이학박사 학위논문

GalaxyWater: Predicting Positions of Water Molecules on Protein Structure

단백질 구조 인근의 물 분자 위치 예측 방법에 대한 연구

2022 년 8 월

서울대학교 대학원

화학부 물리화학 전공

박 상 우

GalaxyWater: Predicting Positions of Water Molecules on Protein Structure

지도교수 석 차 옥

이 논문을 이학박사 학위논문으로 제출함

2022 년 8 월

서울대학교 대학원

화학부 물리화학 전공

박 상 우

박상우의 이학박사 학위논문을 인준함

2022 년 8 월

위원장 신석민 (인)

부위원장 석차옥 (인)

위 원 정연준 (인)

위 원 박한범 (인)

위 원 주기형 (인)

ABSTRACT

GalaxyWater: Predicting Positions of Water Molecules on Protein Structure

Sangwoo Park

Department of Chemistry

The Graduate School

Seoul National University

Most proteins in the living cell function in an aqueous solution, and protein molecules interact closely with water molecules. These interactions play critical roles in determining the structure and physiological function of proteins. Methods for predicting the structure or interaction of proteins consider the interaction between protein and water either implicitly or explicitly. Typical implicit solvent models consider protein-water interaction by treating solvent as a continuous dielectric medium. Such models can effectively evaluate the important electrostatic interactions with much cheaper computational costs than simulating proteins in explicit water by molecular dynamics simulation. Therefore, implicit water models are employed for protein structure prediction and docking, unlike molecular dynamic simulations. However, implicit models do not consider specific, short-range, orientation-dependent hydrogen bonds between water and protein molecules. Specific hydrogen bond interactions with water molecules are known to be involved in the structure and function of some proteins. Therefore, it is essential to consider such water molecules explicitly for detailed description and accurate prediction of protein structure and function even in the framework of implicit

solvent models. 3D-RISM is an elegant statistical mechanical method that can predict essential water molecules making specific interactions with a given protein structure using a molecular mechanics force field and an integral equation theory.

In this thesis, two methods for predicting water positions on a given protein structure are introduced. The first method is based on a new statistical potential that describes interactions between protein atoms and water molecules. The potential was derived from protein structures experimentally resolved with water molecules. A crucial part of the potential that distinguishes from other conventional potentials is consideration of the solvation environment of protein atoms during statistical derivation. This method is about 180 faster than the method based on 3D-RISM and has similar or higher performance.

Further performance improvement was achieved by adopting a machine learning approach. This method trained a convolutional neural network (CNN) on experimentally resolved structures to recognize structural patterns that favor water-binding on the protein surfaces. This method is about 44 times faster than 3D-RISM when GPGPU was used. Furthermore, the performance of locating water molecules at protein-protein interfaces and protein-ligand binding sites is also improved compared to other existing methods.

keywords: protein-water interaction, water site prediction, statistical potential, convolutional neural network

Student Number: *2013-22921*

TABLE OF CONTENTS

LIST OF FIGURES

1. INTRODUCTION

Proteins fold in water and interact with surrounding water molecules. Such interactions are closely related to protein structure and function[1-10]. In computational studies on protein structure and function, protein-water interactions are considered either explicitly or implicitly.

When predicting protein structures from amino acid sequences, the effect of water is frequently considered only implicitly to avoid the complexity of representing water explicitly. Structure prediction methods based on bioinformatics consider the water effect indirectly only using structures of evolutionarily related proteins. Structure prediction methods that rely on physics-based energy functions frequently consider water solvation via implicit solvation free energy terms, such as empirical terms that depend on solvent-accessible surface area and/or electrostatic polarization terms derived by treating water as a continuum dielectric medium[11-13]. However, such implicit solvation models ignore specific atomistic interactions of water molecules with protein atoms that may be critical to protein-ligand interactions as well as to protein folding itself. In particular, orientation-dependent hydrogen bond interactions between water and protein atoms[14] are difficult to describe using implicit solvation models. Such short-range hydrogen bond interactions affect protein-ligand interactions significantly[15, 16], and consideration of explicit water molecules is an important problem in predicting binding pose and binding affinity in protein-ligand interactions[6, 9, 10].

Molecular dynamics (MD) simulations of proteins that are fully solvated by water molecules treat water most realistically, naturally considering water both as a solvent and as molecules forming specific structural and functional interactions. However, MD simulations tend to be highly computationally expensive to be used

for protein structure prediction or protein-ligand docking directly. Notably, MD simulations with explicit water molecules have been successful in refining the protein model structures predicted by the information-based methods in the refinement category of the blind protein structure prediction experiment, CASP[17, 18].

A compromise between explicit and implicit water models is to treat only the water molecules forming specific interactions explicitly and the remaining ones implicitly. In this case, it is important to predict which water molecules will form significant structural and/or functional interactions. Various prediction methods for water positions have been developed, particularly in the protein-ligand docking field, including thermodynamic prediction methods involving MD simulations[19-26], methods based on integral equation theory[4, 27-29], and geometry-based methods considering hydrogen bond geometry[30-33]. Other examples are a protein-water docking method[34], methods refining crystal water positions in protein crystal structures[29, 35], and a method utilizing a statistical potential derived from the Protein Structure Databank (PDB)[36].

Although MD simulations tend to predict bound water positions accurately, the computational costs are high[37]. When short simulations are conducted, the water distribution functions might depend on the initial water positions[4]. A water prediction method based on integral equation theory is more rapid but less accurate than MD-based methods[37]. Although geometry-based methods can be fast, it is difficult to predict water molecules accurately using them. Docking methods can consider specific interactions; however, the prediction results depend on the accuracy of the docking score. Crystal structure refinement methods cannot predict previously unknown water positions. Although methods based on a statistical potential can predict protein-water interactions at a low computational cost, the prediction performance is dependent on the accuracy of the potential.

In this thesis, two methods for water position prediction on given protein structures are presented. In **Chapter 2**, a method called GalaxyWater-wKGB, which is based on a new statistical potential function is introduced. The statistical potential, called water knowledge-based potential based on the generalized Born model (wKGB), describes interactions between water molecules and protein atoms. This wKGB statistical potential describes water-protein interactions more precisely than a previous statistical potential used in water prediction[36] in that the dipole orientation involved in hydrogen bonds and the degree of solvent accessibility of protein atoms are considered in addition to water-protein atomic distances. From wKGB potential, water positions on protein surfaces could be predicted by identifying low-potential regions. GalaxyWater-wKGB recovered a similar or larger fraction of crystallographic water positions than 3D-RISM with a 180 times higher computational speed when the same number of water positions are predicted. In **Chapter 3**, a method called GalaxyWater-CNN, which is based on 3D-Convolutional Neural Network (CNN), is presented. The method considers the 3D structure of the protein as a kind of 3D image with 16 input channels that account for atom types. CNN was made up of 20 convolutional residual neural network[38] layers using atrous kernel[39] and returns probability map of the water molecules which interacts with protein for water position prediction. The water positions predicted by GalaxyWater-CNN showed higher coverage than GalaxyWater-wKGB[40], 3D-RISM[41, 42], and FoldX[43] when the same number of water molecules were predicted. Additionally, the protein-ligand complex version of GalaxyWater-CNN (GalaxyWater-CNN.lig) was introduced, which only uses eight atom type channels and is trained on a protein-ligand complex structure set. The prediction result of GalaxyWater-CNN.lig showed better performance than that from 3D-RISM.

In summary, this thesis presents two methods for predicting water sites from a given protein structure, which are faster and more accurate than the previous water site prediction methods, including those based on 3D-RISM. The first method, GalaxyWater-wKGB, enables fast water site prediction by using statistical potential and considering solvation state into the potential. The statistical potential, wKGB potential, can also be used with other physicochemical energy terms. The other method, GalaxyWater-CNN, enables accurate water site prediction by utilizing CNN. This method can learn protein structure patterns that can accommodate water molecules and bypass some artifacts that can arise in parameter derivation. Also, GalaxyWater-CNN is available for predicting water sites for protein-compound complexes. The methods presented in the thesis can be used to detect water sites in the binding interface of protein-protein or protein-compound complexes. By treating the detected water molecules explicitly, conventional docking methods may be improved to predict protein-protein and protein-compound complex structures and binding free energies more accurately.

2. Prediction of Water Positions on Protein Structure using wKGB Statistical Potential

GalaxyWater-wKGB is a method for predicting water positions on a protein surface, based on a statistical potential, water knowledge-based potential based on the generalized Born model (wKGB). The statistical potential describes specific protein atom-water interactions by considering the dependence on the degree of solvent accessibility of protein atoms as well as on protein atom-water distances and orientations. The introduction of solvent accessibility allows effective consideration of competing nonspecific protein-water and intra-protein interactions. From wKGB potential, water positions on protein surfaces could be predicted by identifying low-potential regions.

2.1. Methods for GalaxyWater-wKGB

2.1.1. Derivation of wKGB potential

For a given protein structure, the interaction potential between a protein atom and a water molecule is derived here by statistical analysis of the experimentally resolved protein structures deposited in PDB. The derivation was motivated by statistical potentials developed for protein structure prediction in previous studies[44, 45]. The current derivation is most similar to that of dDFIRE[44], except that the degree of solvent exposure is newly introduced here as an additional variable.

The statistical potential called as wKGB, $f_p^{\mathrm{wKGB}}(r,\theta;s)$, describes the interaction of a non-hydrogen protein atom *p* and a water oxygen atom in terms of the distance between the two atoms, *r*, and the hydrogen bond orientation, θ, which

is defined below for polar protein atoms. The dependence of the interaction on the solvent accessibility of the protein atom is also considered by a variable *s*, which is defined below. A total of 158 atom types is considered for *p*, treating the atoms in different amino acid types differently, except those that are chemically equivalent (See **Table S1** for the definitions of the atom types).

Assuming the inverse Boltzmann relationship of the potential with the observed interaction density, the statistical potential, f_p^{wKGB}, is expressed as follows, i.e., in terms of the frequency of the interactions between the protein atom and water per unit volume observed in the PDB, $\rho_p^{\mathrm{obs}}(r,\theta;s)$, relative to the reference density, $\rho_p^{\mathrm{ref}}(s)$:

$$f_p^{\mathrm{wKGB}}(r,\theta;s) = -k_{\mathrm{B}}T\log\left(\frac{\rho_p^{\mathrm{obs}}(r,\theta;s)}{\rho_p^{\mathrm{ref}}(s)}\right), \qquad (1)$$

where k_B is the Boltzmann constant and *T* is the temperature.

The hydrogen bond orientation, θ, is defined only for polar atoms, i.e., nitrogen and oxygen, and no dependence on the orientation is considered for other types of atoms, i.e., carbon and sulfur. As illustrated in **Figure 2.1**, angle θ is the angle between the local dipole vector of the protein atom, *p*, defined as $\langle \mathbf{r}_q \rangle - \mathbf{r}_p$, where $\langle \mathbf{r}_q \rangle$ is the average coordinate of the $\{q\}$ non-hydrogen atoms that are connected to *p* by chemical bonds, and vector $\mathbf{r}_{pw} = \mathbf{r}_w - \mathbf{r}_p$, where $\mathbf{r}_w$ is the coordinate of the water oxygen atom.

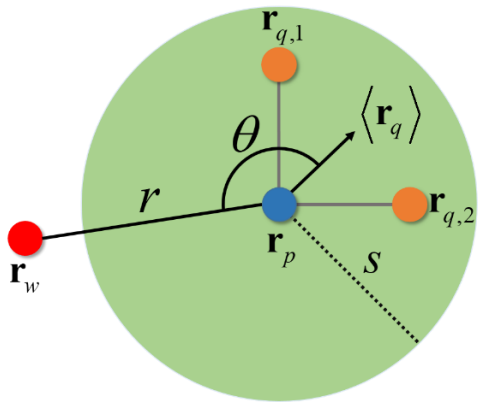


Figure 2.1. Definitions of hydrogen bond distance r and orientation θ. Hydrogen bond distance r is distance between water oxygen atom w whose coordinate vector is represented by $\mathbf{r}_w$ and protein atom p whose coordinate vector is represented by $\mathbf{r}_p$. Angle θ is angle between vector $\langle \mathbf{r}_q \rangle - \mathbf{r}_p$ and $\mathbf{r}_w - \mathbf{r}_p$, where $\langle \mathbf{r}_q \rangle$ is average coordinate of nonhydrogen atoms forming chemical bonds to atom p. Effective Born radius s is determined by local configuration of protein atoms surrounding atom p, as explained in main text.

Variable s describing the degree of solvent accessibility is defined as the effective Born radius of a generalized Born solvation free energy model called as FACTS[46]. The effective Born radius, s, of a protein atom is calculated as follows:

$$s = -\frac{\tau q^2}{2\Delta G^{\mathrm{el}}}, \tag{2}$$

where ΔG^{el} is the solvation free energy of the atom, q is the charge of the atom, and τ is $1/\varepsilon_{\mathrm{m}} - 1/\varepsilon_{\mathrm{s}}$, where ε_{m} is a low dielectric constant (1 here) and ε_{s} is a high electric constant (78.5 here). In FACTS, ΔG^{el} is estimated from the volume and spatial symmetry of the neighboring atoms[46]. Thus, the effective Born radius can be considered as the radius of a sphere with protein atom p at the center

that has the same solvation free energy contribution owing to the solvent shielding by the atoms surrounding *p* in the actual protein structure. Therefore, a large value of *s* implies low solvent accessibility, and a small value indicates high solvent accessibility.

The observed density of interaction, $\rho_p^{\mathrm{obs}}(r,\theta;s)$, is decomposed into a distance-dependent part $\rho_{\mathrm{r},p}^{\mathrm{obs}}(r;s)$ and an angle-dependent factor $\omega_p^{\mathrm{obs}}(\theta\,|\,r;s)$ as follows:

$$\rho_p^{\mathrm{obs}}(r,\theta;s)=\rho_{\mathrm{r},p}^{\mathrm{obs}}(r;s)\omega_p^{\mathrm{obs}}(\theta\,|\,r;s). \tag{3}$$

The angle-dependent factor, $\omega_p^{\mathrm{obs}}(\theta\,|\,r;s)$, is set to 1 for the nonpolar protein atoms that do not form hydrogen bonds. It is defined as the normalized number of interactions at a given hydrogen bond angle for a given distance, and the solvation state in the database as follows:

$$\omega_p^{\mathrm{obs}}(\theta\,|\,r;s)=\Omega_p^{\mathrm{obs}}(\theta\,|\,r;s)/\left\langle\Omega_p^{\mathrm{obs}}(\theta\,|\,r;s)\right\rangle_\theta, \tag{4}$$

where $\Omega_p^{\mathrm{obs}}(\theta\,|\,r;s)$ is the number of observed orientations under a given condition and $\langle\ \rangle_\theta$ denotes the average over the hydrogen bond orientations. The distance-dependent part, $\rho_{\mathrm{r},p}^{\mathrm{obs}}(r;s)$, considers the number of observed protein atom-water interactions, $N_p^{\mathrm{PDB}}(r;s)$, at a given distance for a given solvation state *s* in the PDB. The density also includes the interactions involving the inherent water molecules in the vacant space of volume $V_p^{\mathrm{vacant}}(r;s)$ with no resolved crystal water molecules, which is approximated by $V_p^{\mathrm{vacant}}(r;s)\rho_{\mathrm{w}}^{\mathrm{bulk}}$, where

$\rho_{\mathrm{w}}^{\mathrm{bulk}}$ is the bulk water density. Thus, the density, $\rho_{\mathrm{r},p}^{\mathrm{obs}}(r;s)$, is expressed as follows:

$$\rho_{\mathrm{r},p}^{\mathrm{obs}}(r;s) = N_p^{\mathrm{obs}}(r;s)/V(r;s),$$

$$N_p^{\mathrm{obs}}(r;s) = N_p^{\mathrm{PDB}}(r;s) + V_p^{\mathrm{vacant}}(r;s)\rho_{\mathrm{w}}^{\mathrm{bulk}}. \quad (5)$$

The volume of the vacant space, $V_p^{\mathrm{vacant}}(r;s)$, is estimated by counting the number of grid points in a 3D grid representation of the protein structure with a 0.5 Å spacing. The available volume, $V(r;s)$, at a given distance for a given solvation state is approximated as

$$V(r;s) = \begin{cases} \beta_s (r/\text{Å})^{\alpha_s} / \rho_w^{\mathrm{bulk}}, & r < 4 \text{ Å} \\ < N_p^{\mathrm{obs}}(r;s) >_p / \rho_w^{\mathrm{bulk}}, & r \geq 4 \text{ Å} \end{cases}, \quad (6)$$

where $\langle \ \rangle_p$ denotes the average over protein atoms. At short distances, the volume estimation by grid counting and the estimated occupation of water in the vacant space can be erroneous. Therefore, an approximate formula of $\beta_s r^{\alpha_s}$ is used by extrapolating the data values of $N^{\mathrm{obs}} =< N_p^{\mathrm{obs}}(r;s) >_p$ to short distances. The values of unitless parameters α_s and β_s are provided in Supplementary **Figure S1** with plots of N^{obs}. The reference density, $\rho_p^{\mathrm{ref}}(s)$, is defined as

$$\rho_p^{\mathrm{ref}}(s) = N_p^{\mathrm{obs}}(r^{\mathrm{cut}};s)/V(r^{\mathrm{cut}};s), \quad (7)$$

which is the density at the longest distance bin, $r^{\mathrm{cut}} = 9.75$ Å (see the next paragraph for distance bins).

The wKGB potential was obtained by dividing the distance into 20 bins from 0 to 10 Å with a 0.5 Å spacing. The hydrogen bond angle was separated into six bins with a uniform spacing in the cosine space, and the solvation state into six bins of <4, 4-5, 5-6, 6-7, 7-8, and >8 Å in the effective Born radius. A database of nonredundant, high-resolution crystal structures having resolution better than 1.8 Å and R-free values better than 0.3 with mutual sequence identity less than 30% constructed using PISCES[47] at the beginning of this study (March 14, 2015) was used to derive the potential. The total number of protein structures was 5,329. Only water oxygen atoms with B-factor < 40 were considered. In this high-resolution structure set, the average number of protein atom-water pairs is 2,492,672 per atom type. The atom type with the least (and the largest) number of pairs has 404,131 (and 10,101,461) pairs.

Noises exist in the number of observed interactions for some bins, particularly bins for short distances, owing to the insufficient data points in the database. Therefore, a smoothed potential $\overline{f}_p^{\mathrm{wKGB}}(r,\theta;s)$ is finally obtained as follows:

$$\overline{f}_p^{\mathrm{wKGB}}(r,\theta;s) = \overline{f}_{\mathrm{r},p}^{\mathrm{wKGB}}(r;s) + \overline{f}_{\theta,p}^{\mathrm{wKGB}}(\theta \mid r;s),$$

$$\overline{f}_{\mathrm{r},p}^{\mathrm{wKGB}}(r;s) = -k_{\mathrm{B}}T \log\left[\sigma_{\mathrm{r}}(r) \frac{\rho_{\mathrm{r},p}^{\mathrm{obs}}(r;s)}{\rho_p^{\mathrm{ref}}(s)} \times 0.1 + 0.9 \right],$$

$$\overline{f}_{\theta,p}^{\mathrm{wKGB}}(\theta \mid r;s) = -k_{\mathrm{B}}T \log\left[\sigma_{\theta}(r) \left\{ \omega_p^{\mathrm{obs}}(\theta \mid r;s) - 1 \right\} \times 0.1 + 1 \right], \quad (8)$$

$$\sigma_{\mathrm{r}}(r) = \frac{1}{1 + e^{-3(r/\text{Å}-2)}},$$

$$\sigma_{\theta}(r) = \frac{1}{1 + e^{-10(r/\text{Å}-2)}}.$$

Defined as above, the smoothed radial potential, $\overline{f}_{\mathrm{r},p}^{\mathrm{wKGB}}$, converges to zero at long distances and to a finite value for short distance bins with low apparent densities, playing the role of pseudo-count. The smoothed angular potential, $\overline{f}_{\theta,p}^{\mathrm{wKGB}}(\theta \mid r;s)$, is set to zero when the observed frequency of the hydrogen bond orientation is equal to the average value. This angular term becomes zero for both long-and short-distance bins.

Once the potential is derived by assigning numerical values of the potential for all the bins as in Equation (8), it is evaluated by natural cubic spline interpolation at given values of the variables when it is applied to new protein structures. The numerical values of the potential and data used for deriving the potential are available as supplementary files at http://galaxy.seoklab.org/suppl/wkgb.html.

2.1.2. Prediction of bound water positions with wKGB

As an application of the wKGB potential, the positions of the water oxygen atoms for a given protein structure were predicted as follows. First, the total wKGB potential on the water position, $\mathbf{r}_w$, is defined as

$$\mathrm{F}^{\mathrm{wKGB}}(\mathbf{r}_w) = \sum_p \overline{f}_p^{\mathrm{wKGB}}(r_{pw}, \theta_{pw}; s_p), \tag{9}$$

which is the sum of the smoothed wKGB potential over the protein atoms within a cut-off distance of 9.75 Å. The variables r_{pw} and θ_{pw} are calculated as defined in **Figure 2.1** as r and θ, and the effective Born radius s_p is calculated by using the generalize Born solvation model FACTS from the input protein structure, as described above. The total wKGB potential is calculated on each grid point of a 3D grid box with 0.5 Å spacings around the protein structure. The grid point with the lowest total wKGB potential is then predicted as a water position. The grid point

with the next lowest total wKGB potential is identified, excluding the grid points within 2 Å from the previously predicted water positions. This step is repeated until the desired number of water positions is predicted or until the cutoff value of the total wKGB potential is reached.

The performance of the water position prediction algorithm was tested on 120 protein structures compiled by collecting X-ray crystal structures having resolution better than 1 Å, with 100-500 amino acids, and with mutual sequence identity less than 25%. The test set proteins also do not have a sequence identity more than 30% to any proteins used in deriving the wKGB potential. Only the crystal water positions with crystallographic B-factors less than 40 were considered to evaluate the prediction.

The two evaluation measures for the water prediction considered in this study were the coverage of the crystal water positions and RMSD of the predicted positions from those of the crystal water oxygen atoms. The coverage was defined as the fraction of crystal water positions within 1 Å from the predicted water positions among all the crystal water positions. The coverage and RMSD were calculated after matching at most one crystal water to each predicted water. The coverage and the RMSD were examined at different N_{pred}, the number of predicted water positions, which was set to $N_{\text{pred}} = nN_{\text{cryst}}$, where N_{cryst} is the number of crystal water molecules, varying n from 1 to 10.

The performance of the current method called as GalaxyWater-wKGB was compared with those of two other existing methods FoldX[43] and 3D-RISM[41, 42] on the 98 protein targets for which FoldX successfully generated results out of the 120 test set proteins. The default option for FoldX was used, and Placevent[48] with the default option was used for 3D-RISM. The 3D-RISM calculation was performed using the "rism.snglpnt" program of the AmberTools simulation suite[49]. SPC/E

water[50] with the ff99SB protein force field and the ions94 ion parameter was used. For the solvent, 55.5 M of water and 0.005 M of sodium ions and chloride ions were used. KH was used for the closure of the integral equation. Grid spacings of 0.5 Å, minimum buffer of 14 Å between solute and grid box, and 10,000 iterations with convergence criteria of 10^{-5} were used. All the calculations were performed on an Intel Xeon E5-2620 CPU.

2.2. Performance of GalaxyWater-wKGB

2.2.1. Characteristics of wKGB potential

The radial part of the smoothened wKGB potential, $\overline{f}_{\mathrm{r},p}^{\mathrm{wKGB}}(r;s)$, is shown for different protein atom types in **Figure 2.2** and Supplementary **Figure S2**. The minima of the potential near 3 Å for the polar protein atoms [**Figure 2.2** (a)-(d)] indicates the hydrogen bond interactions at that distance. The depth of the minimum near 3 Å is shallower for neutral N or O atoms, such as main chain N/O, amide N/O, or hydroxyl O [**Figures 2.2** (a) and (b), Supplementary **Figure S2**], than that for charged N or O [**Figures 2.2** (c) and (d), Supplementary **Figure S2**]. The minimum of the potential for the lysine side chain N atoms is particularly deep with three hydrogen atoms that can form hydrogen bonds. The minima for the N atoms acting as hydrogen bond donors are shallower than those for the O atoms acting as hydrogen bond acceptors (Supplementary **Figure S2**), which is consistent with the weaker hydrogen bond energy of N than that of O[51].

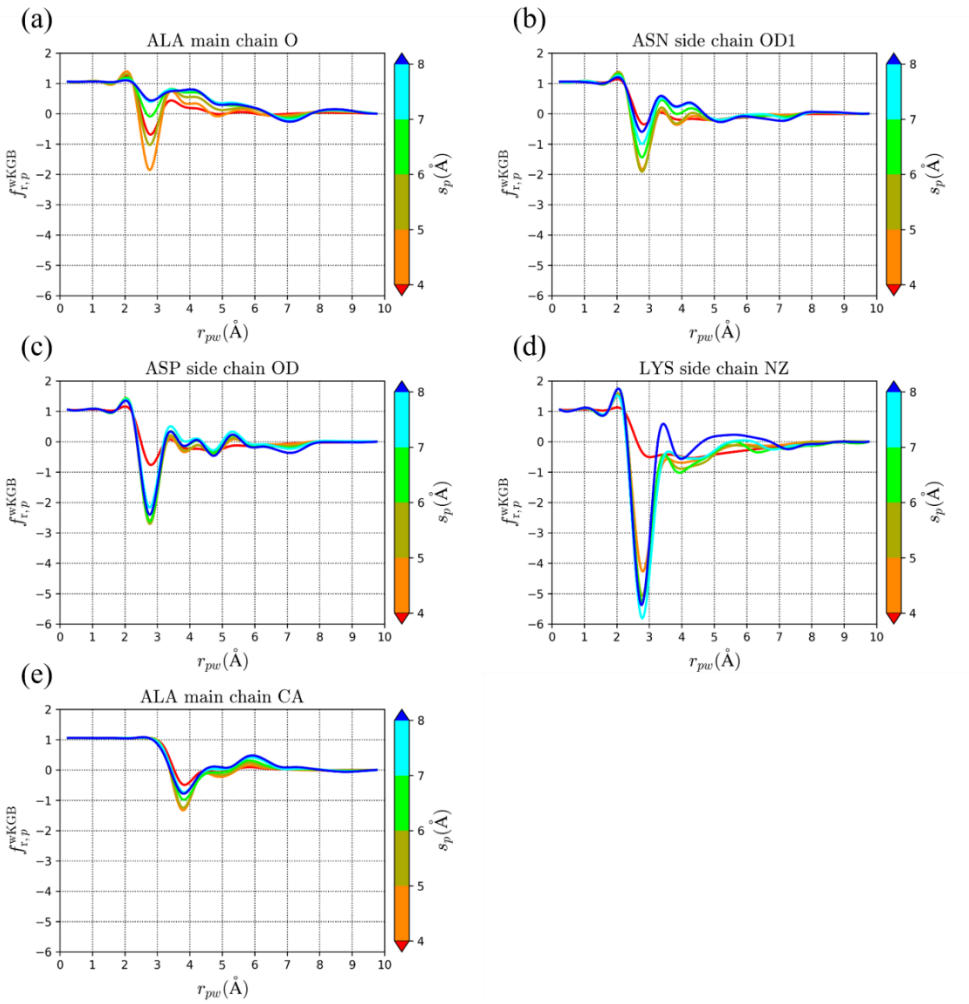


Figure 2.2. Radial part of smoothed wKGB potential for (a) ALA main chain O, (b) ASN side chain OD1, (c) ASP side chain OD, (d) LYS side chain NZ, and (e) ALA main chain CA.

The solvation state of the protein atoms strongly affects the depth of the potential minimum near 3 Å. The potentials for the neutral N or O atoms show stronger minima when the atoms are slightly exposed (orange lines) than when they are more buried (green, cyan, and blue lines) [**Figures 2.2** (a) and (b)]. This can be interpreted as a result of the fact that neutral polar atoms tend to form hydrogen

bonds with other protein atoms when buried in the protein structure. This is in particular for the main chain atoms that form hydrogen bonds with other main chain atoms to generate secondary structures. The potentials for the charged N or O atoms show much stronger minima when the atoms are buried (blue lines) than for the neutral N or O atoms [**Figures 2.2** (c) and (d)]. This fact may be explained by the stronger Coulomb interactions of water molecules with charged atoms in a less shielded, buried condition.

Although carbon atoms are not expected to form hydrogen bond interactions with water molecules, the potentials for C atoms still show weak minima at distances longer than 3 Å [**Figure 2.2** (e)]. This type of minima originates from water molecules interacting with nearby polar atoms in the protein structure. Such minima may be considered as artificial.

The angular part of the smoothed wKGB potential, $\overline{f}_{\theta,p}^{\mathrm{wKGB}}(\theta \mid r;s)$, is shown for a moderately exposed state (effective Born radius 4-5 Å) in **Figure 2.3** as a function of r, instead of θ, for convenience. The angular term shows the strongest minimum near 3 Å when θ approaches the ideal hydrogen bond angle, which depends on the atom type. For a main chain N atom whose ideal θ = 180° because the dipole vector of N is along the N-H bond, the angle bin having >131° shows the strongest potential [**Figure 2.3** (a)]. For an sp^2 hybridized atom whose ideal θ = 120°, the angle bin of 109°-131° is favored the most [**Figure 2.3** (b)]. For an sp^3 hybridized atom whose ideal θ is close to 109.5°, two angle bins of 109°-131° and 90°-109° show the strongest potential minima [**Figure 2.3** (c)]. Angular part of the potential for other atom types are provided in **Figure S3**.

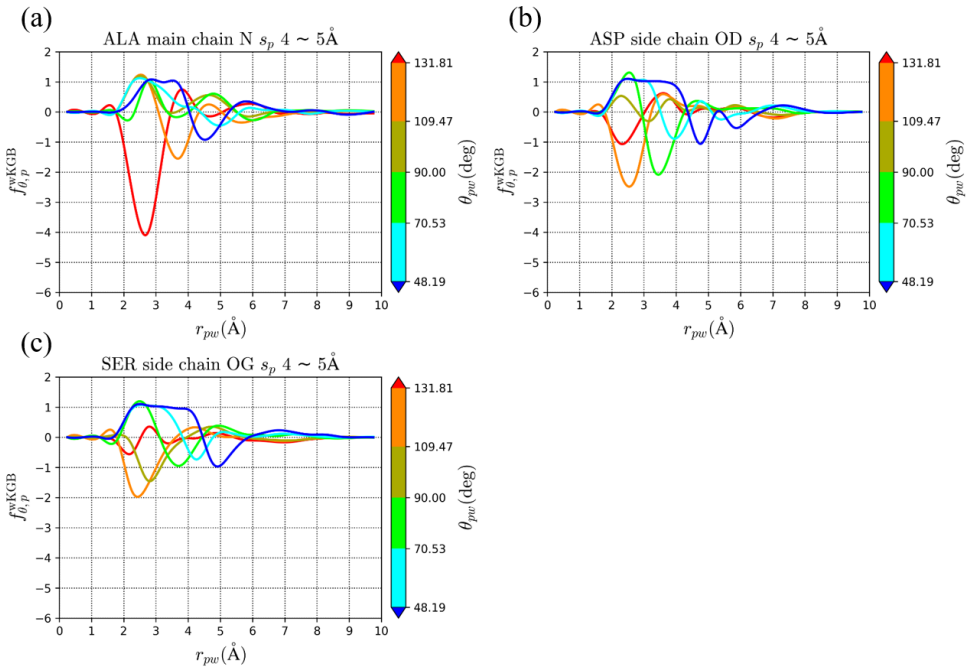


Figure 2.3. Angular part of smoothed wKGB potential for (a) ALA main chain N, (b) ASP side chain OD, and (c) SER side chain OG. Potential is drawn for s_p of 4-5 Å, and different colors are used for different θ_{pw} bins.

If the solvation state of protein atoms is not considered in the derivation of the potential by excluding the dependence on the solvent state represented by the effective Born radius, *s*, the resulting potential will be an averaged potential over the solvation states. This lacks the important information on the strong dependency of the potential on the degree of solvent accessibility, as can be seen from the comparison of **Figures 2.4** (b) and (a).

If the inherent occupations of the water molecules in the vacant space in the protein structure are not considered when deriving the potential, the potential for the more exposed states would be much weaker, as shown in **Figure 2.4** (c). The potential obtained without considering the dependence on the solvation state, *s*, and

the water molecules in the vacant space is also shown in **Figure 2.4** (d). Predictions of the water positions with the potentials without considering the solvation state and/or the vacant space are less successful, as discussed in the next subsection.

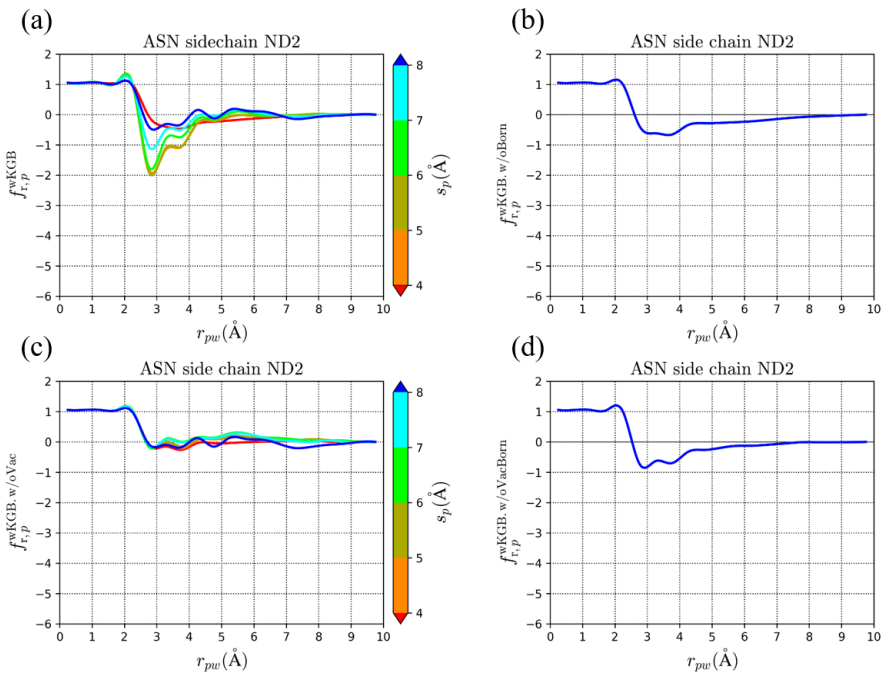


Figure 2.4. Effect of considering solvation state and water occupation in vacant space on resulting wKGB potential, illustrated for ASN side chain ND2. (a) current wKGB potential, (b) potential derived without consideration of dependence on solvation state *s* (represented as w/o Born), (c) potential derived without consideration of water occupation in vacant space (represented as w/o Vac), and (d) potential derived without consideration of solvation state and water occupation in vacant space (represented as w/o Vac Born).

2.2.2. Results of water site prediction

The prediction results of the water sites on the crystal structures of the 120 test proteins obtained using GalaxyWater-wKGB are presented in **Figure 2.5**. The crystal water sites recovered by the prediction converged to 80% as a larger number of water positions were predicted [**Figure 2.5** (a)]. The coverage increased to 96% if a distance cutoff of 1.5 Å was used as a criterion of recovery instead of 1 Å [**Figure S4** (a)]. The RMSD of the predicted water positions converged to 0.8 Å [**Figure 2.5** (b)]. As can be seen from the figures, without the consideration of either the dependence on the solvation state (green lines) or the water occupation in the vacant space (orange lines), the prediction performance deteriorates. Considering the solvation state has a larger effect than including the vacant space. Random rotation and translation of the protein structure in the 3D grid box did not affect the performance, showing deviations of < 4% in the coverage (See **Table S2** for detailed results). The water sites predicted for two protein targets using the wKGB potential and the potential derived without considering the solvation states (wKGB.w/o Born) are compared in **Figure 2.6**. The figure shows that using the potential not considering the solvation state, the water sites are less precise, with over-prediction of the sites on the protein surface.

In the actual prediction of water sites, the number of crystal waters shown on the abscissa of the figures is not known. A cutoff value of the wKGB potential was used instead to determine the number of water sites to predict. The dependence of the prediction performance on the potential cutoff value is summarized in Supplementary **Table S3**.

The prediction performance of GalaxyWater-wKGB is compared to those of FoldX and 3D-RISM on 92 proteins, as shown in **Figure 2.7**. At a small number of predicted water sites, GalaxyWater-wKGB performs comparably to 3D-RISM and

better than FoldX in terms of the coverage and slightly worse than 3D-RISM in terms of RMSD on the crystal structures [**Figures 2.7** (a) and (b)]. At a large number of predicted sites, GalaxyWater-wKGB performs better than 3D-RISM.

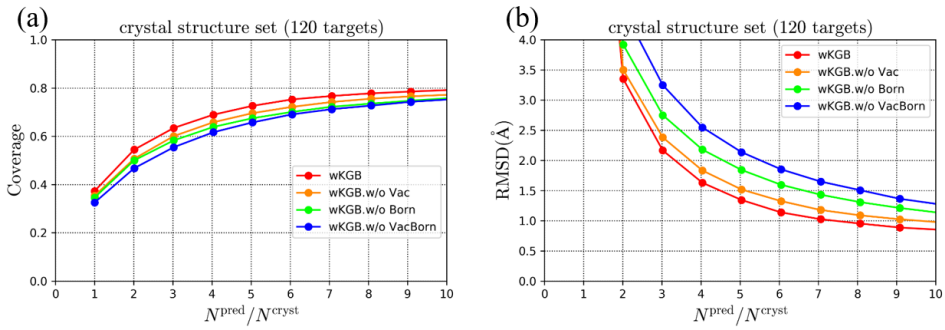


Figure 2.5. Prediction results of water sites on crystal structure sets. X-axis is number of predicted water sites (N_{pred}) divided by number of well-resolved crystallographic water molecules (N_{cryst}). (a) average coverage and (b) average RMSD for crystal structure set. Results for potential derived without consideration of water occupation in vacant space (represented as w/o Vac), without consideration of dependence on solvation state *s* (represented as w/o Born), and without consideration of both solvation state and water occupation in vacant space (represented as w/o Vac Born) are shown in different colors for comparison.

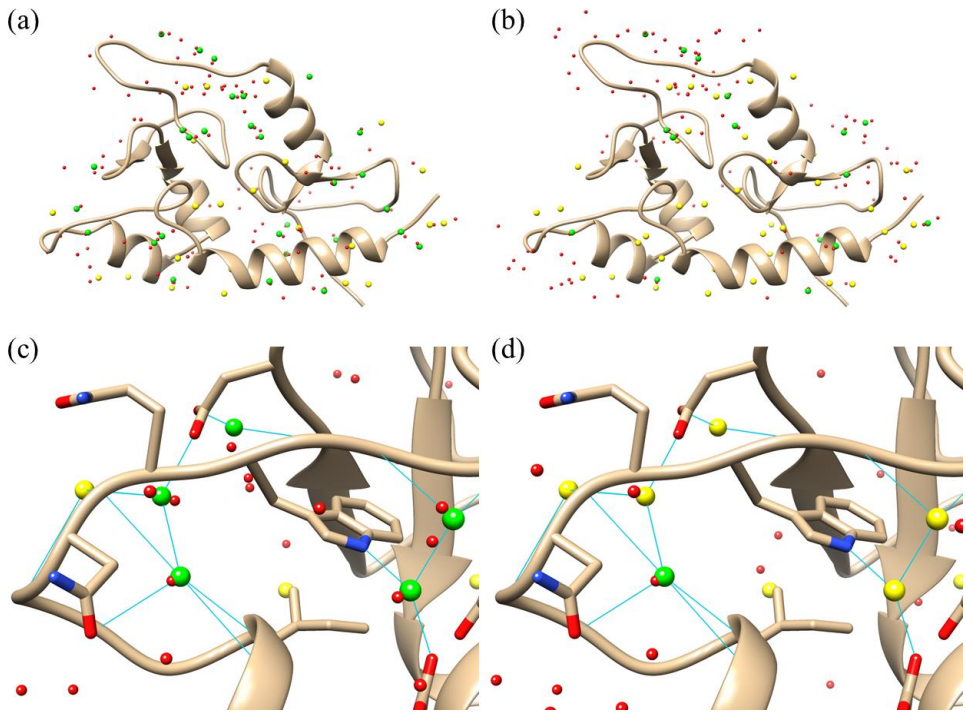


Figure 2.6. Example case of water site prediction (PDB ID: 3QL9) when twice the number of crystallographic water molecules are predicted. Red spheres represent predicted water sites, and green and yellow spheres crystallographic water sites. Green spheres indicate crystallographic water molecules with corresponding predicted water sites within 1 Å, and yellow spheres indicate crystallographic water molecules for which no predicted water site is found within 1 Å. Blue lines indicate hydrogen bonds between crystallographic water molecules and neighboring protein atoms. (a) and (c) show water site prediction results with wKGB potential, and (b) and (d) with potential derived without considering dependence on solvation state. Region with pronounced difference between two methods is enlarged in (c) and (d).

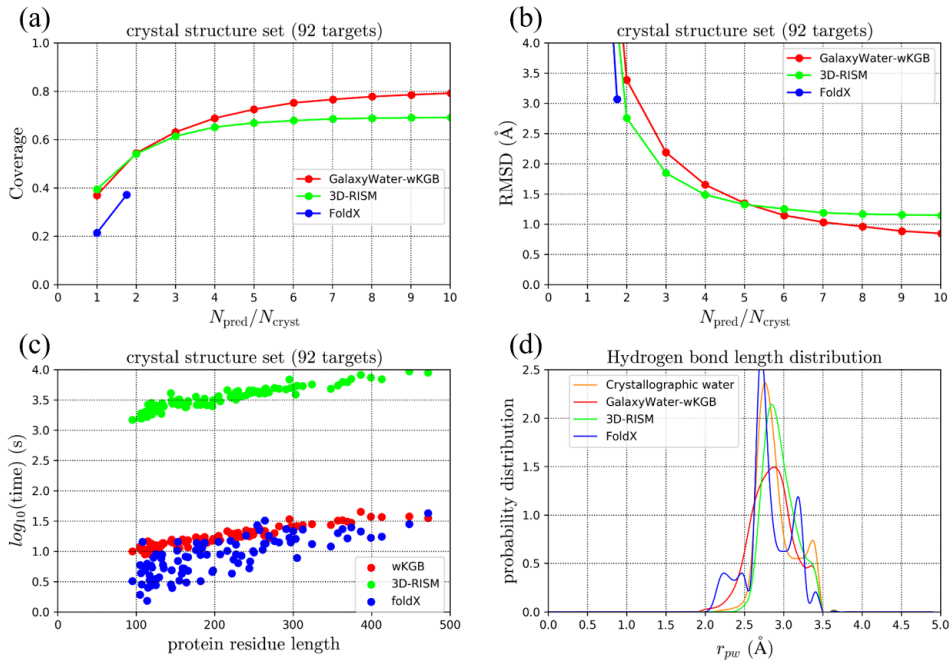


Figure 2.7. Coverage of crystallographic water positions and RMSD of predicted positions for crystal structure set is shown in (a) and (b). X-axis is number of predicted water sites (N_{pred}) divided by number of well-resolved crystallographic water molecules (N_{cryst}) for both (a) and (b). Computational time consumed by prediction is presented in (c) as function of protein size. Probability distribution of hydrogen bond length (r_{pw}) between predicted water and protein atom is presented in (d). Hydrogen bonds are defined here as interactions between water O atom and protein N or O atoms with $\theta_{pw} > 100°$ and $r_{pw} < 3.5$ Å.

3. Prediction of Water Positions on Protein Structure using 3D-CNN

In previous chapter, statistical potential based method showed better water site prediction performance than 3D-RISM. However, statistical potential could make inaccurate predictions due to limited degree of freedom of the potential function used to predict water. For the freedom of water prediction method, Deep learning[52] was applied in this chapter. Deep learning is a machine learning method that enables learning about complex functions through a large number of artificial neural network layers, of which the Convolutional Neural Network (CNN)[52] is a type of network used in deep learning and is mainly used in image detection. CNN is characterized by recognizing features for narrow areas in shallow layers, being able to recognize the overall patterns that make up the image as the layer deepens, and being able to recognize features that make up the image with a relatively small number of parameters using the solution kernel. These CNN networks are the best way to recognize images, and methods using CNN such as GoogLeNet have won ImageNet Large Scale Visual Recognition Challenge of 2014[53]. The 3D structure of a protein can also be regarded as a kind of 3D image, and it has been suggested using CNN to recognize the structure of a protein that is likely to interact with a particular ligand molecule, indicating that it is possible to consider the structure of a protein as a kind of image and analyze its structure and function[54]. GalaxyWater-CNN is a 3D-CNN network that could recognize patterns of protein structures that could accommodate water molecules by considering the structure of proteins as a kind of 3D image, thereby predicting the location of water interacting with proteins. In this thesis, the performance of GalaxyWater-CNN was tested by calculating the coverage of the crystal water molecule position from the water site prediction for the given protein from the high resolution PDB set and calculating the coverage of

the protein-protein bridging crystal water molecule position from the water site prediction for the given protein from protein-protein complex set. GalaxyWater-CNN showed coverage of 75% and shows the coverage of 78% for inter-protein bridging water prediction, when three times of the bridging water in the crystal is predicted. Additionally, GalaxyWater-CNN which trained on protein-ligand complex set was tested by calculating the coverage of the ligand-neighboring crystal water molecule position from the water site prediction for the given protein from protein-ligand complex set, which showed 81% coverage when three times of the bridging water in the crystal is predicted.

3.1. Methods for GalaxyWater-CNN

3.1.1. Overview of the overall method

GalaxyWater-CNN places water molecules on the surface of a protein chain, a protein-protein complex, or a protein-compound complex using a CNN model that generates a water distribution map, as illustrated in **Figure 3.1**. The CNN takes a protein or complex structure as input and generates a water distribution map in a 3D grid box surrounding the structure. The positions of water molecules on the protein surface are predicted by locating high-probability regions on the map.

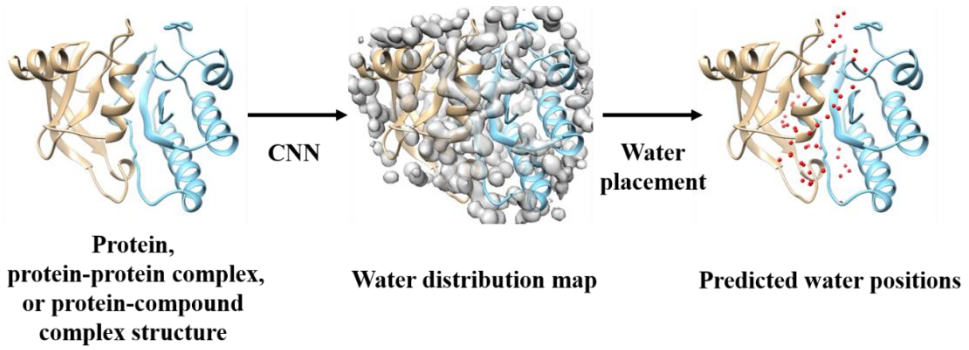


Figure 3.1. GalaxyWater-CNN places water molecules around a given structure of a protein, a protein-protein complex, or a protein-compound complex by locating high-probability regions of a water distribution map generated by a convolutional neural network.

3.1.2. The CNN architecture

The CNN architecture is shown in **Figure 3.2**. The input of the network is the atomic distribution of n channels on a 3D $N \times N \times N$ cubic grid box represented in the dimension of $n \times N \times N \times N$. The spacing between the grid points is set to 0.5 Å. For the box size, N = 32 (volume = 16^3 Å^3) is used when training the network, and a larger size of N = 64 (volume = 32^3 Å^3) is used for actual prediction to avoid memory problems during training. To generate a water map on a protein-compound complex, $n = 8$, corresponding to eight atom types (i.e., C, N, O, S, P, halogen, metal, and others) is used. For a protein chain or a protein-protein complex, $n = 16$, corresponding to C, N, O, and S channels and an additional 12 channels that describe more detailed protein atom types, as listed **in Table S4**. Atomic distributions are generated from the atomic coordinates of the input structure. For each atom in the input structure, the contribution to the atomic distribution at a grid point for each channel is given by a shifted Gaussian function $A(d,r)$ as[54]

$$A(d,r)=\begin{cases} e^{-2d^2/r^2} & 0 \le d < r \\ \dfrac{4}{e^2 r^2} d^2 - \dfrac{12}{e^2 r} d + \dfrac{9}{e^2} & r \le d < 1.5r \\ 0 & d \ge 1.5r \end{cases}, \tag{10}$$

where d is the distance from the atom center to the grid point, and r is the atomic radius.

As shown in **Figure 3.2 (a)**, the $n \times N \times N \times N$ input is first increased to $64 \times N \times N \times N$ by $1 \times 1 \times 1$ convolution and passes through 10 residual blocks, as explained next. The network is then split into two branches. One branch generates a coarse water map that is used to calculate a cross-entropy loss function, Loss1, with the definition of a larger water radius of $r_1 = 2.28$ Å (150 % of the more precise radius, 1.52 Å). The other branch passes through an additional 10 residual blocks to generate a fine water map, which gives another cross-entropy loss function, Loss2, with a water radius of $r_2 = 1.52$ Å. The loss function is explained in more detail in the next subsection. A fine water map is used to place water molecules.

The residual block, shown in **Figure 3.2 (b)**, is a convolutional residual neural network that uses the atrous convolutions[39] to effectively process information encompassing a wider space with fewer layers. After $3 \times 3 \times 3$ convolutions of different dilations of one and two with zero-paddings of one and two, respectively, the two grids were concatenated and reduced to the original size of $64 \times N \times N \times N$ by $1 \times 1 \times 1$ convolution. After batch normalization, followed by the addition of input grid values, an ELU activation layer is applied to generate an output grid of the same dimension: $64 \times N \times N \times N$.

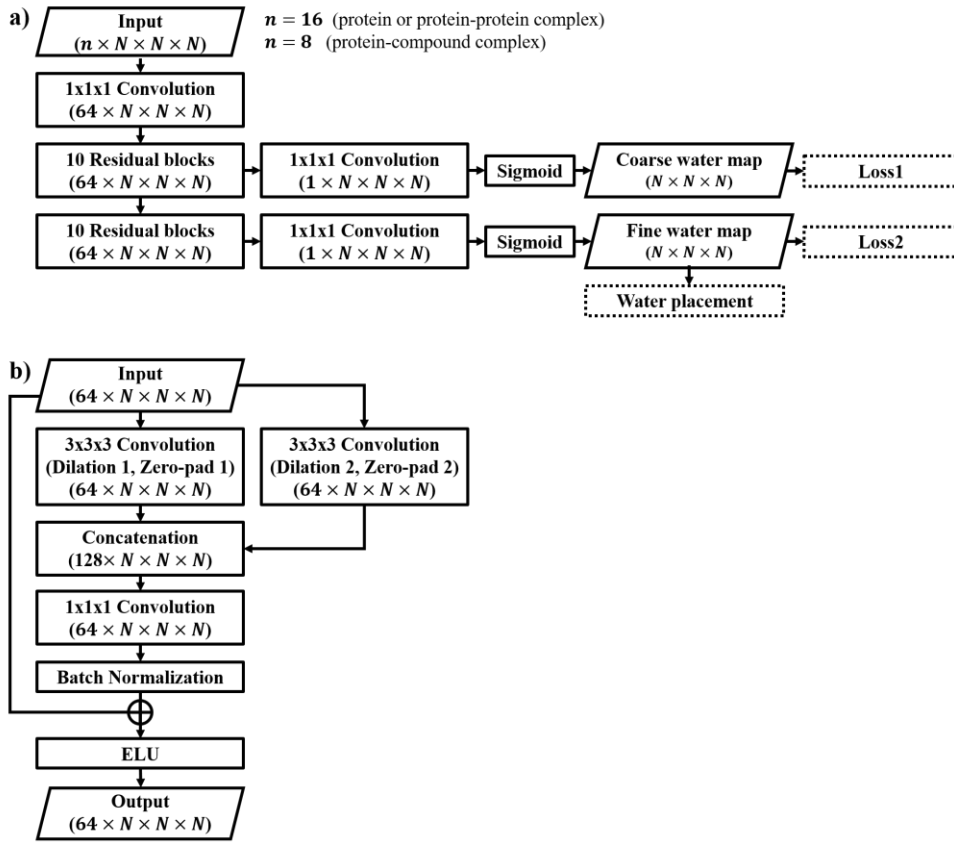


Figure 3.2. Network structure of GalaxyWater-CNN. a) The overall network structure, where *n* denotes the number of channels representing atom types and *N* is the number of grid points along each spatial dimension. b) Structure of the residual block in a).

3.1.3. Training of the neural network

3.1.3.1. Training set proteins and complexes

To train and evaluate the network for proteins and protein-protein complexes, a database of non-redundant, high-resolution X-ray crystal structures were constructed. Protein structures having a crystallographic resolution better than 2 Å and R-free values lower than 0.3 with mutual sequence identity less than 25 % with a sequence length between 100 and 500 were collected using PISCES protein sequence culling server[47] on July 18th, 2019. The symmetry operators given in the Protein Data Bank (PDB) structure files were used to generate protein and water oxygen atoms surrounding the center protein.

The interactions present in the crystal environments were considered explicitly by including all the surrounding proteins and water atoms during training. For convenience, only those proteins and complexes having fewer than 10 symmetry operators, no nucleic acids, no alternate structures, and no missing residues were included for method training and test. Only well-resolved water oxygen atoms having a B-factor < 40 were considered. Additionally, only those having the number of water molecules with B-factor < 40 between 5 and 20 % of that of protein atoms were included.

The total number of compiled protein and complex structures was 312, of which 160 structures were used as a training set, and 152 structures were used as a validation set. Separate test sets were employed for comparison with the previous methods, as explained next.

For protein-compound complexes, another set was curated from PDBBind[55] v2019 refined set. It is a set of non-redundant, high-resolution X-ray crystal structures with resolution better than 2 Å, mutual sequence identity of proteins less

than 30 %, and Tanimoto similarity of compounds less than 50 %. Only those having water molecules between 5 and 20 % of the number of protein atoms were considered. Among the 1,189 protein-complex structures, 792 were used for training, and the rest were used for independent evaluation of the method.

3.1.3.2. Training method

The parameters of the network are optimized using the training loss, defined as follows. Each of the cross-entropy loss functions, Loss1 and Loss2 (see **Figure 3.2 (a)**), is calculated from a predicted and a reference water map. Two kinds of reference maps, coarse and fine, are generated for Loss1 and Loss2, respectively, from the coordinates of well-resolved water oxygen atoms (B-factor < 40) for the training set structures described above using two different definitions for hard-sphere water radius, 2.28 Å and 1.52 Å, respectively. The total loss function is defined as the average of the cross-entropy terms with a weight of 1 for the grid points without any assigned water oxygen, and 20 and 64 for the grid points with assigned water oxygen for the coarse and fine maps, respectively.

Pytorch is used as a platform for building and training the neural network. The Adam optimizer is used with a learning rate of 0.0001 for the parameter optimization. To minimize the dependence on rotation and translation of the input structure, the structures are randomly rotated and cropped during training. To account for the interactions with adjacent symmetric units in the crystal environment, nearby protein and water oxygen atoms placed by symmetric operations are included during training.

3.1.4. Placement of water molecules from the water map

Water molecules were placed sequentially on the 3D box of $N = 64$ (volume = 32^3 Å^3) by locating high-score regions in the water distribution map, as outlined in

Figure 3.3. First, the generated water map is convoluted with a shifted Gaussian kernel (See **Equation 10**) with $r =$ 1.52 Å to remove possible noise. During the convolution, a reflecting padding was applied to the boundary of the water distribution map to keep the size of the convoluted map the same as that of the distribution map. Water oxygen atoms are then iteratively placed on the highest distribution grid point after evacuating neighboring grid points within 2.28 Å.

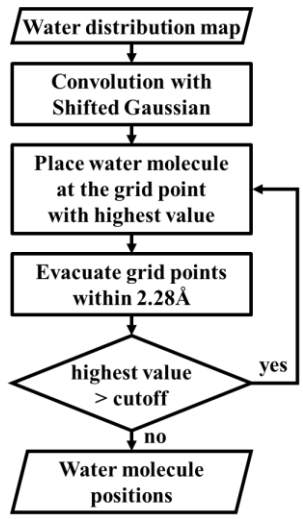


Figure 3.3. Water placement method on the water distribution map generated by CNN.

For proteins larger than the size of the grid box (32^3 Å^3), a larger box covering all protein atoms are sliced into boxes of 32^3 Å^3 shifted by 24 Å in each dimension. To account for possible inaccuracy in the boundary, only the internal regions of the water maps of size 24^3 Å^3 were merged before water positions are predicted.

3.1.5. Methods for performance comparison

3.1.5.1. Evaluation measures

To evaluate the performance of water position prediction, the extent to which the

predicted water recovers the crystal waters in the PDB structure (coverage) and the deviation of the predicted positions from the crystal water positions (RMSD) are considered. Only the crystal waters having a B-factor < 40 with no contacts with neighboring symmetric units within 4 Å are considered. The coverage and RMSD are calculated after matching at most one crystal water to each predicted water within a distance cutoff of 1 Å. The coverage is defined as the fraction of matched crystal water positions among all crystal water positions. The coverage and RMSD are examined at different N_{pred}. The number of predicted water positions is set to $N_{pred} = N_{cryst}$, where N_{cryst} is the number of crystal water molecules, which varies from 1 to 10.

In actual predictions, the number of crystal water molecules is not known. Therefore, in the web server described below, a different parameter, a cut-off value for the water map probability score, is used instead of n to control the number of predicted water molecules.

3.1.5.2. Test sets

GalaxyWater-CNN was tested on a single-protein test set, protein-protein complex test set, and protein-compound complex test set for placing water on single protein chains, protein-protein interfaces, and protein-compound binding sites, respectively. The test sets do not overlap with the training sets.

The single-protein test set of high-resolution crystal structures consists of 120 X-ray crystal structures having a resolution better than 1 Å, released before March 14th, 2015. This set was also used as a benchmark set in **Chapter 2**. For comparison with other methods, results were analyzed for 92 structures out of 120 structures after excluding the proteins for which water prediction failed with FoldX.

The protein-protein complex test set comprises X-ray crystal structures with

resolution better than 2.5 Å with residues between 50 and 450 and interface residues between 30 and 500 amino acids, which was constructed from Propairs[56] prepared on April 6th, 2018. Only those with no disorder and no ligand molecules at the binding interface and with more than four bridging water molecules at the interface are considered to evaluate the performance of predicting interface bridging water molecules. Bridging waters are defined as those with a B-factor <40 and those that form hydrogen bonds with both the receptor and ligand proteins. The hydrogen bonds are assigned to each water molecule with any neighboring hydrogen donor/acceptor heavy atom within 3.5 Å, and the angle made by the water oxygen atom, the hydrogen donor/acceptor atom, and the atom bonded to the donor/acceptor atom larger than 100°. For predicted waters, a relaxed criterion of 4.5 Å and 80° is used to match the crystal bridging waters.

The protein-compound complex test set has 397 non-redundant, high-resolution X-ray crystal protein-compound complex structures curated from the PDBBind refined set, as explained above. Crystal water molecules within 4 Å of the compound atoms are considered for performance evaluation. For predicted waters, a criterion of 5 Å is used to match the crystal waters near the compounds.

3.1.5.3. Running other methods for comparison

The Performances of GalaxyWater-CNN was compared to those of GalaxyWater-wKGB[40], 3D-RISM[41, 42], and FoldX[43]. For water placement on protein-compound complexes, only 3D-RISM was employed for comparison because the other two methods cannot deal with protein-compound complexes.

Prediction by 3D-RISM was carried out using the “rism.snglpnt” program of the AmberTools simulation suite[49]. For the force field, SPC/E for water[50], ff99SB for protein, ions94 for ions, and General AMBER Force Field (GAFF) for compounds

were used. Antechamber was used to set the charges for the compounds with AM1-BCC partial charge model. Protonation of compounds followed the protonation state given in the PDBBind refined set. For the solvent, 55.5 M of water and 0.005 M of sodium ions and chloride ions were used. KH was used for the closure of the integral equation. Grid spacings of 0.5 Å, minimum buffer of 14 Å between solute and grid box, and 10,000 iterations with convergence criteria of 10^{-5} were used. FoldX and GalaxyWater-wKGB were performed with the default options.

All CPU calculations were performed on Intel Xeon E5-2650 CPU, and GPU calculations were performed on GeForce GTX 1080 Ti.

3.2. Performance of GalaxyWater-CNN

3.2.1. Results of network training

The network that generates a water distribution map for a protein or a protein-protein complex structure was trained and validated on the single-protein sets. The training and validation losses plotted in **Figure 3.4** show similar trends with the increasing number of epochs with no indication of overtraining.

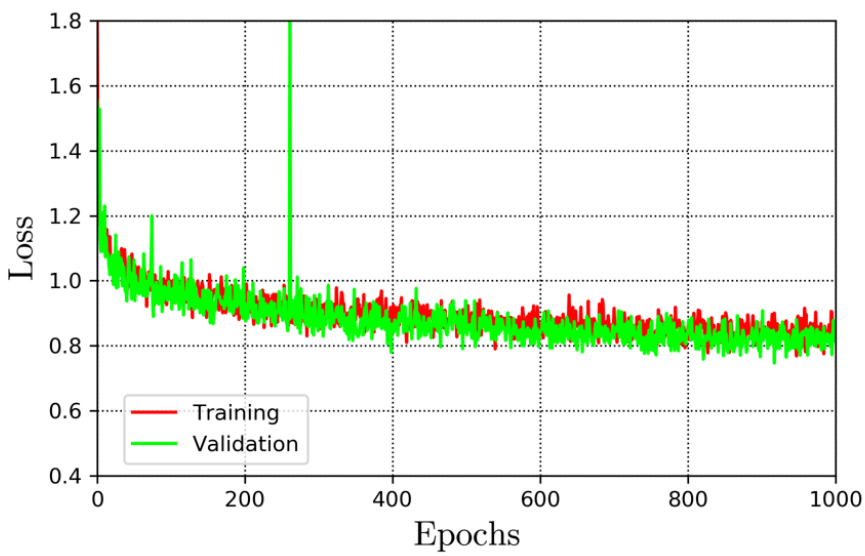


Figure 3.4. Evolution of the loss values for the single-protein training and validation sets with respect to the number of epochs.

When the positions of the water molecules were predicted using the trained network, the prediction results were very close between the training and validation sets. For example, water molecules were predicted with the average coverage of (73%, 82%) and (75, 83%) when the number of predicted waters, N_{pred}, was (3, 10)

times that of the crystallographic waters, N_{cryst}, for the training set and the validation set, respectively.

Random rotation and translation of the protein structure in the 3D grid box did not affect the performance, showing deviations of < 3% in the coverage (See **Table S5** for detailed results).

The contribution of distinctive features of the network to the prediction performance was analyzed by examining the prediction results when each feature was removed or modified, as shown in **Supplementary Figure S5**. The most significant impact on the performance of the network was achieved using the dilated convolution kernel, followed by using 20 residual layers. For example, water molecules were predicted with an average coverage of (68 %, 78 %) and (70 %, 79 %) for N_{pred}/N_{cryst} = (3, 10) when the dilated convolution layers were removed from the residual block and when only half of the residual layers were used, respectively, compared to (75 %, 83 %) for the full network. Additional atom channels of detailed functional groups and using the auxiliary loss (Loss 1) showed a relatively small impact, but there was some when small numbers of water molecules (< $2N_{cryst}$) were predicted.

3.2.2. Results on the single-protein test set

The prediction performance of GalaxyWater-CNN on an independent single-protein test set was compared with those of GalaxyWater-wKGB, 3D-RISM, and FoldX. The test set consists of proteins with sequence identity < 25 % from those used for training and validating the GalaxyWater-CNN. As can be seen from the performance comparison presented in **Figure 3.5 (a)**, GalaxyWater-CNN can predict more crystallographic water molecules than other methods with an average coverage of (75 %, 86 %) for N_{pred}/N_{cryst} = (3, 10), compared with (63 %, 79 %) and (61 %, 69 %) for GalaxyWater-wKGB and 3D-RISM, respectively. FoldX generated only a small number of water molecules, and its performance was worse

than that of the other methods.

GalaxyWater-CNN can predict the crystallographic water molecules more precisely when larger number of waters are predicted with average RMSD of (1.99 Å, 0.78 Å) for N_{pred}/N_{cryst} = (3, 10), compared with (2.19 Å, 0.85 Å) and (1.84 Å, 1.15 Å) for GalaxyWater-wKGB and 3D-RISM, respectively, as shown in **Figure 3.5 (b)**.

The average calculation time for GalaxyWater-CNN was 83 s on a GPGPU using CUDA and 2,308 s on a CPU, compared with 18 s, 3,641 s, and 11 s on a CPU for GalaxyWater-wKGB, 3D-RISM, and FoldX, respectively. **Figure 3.5 (c)** shows the dependence of the calculation time on protein size.

Detailed analysis showed that GalaxyWater-CNN predicts water positions more precisely than GalaxyWater-wKGB when hydrogen bond networks exist, as illustrated in **Figure 3.6**. In **Figure 3.6 (c)** and **(d)**, GalaxyWater-CNN predicted three crystal water molecules forming a hydrogen bond network correctly, whereas GalaxyWater-wKGB did not. This demonstrates the limitation of the statistical potential used in GalaxyWater-wKGB. In a statistical potential like wKGB, nearby atoms chemically bonded to hydrogen acceptors or donors tend to show preferable potential because of their proximity to the observed hydrogen bonds. This artifact may blur the total potential map between the protein and water, which sometimes decreases the water site prediction accuracy.

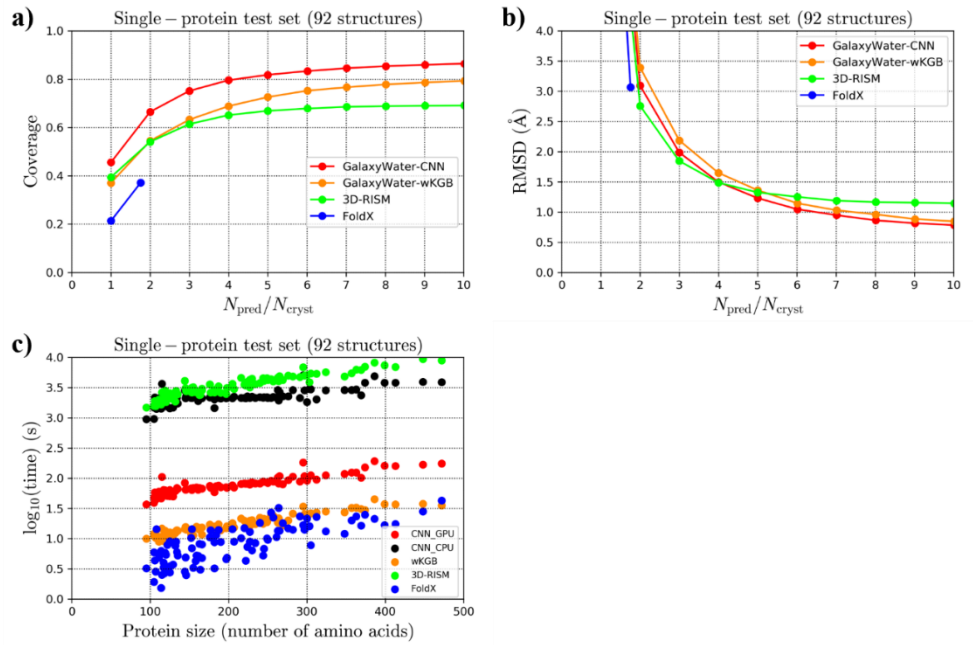


Figure 3.5. Performance comparison of GalaxyWater-CNN, GalaxyWater-wKGB, 3D-RISM, and FoldX on the single-protein test set. a) Average coverage and b) average RMSD of the predicted water molecules vs. the number of predicted waters (N_{pred} / N_{cryst}); c) Computation time in log-scale versus the protein size. CNN_GPU and CNN_GPU refer to GalaxyWater-CNN using GPU and CPU, respectively.

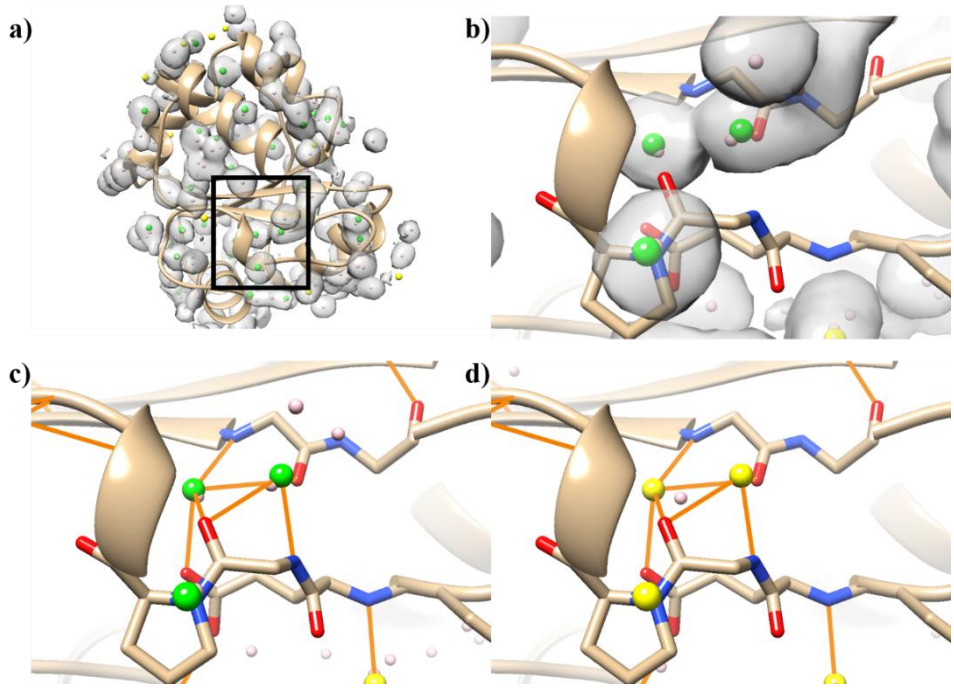


Figure 3.6. An example case PDB ID: 2FWH which emphasizes a case in which GalaxyWater-CNN predicts water sites precisely, as in (a), (b), and (c), whereas the potential-based method GalaxyWater-wKGB does not as in (d), for N_{pred}/N_{cryst} = 3. The gray contour in (a) and that magnified in (b) show water distribution map with probability score > 0.9. Green and pink spheres represent predicted water sites with and without corresponding crystallographic waters within 1 Å. Yellow spheres represent crystallographic water sites without corresponding predicted water sites within 1 Å. Orange lines show hydrogen bonds.

3.2.3. Results on the protein-protein complex test set

A comparison of the prediction performance of the bridging water position at the protein-protein interface on the protein-protein complex test set (151 structures) is shown in **Figure 3.7**. GalaxyWater-CNN shows a coverage of 78 %, compared with 68, 57, and 46 % for GalaxyWater-wKGB, 3D-RISM, and FoldX, respectively, at N_{pred}/N_{cryst} = 3. The coverage converged to 90 and 82 % for GalaxyWater-CNN and GalaxyWater-wKGB, respectively, for large N_{pred}/N_{cryst}. GalaxyWater-CNN also showed enhanced performance in terms of average RMSD of 0.90 Å, compared with 1.25, 1.20, and 1.12 Å for GalaxyWater-wKGB, 3D-RISM, and FoldX, respectively. It is notable that protein-protein interface waters were predicted with high performance using the water distribution map generated by the same network trained on a single-protein structure set.

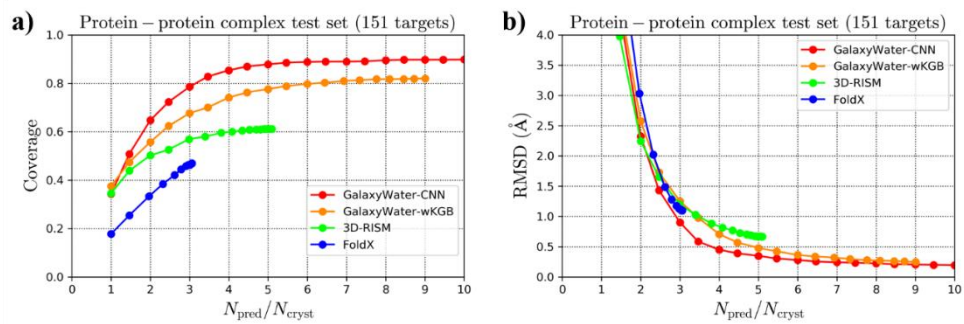


Figure 3.7. Performance comparison of GalaxyWater-CNN, GalaxyWater-wKGB, 3D-RISM, and FoldX on the protein-protein complex test set. a) Average coverage; b) Average RMSD of the predicted crystallographic waters bridging in the protein-protein interface.

GalaxyWater-CNN also showed higher performance in predicting bridging waters in protein-protein interfaces (**Figure 3.7**) than waters on protein surfaces (**Figure 3.5**) in terms of absolute measures (coverage/RMSD of 90 %/0.19 Å compared with 86 %/0.78 Å at N_{pred}/N_{cryst} = 10). This is consistent with the observation made on the single-protein chain set in which water molecules making hydrogen bond networks were more precisely predicted. An example of a water distribution map and predicted water sites is shown in **Figure 3.8**. In the figure, GalaxyWater-CNN correctly predicted two crystal water molecules forming hydrogen-bond bridges with main-chain protein atoms, whereas GalaxyWater-wKGB failed. This failure of GalaxyWater-wKGB may be ascribed to an artifact of using statistical potential, as explained regarding **Figure 3.6**.

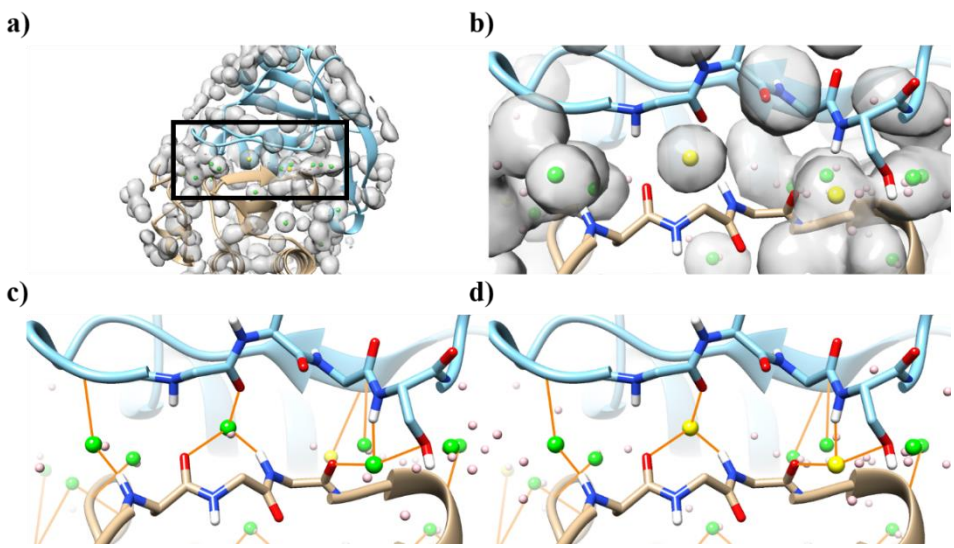


Figure 3.8. Example case of predicting protein interface bridging water molecules (PDB ID: 2FHZ) at N_{pred}/N_{cryst} = 3. Gray contour in (a) and that magnified in (b) show GalaxyWater-CNN water distribution map with probability > 0.9. Green and pink spheres represent predicted water sites with and without corresponding

crystallographic waters within 1 Å. Yellow spheres represent crystallographic water sites without corresponding predicted water sites within 1 Å. Orange lines show hydrogen bonds. Prediction by GalaxyWater-CNN shown in (c) recovers more bridging waters than that by GalaxyWater-wKGB shown in (d).

3.2.4. Result on the protein-compound complex set

A separate CNN that generates a water distribution map on the surface of protein-compound complexes was trained in the same manner as that for single-protein chains. The performance of using this network for predicting water sites in the compound binding pockets was compared to that of 3D-RISM as shown in **Figure 3.9**. GalaxyWater-wKGB and FoldX, compared in the above sections, did not handle protein-compound complexes. GalaxyWater-CNN showed a higher coverage of 81 % and a lower RMSD of 0.96 Å compared with 48 % and 1.34 Å with 3D-RISM at $N_{pred}/N_{cryst} = 3$ when the distance cutoff of 1 Å was used to match predicted and crystallographic water molecules. Higher coverages of 88 and 67 % were obtained for GalaxyWater-CNN and 3D-RISM, respectively, with a more relaxed cutoff value of 1.5 Å.

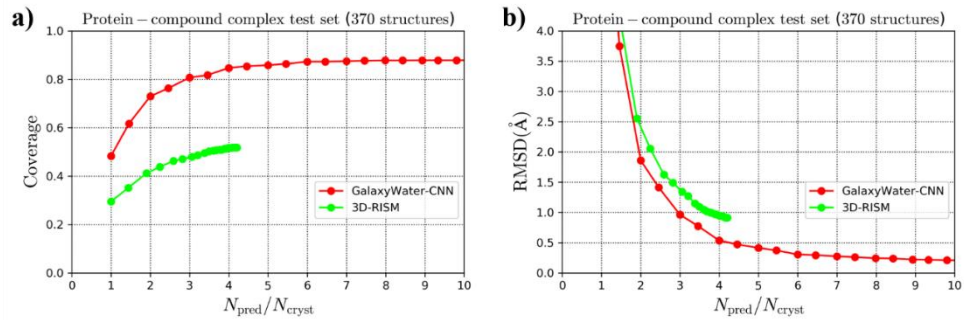


Figure 3.9. Performance comparison of GalaxyWater-CNN and 3D-RISM for predicting water molecules in the compound binding sites of proteins: a) Average coverage; b) RMSD of the predicted water molecules in the binding site.

GalaxyWater-CNN also showed higher absolute performance in predicting binding site waters in protein-compound complexes (**Figure 3.9**) than waters on protein surfaces (**Figure 3.5**) (coverage/RMSD of 88 %/0.21 Å compared with 86 %/0.78 Å at N_{pred}/N_{cryst} = 10). This also seems to have been caused by the binding site waters forming more hydrogen bonds than those on protein surfaces. An example of this is shown in **Figure 3.10**. In the figure, GalaxyWater-CNN predicted three crystal water molecules, making hydrogen bond networks in the binding site correctly for which 3D-RISM failed. The results of 3D-RISM are affected by many factors, including fluctuations of heavy atoms and hydrogen atoms involved in hydrogen bonds, accuracy of the force field, and accuracy of the integral equation closure. CNNs may overcome such issues by learning from 3D structure data directly, especially for water molecules, making stable enough hydrogen bond networks to be observed in the crystal structures.

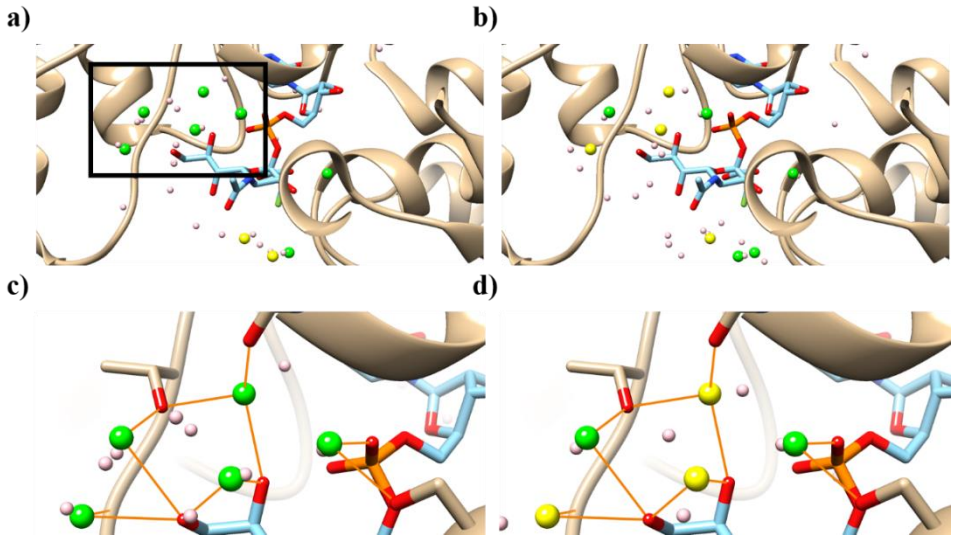


Figure 3.10. Example case of predicting binding-site water molecules on a protein-compound complex (PDB ID: 2ihj) at N_{pred}/N_{cryst} = 3. Green and pink spheres represent predicted water sites with and without corresponding crystallographic waters within 1 Å. Yellow spheres represent crystallographic water sites without corresponding predicted water sites within 1 Å. Orange lines show hydrogen bonds. Prediction results of GalaxyWater-CNN are shown in (a) and magnified in (c), and those of 3D-RISM are in (b) and (d).

4. CONCLUSION

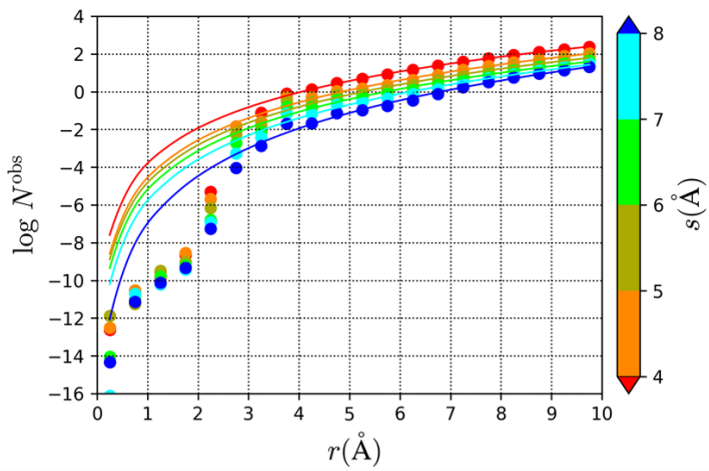
Two new water prediction methods on given protein structures were introduced in this thesis. In **Chapter 2**, water prediction method based on the new protein-water statistical potential called wKGB was introduced. wKGB statistical potential considers the degree of solvent accessibility of protein atoms as well as the orientation and distance of protein-water hydrogen bonds. The positive effect of including the solvent accessibility of protein atoms was illustrated well in the examples of the stronger interactions (1) in buried states than in exposed states for charged protein atoms owing to the less solvent shielding and (2) in slightly exposed states than in buried states for neutral polar protein atoms owing to the less competition with other protein atoms for hydrogen bonding. The positions of water molecules were predicted by identifying low-energy regions on the protein structures by GalaxyWater-wKGB using the wKGB potential. GalaxyWater-wKGB performs better than FoldX and is comparable to or better than 3D-RISM. To achieve higher performance in water placement, a method adopting Convolutional Neural Network, GalaxyWater-CNN, was developed, as introduced in **Chapter 3**.

GalaxyWater-CNN is a water position prediction method based on Convolutional Neural Network, and was able to cover a significant number of crystal waters with less prediction of water positions compared to the existing methods. Even with a small number of training sets, it showed robust performance for various test sets. The Positions of water molecules were predicted by placing water molecules at high water-probability regions on protein structures by GalaxyWater-CNN. Performance of GalaxyWater-CNN was generally better than GalaxyWater-wKGB, 3D-RISM, and FoldX, especially when a small number of water molecules are predicted with stricter criteria. This result can be ascribed to the higher amount of parameters of GalaxyWater-CNN enables us to depict water

distribution more accurately. Additionally, using atrous convolution, residual network, and many layers considerably increases the performance of GalaxyWater-CNN network, while using functional group channels and an auxiliary loss affect the performance relatively less. The faster and more accurate water position prediction methods presented here could be used to improve molecular docking methods by considering those water molecules explicitly.

SUPPLEMENTARY INFORMATION

(a)



(b)

Born radius (Å)	α_s	log β_s
< 4	2.72570	-3.81795
4~5	2.89524	-4.57753
5~6	2.91516	-4.84735
6~7	2.98885	-5.21368
7~8	3.17649	-5.80271
>8	3.65254	-7.00703

Figure S1. Estimation of available volume at given distance and solvation state for short distances used in derivation of wKGB statistical potential. (a) Average number of observed crystal water molecules per protein atom (N^{obs}) as function of protein-water distance (r) for varying solvation states that are represented as effective Born radius s of protein atom. Data at longer distances ($r \geq 4$ Å) were fitted with formula of $\beta_s r^{\alpha_s}$, and parameters obtained by this fitting are presented in (b).

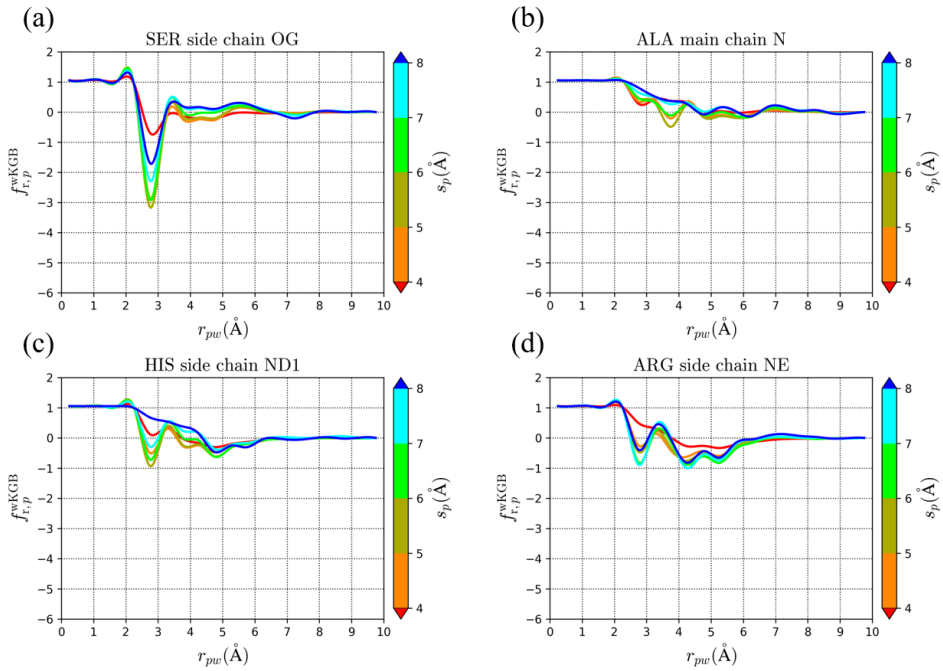


Figure S2. Radial part of smoothed wKGB potential for (a) SER side chain OG, (b) ALA main chain N, (c) HIS side chain ND1, and (d) ARG side chain NE

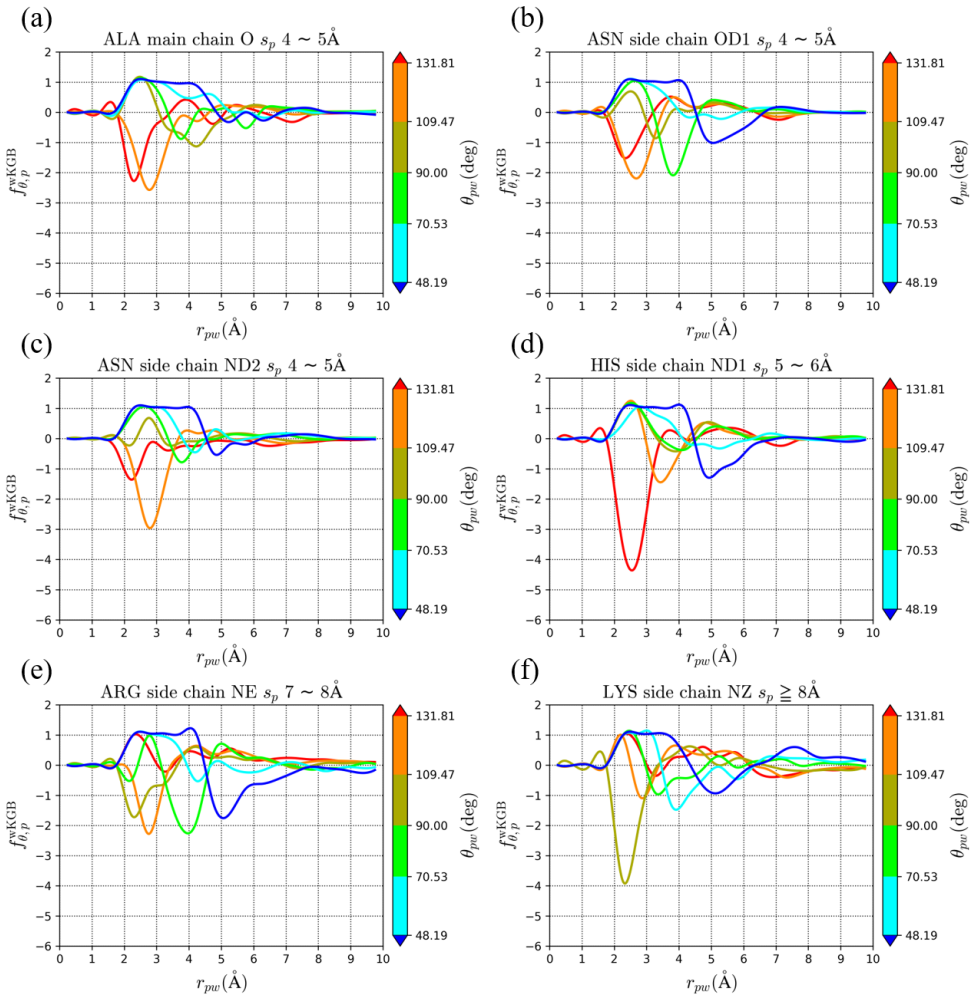


Figure S3. Angular part of smoothed wKGB potential for (a) ALA main chain O, (b) ASN side chain OD1, (c) ASN side chain ND2, (d) HIS side chain ND1, (e) ARG side chain NE, and (f) LYS side chain NZ

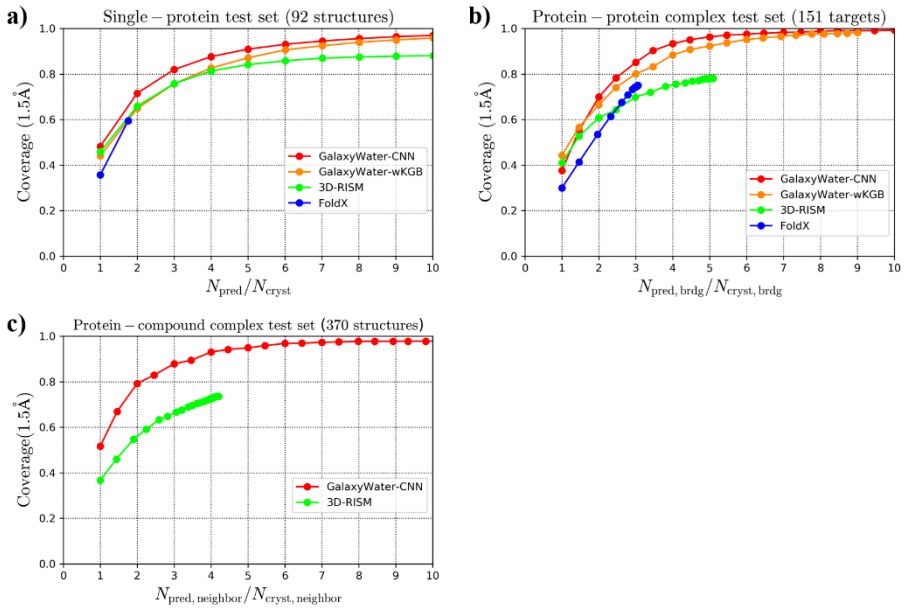


Figure S4. Coverage (cutoff = 1.5 Å) comparison between GalaxyWater-CNN, GalaxyWater-wKGB, 3D-RISM, and FoldX for test sets. Each graph plotted N_{pred} / N_{cryst} versus coverage for different test set, where a) plotted for single-protein test set, b) plotted for protein-protein complex structure set, and c) plotted for protein-compound complex test set

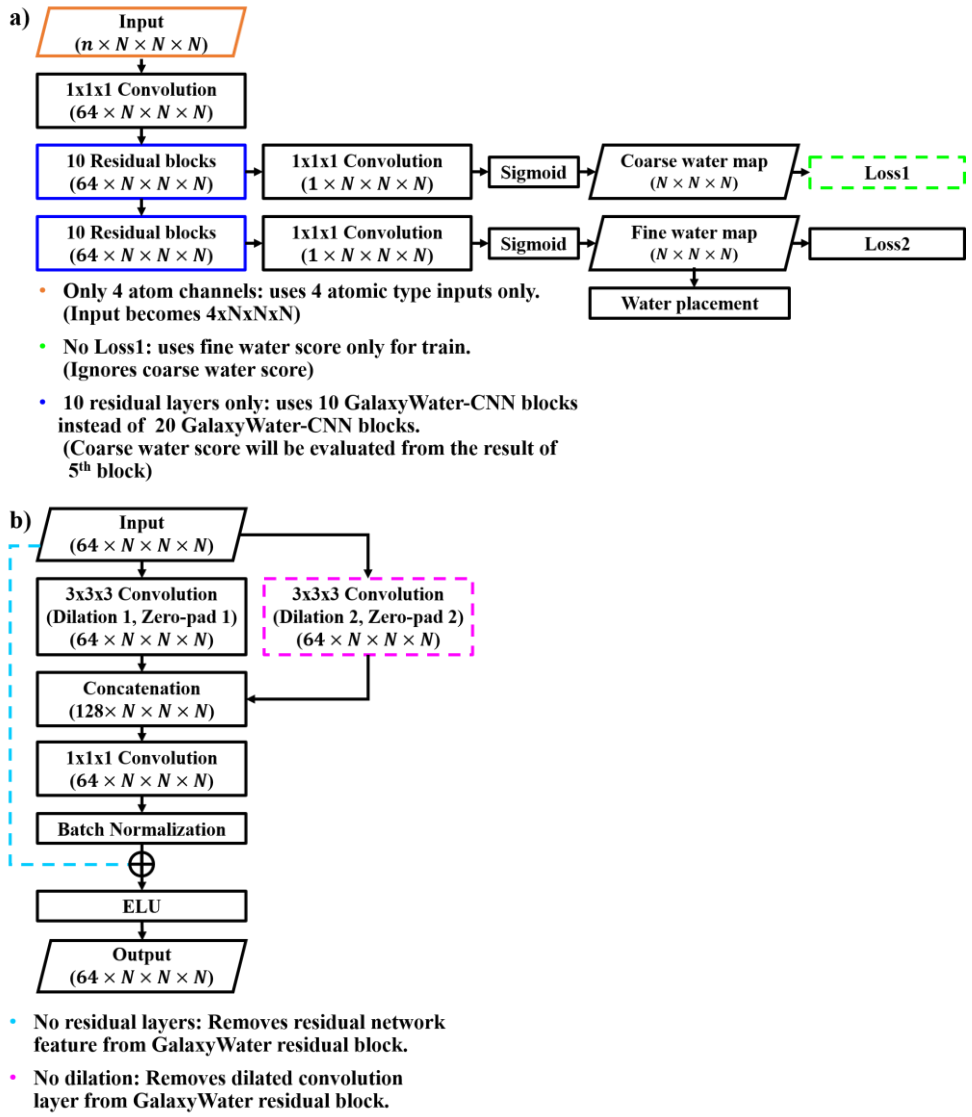


Figure S5. Detailed structures of GalaxyWater-CNN network without specific feature. Each colored object is modified or removed from the original network with corresponding color. Colored object with plain line is modified from the original network, while objects with dashed line is removed from the original network. Detailed information is written in the figure. a) Networks with modification with overall GalaxyWater-CNN procedure. b) Networks with modification inside GalaxyWater-CNN layer.

	Backbone				Sidechain					
ALA	N	CA	C	O	CB					
CYS	N	CA	C	O	CB	SG				
ASP	N	CA	C	O	CB	CG	OD1/2			
GLU	N	CA	C	O	CB	CG	CD	OD1/2		
PHE	N	CA	C	O	CB	CG	CD1/2	CE1/2	CZ	
GLY	N	CA	C	O						
HIS	N	CA	C	O	CB	CG	ND1	CD2	CE1	NE2
ILE	N	CA	C	O	CB	CG1	CG2	CD1		
LYS	N	CA	C	O	CB	CG	CD	CE	NZ	
LEU	N	CA	C	O	CB	CG	CD1/2			
MET	N	CA	C	O	CB	CG	SD	CE		
ASN	N	CA	C	O	CB	CG	OD1/2	ND2		
PRO	N	CA	C	O	CB	CG	CD			
GLN	N	CA	C	O	CB	CG	CD	OE1	NE2	
ARG	N	CA	C	O	CB	CG	CD	NE	CZ	NH1/2
SER	N	CA	C	O	CB	OG				
THR	N	CA	C	O	CB	OG1	CG2			
VAL	N	CA	C	O	CB	CG1/2				
TRP	N	CA	C	O	CB	CG	CD1	CD2	NE1	
					CE2	CE3	CZ2	CZ3	CH2	
TYR	N	CA	C	O	CB	CG	CD1/2	CE1/2	CZ	OH

Table S1. Details of 158 atom types used in wKGB statistical potential.

N_{pred}/N_{cryst}	σ_{RMSD} (Å)	$\sigma_{Coverage}$ (cutoff: 1Å)	$\sigma_{Coverage}$ (cutoff: 1.5Å)
1	0.681	0.0269	0.0228
3	0.141	0.0334	0.0235
5	0.095	0.0356	0.0218
7	0.076	0.0359	0.0196
9	0.057	0.036	0.0178

Table S2. Effect of rotation and translation of input protein crystal structure on the water prediction performance. Standard deviations of the prediction performance in terms of RMSD and coverage (for different cut-off values for correct prediction) for 25 random rotation-translations averaged over the test set proteins are presented for different number predicted water molecules.

wKGB score cutoff	N_{pred}/N_{cryst}	RMSD (Å)	Coverage (cutoff: 1Å)
-4.0	10.7	0.815	0.800
-6.0	6.62	0.993	0.769
-8.0	4.06	1.549	0.700
-10.0	2.52	2.616	0.598

Table S3. Dependence of average coverage and RMSD of crystallographic water positions on test set for different score cutoff values for crystal structure set..

Common atom channels	
C atom	**O atom**
N atom	**S atom**
Extra atom channels (protein-compound version only)	
P atom	halogen atoms
Metal atoms	All atoms except above
Functional group channels (protein version only)	
Main chain amide C/N	Main chain amide C/O
Side chain amide C/N	Side chain LYS/ARG amine C/N
Side chain HIS amine C/N	Side chain TRP amine C/N
Side chain phenol C/O	Side chain hydroxyl C/O
Side chain carboxyl C/O	Side chain carbonyl C/O
Side chain thiol/sulfide C/S	Side chain aliphatic/aromatic C

Table S4. Type of input channels in GalaxyWater-CNN. Channels with blue background corresponds to the atom type channels used both in protein version and protein-compound complexversion of GalaxyWater-CNN, while red background corresponds to the extra atom channels used for protein-compound complex version of GalaxyWater-CNN only, and channels with green background corresponds to the functional group channels used for protein version of GalaxyWater-CNN only.

N_{pred}/N_{cryst}	σ_{RMSD} (Å)	$\sigma_{Coverage}$ (cutoff: 0.5Å)	$\sigma_{Coverage}$ (cutoff: 1Å)	$\sigma_{Coverage}$ (cutoff: 1.5Å)
1	1.103	0.0355	0.0271	0.0260
3	0.179	0.0420	0.0274	0.0227
5	0.125	0.0437	0.0271	0.0188
7	0.093	0.0441	0.0276	0.0160
9	0.070	0.0445	0.0279	0.0138

Table S5. Effect of rotation and translation of input protein crystal structure on the water prediction performance. Standard deviations of the prediction performance in terms of RMSD and coverage for 25 random rotation-translations averaged over the test set of wkgb crystal structure set.

	Score cutoff	N_{pred}/N_{cryst}	RMSD (Å)	Coverage
Single-chain proteins	**34**	**5.45**	**1.07**	**83.6%**
	38	**4.05**	**1.37**	**80.6%**
	42	**2.80**	**2.05**	**74.8%**
Protein-protein complexes	**34**	**6.62**	**0.26**	**89.3%**
	38	**5.57**	**0.29**	**89.0%**
	42	**4.48**	**0.37**	**87.9%**
Protein-compound complexes	**34**	**3.96**	**0.53**	**85.9%**
	38	**3.26**	**0.71**	**84.5%**
	42	**2.66**	**1.14**	**81.7%**

Table S6. Performance of GalaxtWater-CNN at three different score cutoff values on the single-protein, protein-protein complex, and protein-compound complex test sets.

BIBLIOGRAPHY

1. Bhat, T. N.; Bentley, G. A.; Boulot, G.; Greene, M. I.; Tello, D.; Dall'Acqua, W.; Souchon, H.; Schwarz, F. P.; Mariuzza, R. A.; Poljak, R. J., Bound water molecules and conformational stabilization help mediate an antigen–antibody association. *Proc Natl Acad Sci U S A* **1994**, 91, 1089-1093.
2. Park, S.; Saven, J. G., Statistical and molecular dynamics studies of buried waters in globular proteins. *Proteins* **2005**, 60, 450-463.
3. Levy, Y.; Onuchic, J. N., Water mediation in protein folding and molecular recognition. *Annu Rev Biophys Biomol Struct* **2006**, 35, 389-415.
4. Imai, T.; Hiraoka, R.; Kovalenko, A.; Hirata, F., Locating missing water molecules in protein cavities by the three-dimensional reference interaction site model theory of molecular solvation. *Proteins* **2007**, 66, 804-813.
5. Reichmann, D.; Phillip, Y.; Carmi, A.; Schreiber, G., On the contribution of water-mediated interactions to protein-complex stability. *Biochemistry* **2008**, 47, 1051-1060.
6. Villacanas, O.; Madurga, S.; Giralt, E.; Belda, I., Explicit treatment of water molecules in protein-ligand docking. *Curr Comput-Aid Drug* **2009**, 5, 145-154.
7. Sabarinathan, R.; Aishwarya, K.; Sarani, R.; Vaishnavi, M. K.; Sekar, K., Water-mediated ionic interactions in protein structures. *J Biosci* **2011**, 36, 253-263.
8. Li, S.; Bradley, P., Probing the role of interfacial waters in protein-DNA recognition using a hybrid implicit/explicit solvation model. *Proteins* **2013**, 81, 1318-1329.
9. Sousa, S. F.; Ribeiro, A. J. M.; Coimbra, J. T. S.; Neves, R. P. P.; Martins, S. A.; Moorthy, N. S. H. N.; Fernandes, P. A.; Ramos, M. J., Protein-ligand docking in the new millennium - A retrospective of 10 years in the field. *Curr Med Chem* **2013**, 20, 2296-2314.
10. Gathiaka, S.; Liu, S.; Chiu, M.; Yang, H. W.; Stuckey, J. A.; Kang, Y. N.; Delproposto, J.; Kubish, G.; Dunbar, J. B.; Carlson, H. A.; Burley, S. K.; Walters, W. P.; Amaro, R. E.; Feher, V. A.; Gilson, M. K., D3R grand challenge 2015: Evaluation of protein-ligand pose and affinity predictions. *J Comput Aided Mol Des* **2016**, 30, 651-668.
11. Das, R.; Baker, D., Macromolecular modeling with Rosetta. *Annu Rev Biochem* **2008**, 77, 363-382.
12. Fiser, A.; Do, R. K. G.; Sali, A., Modeling of loops in protein structures. *Protein Sci* **2000**, 9, 1753-1773.
13. Shin, W. H.; Lee, G. R.; Heo, L.; Lee, H.; Seok, C., Prediction of Protein Structure and Interaction by Galaxy protein modeling programs. *Bio Des* **2014**, 2, 1-11.
14. Kleinjung, J.; Fraternali, F., Design and application of implicit solvent models in biomolecular simulations. *Curr Opin Struct Biol* **2014**, 25, 126-134.
15. Foloppe, N.; Fisher, L. M.; Howes, R.; Kierstan, P.; Potter, A.; Robertson, A. G. S.; Surgenor, A. E., Structure-based design of novel Chk1 inhibitors: Insights into hydrogen bonding and protein-ligand affinity. *J Med Chem* **2005**, 48, 4332-4345.

16. 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.
17. Feig, M.; Mirjalili, V., Protein structure refinement via molecular-dynamics simulations: What works and what does not? *Proteins* **2016**, 84 Suppl 1, 282-1292.
18. Park, H.; Lee, G. R.; Kim, D. E.; Anishchenko, I.; Cong, Q.; Baker, D., High-accuracy refinement using Rosetta in CASP13. *Proteins* **2019**, 87, 1276-1282.
19. Uehara, S.; Tanaka, S., AutoDock-GIST: Incorporating thermodynamics of active-site water into scoring function for accurate protein-ligand docking. *Molecules* **2016**, 21.
20. Young, T.; Abel, R.; Kim, B.; Berne, B. J.; Friesner, R. A., Motifs for molecular recognition exploiting hydrophobic enclosure in protein-ligand binding. *Proc Natl Acad Sci U S A* **2007**, 104, 808-813.
21. Beuming, T.; Che, Y.; Abel, R.; Kim, B.; Shanmugasundaram, V.; Sherman, W., Thermodynamic analysis of water molecules at the surface of proteins and applications to binding site prediction and characterization. *Proteins* **2012**, 80, 871-883.
22. Hu, B.; Lill, M. A., WATsite: Hydration site prediction program with PyMOL interface. *J Comput Chem* **2014**, 35, 1255-1260.
23. Lynch, G. C.; Perkyns, J. S.; Nguyen, B. L.; Pettitt, B. M., Solvation and cavity occupation in biomolecules. *Biochim Biophys Acta* **2015**, 1850, 923-931.
24. Masters, M. R.; Mahmoud, A. H.; Yang, Y.; Lill, M. A., Efficient and accurate hydration site profiling for enclosed binding sites. *J Chem Inf Model* **2018**, 58, 2183-2188.
25. Pradhan, M. R.; Nguyen, M. N.; Kannan, S.; Fox, S. J.; Kwoh, C. K.; Lane, D. P.; Verma, C. S., Characterization of Hydration Properties in Structural Ensembles of Biomolecules. *J Chem Inf Model* **2019**, 59, 3316-3329.
26. Yang, Y.; Hu, B.; Lill, M. A., Analysis of factors influencing hydration site prediction based on molecular dynamics simulations. *J Chem Inf Model* **2014**, 54, 2987-2995.
27. Imai, T.; Oda, K.; Kovalenko, A.; Hirata, F.; Kidera, A., Ligand mapping on protein surfaces by the 3D-RISM theory: Toward computational fragment-based drug design. *J Am Chem Soc* **2009**, 131, 12430-12440.
28. Güssregen, S.; Matter, H.; Hessler, G.; Lionta, E.; Heil, J.; Kast, S. M., Thermodynamic characterization of hydration sites from integral equation-derived free energy densities: Application to protein binding sites and ligand series. *J Chem Inf Model* **2017**, 57, 1652-1666.
29. Huang, W.; Blinov, N.; Wishart, D. S.; Kovalenko, A., Role of water in ligand binding to maltose-binding protein: Insight from a new docking protocol based on the 3D-RISM-KH molecular theory of solvation. *J Chem Inf Model* **2015**, 55, 317-328.
30. Petukhov, M.; Cregut, D.; Soares, C. M.; Serrano, L., Local water bridges and protein conformational stability. *Protein Sci* **1999**, 8, 1982-1989.
31. Rarey, M.; Kramer, B.; Lengauer, T., The particle concept: Placing discrete water molecules during protein-ligand docking predictions. *Proteins* **1999**,

34, 17-28.
32. Nittinger, E.; Flachsenberg, F.; Bietz, S.; Lange, G.; Klein, R.; Rarey, M., Placement of water molecules in protein structures: From large-scale evaluations to single-case examples. *J Chem Inf Model* **2018**, 58, 1625-1637.
33. Rossato, G.; Ernst, B.; Vedani, A.; Smiesko, M., AcquaAlta: A directional approach to the solvation of ligand-protein complexes. *J Chem Inf Model* **2011**, 51, 1867-1881.
34. Ross, G. A.; Morris, G. M.; Biggin, P. C., Rapid and accurate prediction and scoring of water molecules in protein binding sites. *PLOS ONE* **2012**, 7, e32036.
35. Verdonk, M. L.; Chessari, G.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Nissink, J. W. M.; Taylor, R. D.; Taylor, R., Modeling water molecules in protein-ligand docking using GOLD. *J Med Chem* **2005**, 48, 6504-6515.
36. Li, Y.; Gao, Y.; Holloway, M. K.; Wang, R., Prediction of the favorable hydration sites in a protein binding pocket and its application to scoring function formulation. *J Chem Inf Model* **2020**, 60, 4359-4375.
37. Nguyen, C.; Yamazaki, T.; Kovalenko, A.; Case, D. A.; Gilson, M. K.; Kurtzman, T.; Luchko, T., A molecular reconstruction approach to site-based 3D-RISM and comparison to GIST hydration thermodynamic maps in an enzyme active site. *PLOS ONE* **2019**, 14, e0219473.
38. Kaiming He, X. Z.; Ren, S.; Sun, J. **2016**, Identity mappings in deep residual networks. *Eur Conference on Computer Vision*, 630-645.
39. Chen, L. C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A. L., DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Trans Pattern Anal Mach Intell* **2018**, 40, 834-848.
40. Heo, L.; Park, S.; Seok, C., GalaxyWater-wKGB: Prediction of water positions on protein structure using wKGB statistical potential. *J Chem Inf Model* **2021**, 61, 2283-2293.
41. 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.
42. Kovalenko, A.; Hirata, F., Potential of mean force between two molecular ions in a polar molecular solvent: A study by the three-dimensional reference interaction site model. *J Phys Chem B* **1999**, 103, 7942-7957.
43. Delgado, J.; Radusky, L. G.; Cianferoni, D.; Serrano, L., FoldX 5.0: Working with RNA, small molecules and a new graphical interface. *Bioinformatics* **2019**, 35, 4168-4169.
44. Yang, Y.; Zhou, Y., Specific interactions for ab initio folding of protein terminal regions with secondary structures. *Proteins* **2008**, 72, 793-803.
45. Zhou, H.; Skolnick, J., GOAP: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction. *Biophys J* **2011**, 101, 2043-2052.
46. Haberthur, U.; Caflisch, A., FACTS: Fast analytical continuum treatment of solvation. *Journal of Computational Chemistry* **2008**, 29, 701-715.
47. Wang, G. L.; Dunbrack, R. L., PISCES: A protein sequence culling server. *Bioinformatics* **2003**, 19, 1589-1591.

48. Sindhikara, D. J.; Yoshida, N.; Hirata, F., Placevent: An algorithm for prediction of explicit solvent atom distributionuApplication to HIV-1 protease and F-ATP synthase. *Journal of Computational Chemistry* **2012**, 33, 1536-1543.
49. Case, D. A.; Darden, T. A.; Cheatham, T. E., III; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Walker, R. C.; Zhang, W.; Merz, K. M.; Roberts, B.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Kolossváry, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Liu, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Cai, Q.; Ye, X.; Wang, J.; Hsieh, M.-J.; Cui, G.; Roe, D. R.; Mathews, D. H.; Seetin, M. G.; Sagui, C.; Babin, V.; Luchko, T.; Gusarov, S.; Kovalenko, A.; Kollman, P. A., Amber 11. *University of California* **2010**.
50. Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P., The missing term in effective pair potentials. *J Phys Chem* **1987**, 91, 6269-6271.
51. Grabowski, S. J., Hydrogen bonding strength - measures based on geometric and topological parameters. *J Phys Org Chem* **2004**, 17, 18-31.
52. LeCun, Y.; Bengio, Y.; Hinton, G., Deep learning. *Nature* **2015**, 521, 436-444.
53. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A., Going deeper with convolutions, *2015 IEEE Conference on Computer Vision and Pattern Recognition*, **2015**, 1-9.
54. Ragoza, M.; Hochuli, J.; Idrobo, E.; Sunseri, J.; Koes, D. R., Protein-ligand scoring with convolutional neural networks. *J Chem Inf Model* **2017**, 57, 942-957.
55. 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.
56. Krull, F.; Korff, G.; Elghobashi-Meinhardt, N.; Knapp, E. W., ProPairs: A data set for protein–protein docking. *J Chem Inf Model* **2015**, 55, 1495-1507.

국문초록

대부분의 생체 단백질은 수용액 상태에서 존재하며, 단백질 분자는 물 분자와 많은 상호작용을 일으킨다. 이러한 상호작용은 단백질의 구조나 기능에 중요한 역할을 한다. 따라서 단백질의 구조와 기능을 예측하는 방법들은 단백질과 물 분자 사이의 상호작용을 직, 간접적으로 고려하게 된다. 간접적으로 물과 단백질 분자 사이의 상호작용을 고려하는 방법으로는 물을 일종의 유전체로 가정하는 방법을 사용하는데, 이러한 방법은 각각의 물 분자의 위치를 고려할 필요가 없기 때문에 비교적 계산 비용이 낮고, 물과 단백질 분자 사이의 상호작용 중 많은 부분을 차지하는 정전기적 상호작용을 모사할 수는 있지만, 물 분자의 위치에 따라 크게 달라질 수 있는 물과 단백질 사이의 수소결합과 같은 근거리 상호작용을 모사하기 어렵다는 문제점이 있다. 특히 물과 단백질 분자 사이의 근거리 상호작용은 단백질의 기능에 영향을 끼치기 때문에 단백질의 기능을 예측하는 방법에서는 단백질과 근거리 상호작용을 할 가능성이 높은 물 분자들의 위치와 단백질과의 상호작용을 예측하는 것이 중요할 수 있다. 물과 단백질 분자 사이의 근거리 상호작용을 고려하기 위해서는 물 분자의 위치를 직접적으로 반영하여 물과 단백질 사이의 상호작용을 모사하며, 주로 분자동역학 시뮬레이션이나 3D-RISM이 사용된다. 이러한 방법들은 물과 단백질 사이의 상호작용을 더욱 자세하게 모사할 수 있지만 계산비용이 높다는 문제가 있으며, 단백질과 물 사이의 상호작용에 상당한 기여를 하는 단백질에 결합된 물의 위치를 잘 예측하지 못한다는 문제도 존재한다.

따라서, 본 학위 논문에서는 단백질 주변의 물 분자의 위치를 예측하는 2가지의 방법을 제시하였다. 첫번째 시도는 단백질을 구성하는 원자의 용매화 상태를 고려하여 물과 단백질 사이의 통계기반 포텐셜 함수를 이용하여 단백질 주변의 물의 위치를 예측하는 방법이었다. 이 방법은 3D-RISM 방법에 비해서 평균적으로 180배의 계산 속도 향상을 보여주었으며, 단백질에 결합된 물 분자의 위치를 예측하는 성능은 3D-RISM과 비슷하거나 더 높았다. 그러나 이 방법은 수소결합에 직접적으로 참여하지 않는 단백질 원자와 물 분자 사이의 포텐셜 우물을 만들어지는 현상이 존재하였기 때문에 제한된 예측 성능을 보여주었다. 이러한 문제로 인하여 물 분자를 수용할 수 있는 단백질의 구조 패턴을 인식할 수 있는 Convolutional neural network를 이용한 물 분자 위치 예측 방법을 만들었고, 통계 기반 포텐셜 함수를 이용한 물 분자 위치 예측 방법에 비해 더욱 높은 예측 성능을 보였다. 이 방법은 GPGPU를 사용하였을 경우, 3D-RISM을 사용한 방법에 비해 44배의 속도 향상을 보였고, CPU만을 사용했을 때에도 58%의 속도 향상을 보였다. 예측 성능의 경우, 단백질 분자의 결정 구조에 포함된 물 분자의 수의 3배의 물 분자의 위치를 예측했을 때, 예측된 위치가 결정 구조에 존재하는 물 분자의 위치의 1Å 이내에 있을 확률이 75% 이상이었다.

이 논문에서 제시된 방법들을 이용하여 단백질 주변의 물의 위치를 더 정확히 예측할 수 있다. 나아가서 단백질-리간드 도킹을 할 때, 단백질에 붙잡혀있는 물 분자의 위치를 고려하여 더욱 단백질-리간드 도킹을 할 수 있을 것으로 예상된다.

주요어: 단백질-물 상호작용, 물 분자 위치 예측, 통계기반 포텐셜 함수, Convolutional Neural Network

학 번: 2013-22921

감사의 글

7학기로 학부과정을 마치고 계산화학 석박사 통합과정에 들어설 때, 저는 양자 시뮬레이션 정도밖에 떠올리지 못하는 이 분야에 막연한 기대와 호기심만으로 가득찬 새내기였습니다. 그런데 알면 알수록 모르는 것도 많아진다는 소크라테스의 말처럼, 2편의 졸업 논문을 낸 지금도 여전히 계산화학의 깊이와 가능성을 헤아리기 어렵습니다. 그럼에도 연구과정을 마치며 지나온 시간을 돌이켜보니, 많은 분들의 도움으로 제가 여기까지 올 수 있었습니다. 짧지만 큰 감사의 마음을 담아 이 글을 씁니다.

먼저, 오랜 시간동안 저를 지도해주시고 많은 리비전 과정 동안 격려와 용기를 주신 석차옥 교수님께 진심으로 감사드립니다. 연구하는 방법에 대해 큰 도움을 주신 한범이 형, 학회 포스터 발표 과정에서 조언과 비전을 주신 웅희 형, 졸업논문에 적히지는 않았지만 protein-protein docking에 대해서 알려주고 박사 졸업 이후의 진로에 대해서 도움을 주신 하섭이 형, 졸업논문에 사용된 연구의 근간에 많은 도움을 주신 림 형, 어찌보면 불안했던 대학원 과정 중에서 심적으로 도와주신 규리 누나, 여러 방면으로 도움을 주신 바로 윗 선배 민경 누나, 훈련소에서 같이 지냈던 태용이, 클러스터 관리에 애쓴 진솔이, 졸업학기에 많은 도움을 준 현욱이와 소희, 그리고 좋은 말벗이 되어주신 마틴 슈타이네거 교수님과 마음 따듯한 종훈이 형에게도 감사드립니다.

같은 취미를 공유하며 서로에게 인생의 힘이 되어 준 화학부 선배와 동기생들, 문기 형, 창규 형, 한결이 형, 인환이, 해외에서 무탈하게 지낼 수 있도록 도와주신 범창이 형과 주안이, 투자에 대한 생각을 공유했던 토니, 건강에 대해서 좋은 조언을 준 완상이, 어렸을 때부터 벗이 되어 함께 응원해 준 영준이, 홍일이, 재원이 그리고 저를 믿어주시고 과학자의 꿈을 키워주신 양근하 선생님과 고광진 선생님, 상명고등학교 선생님들께 감사드립니다. 또한 모교에서 보낸 저의 청춘, 20대에게도 고맙다는 말을 전합니다.

마지막으로 여러 어려운 상황에서도 물심양면으로 도와주신 부모님께도 감사하다는 말씀을 드리며 이 글을 마칩니다.