R^2-score values and root mean square error (RMSE) are indicated.

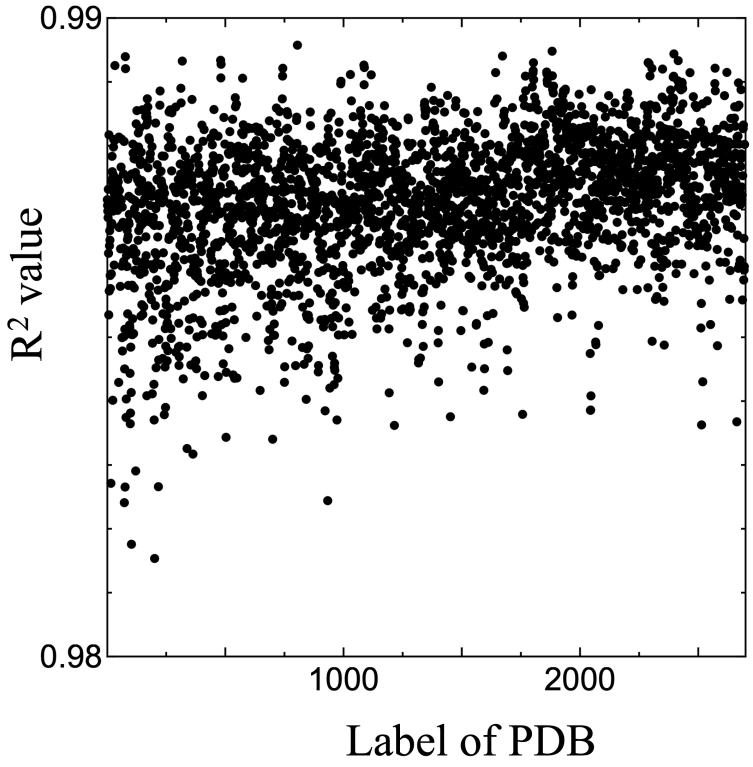


Fig. 9. R^2-score values for the 2696 proteins. The label of each PDB is reported in the file "Data2718-SI-Forsubmit.xlsx".

Table of Contents graphic

gr Predictor: A Deep-Learning Model for Predicting the Hydration Structures around Proteins

Kosuke Kawama, Yusaku Fukushima, Mitsunori Ikeguchi, Masateru Ohta, and Takashi Yoshidome*