

# SuperWater: Predicting Water Molecule Positions on Protein Structures by Generative AI

Xiaohan Kuang<sup>†1</sup>, Zhaoqian Su<sup>†1\*</sup>, Yunchao (Lance) Liu<sup>3</sup>, Xiaobo Lin<sup>1</sup>,  
Jesse Spencer-Smith<sup>1</sup>, Tyler Derr<sup>3</sup>, Yinghao Wu<sup>4</sup>, Jens Meiler<sup>2\*</sup>

<sup>1</sup>Data Science Institute, Vanderbilt University, Nashville, 37212, TN, USA.

<sup>2</sup>Chemical and Biomolecular Engineering Department, Vanderbilt University, Nashville, 37235, TN, USA.

<sup>3</sup>Computer Science Department, Vanderbilt University, Nashville, 37240, TN, USA.

<sup>4</sup>Department of Systems and Computational Biology, Albert Einstein College of Medicine, Bronx, 10461, NY, USA.

\*Corresponding author(s). E-mail(s): [zhaoqian.su@vanderbilt.edu](mailto:zhaoqian.su@vanderbilt.edu); [jens@meilerlab.org](mailto:jens@meilerlab.org);

## Abstract

Water molecules play a significant role in maintaining protein structural stability and facilitating molecular interactions. Accurate prediction of water molecule positions around protein structures is essential for understanding their biological roles and has significant implications for protein engineering and drug discovery. Here, we introduce SuperWater, a novel generative AI framework that integrates a score-based diffusion model with equivariant graph neural networks to predict water molecule placements around proteins with high accuracy. SuperWater surpasses existing methods, delivering state-of-the-art performance in both crystal water coverage and prediction precision, achieving water localization within  $0.3 \pm 0.06$  Å of experimentally validated positions. We demonstrate the capabilities of SuperWater through case studies involving protein hydration, protein-ligand binding, and protein-protein binding sites. This framework can be adapted for various applications, including structural biology, binding site prediction, multi-body docking, and water-mediated drug design.

**Keywords:** Generative AI, diffusion model, water-binding sites, drug discovery

## 1 Introduction

Water is essential for life, with most biological processes occurring in aqueous environments [1]. It profoundly influences the structure, stability, dynamics, and function of biomolecules within it [2, 3, 4, 5]. During protein folding, water facilitates hydrophobic interactions, drawing non-polar residues together, while also participating in hydrogen bond networks and modulating interactions between polar residues [6, 7, 8]. In protein-protein and protein-ligand interactions, water molecules not only compete with ligands for binding sites but also play a crucial role in ligand dissociation [9, 10, 11, 12]. Overlooking the role of water molecules in the process of molecular recognition has led to failures in structure-based drug discovery campaigns [13]. Thus, precise modeling of water-biomolecule interactions is crucial for effective protein design and structure-based drug development.

Despite the crucial role of water molecules, accurately determining their positions remains challenging due to the limitations of current techniques. Experimental methods often suffer from inadequate spatial and temporal resolution [3], making it difficult to fully characterize the hydration structures across entire protein surfaces [14]. To address these limitations, computational approaches, such as Molecular Dynamics and Monte Carlo simulations, have been widely employed [15, 16, 17, 18, 19]. However, these physics-based methods are heavily dependent on accurate potential energy functions and become increasingly inefficient when applied to large complexes due to their computational intensity.

---

<sup>†</sup>Equal contribution.

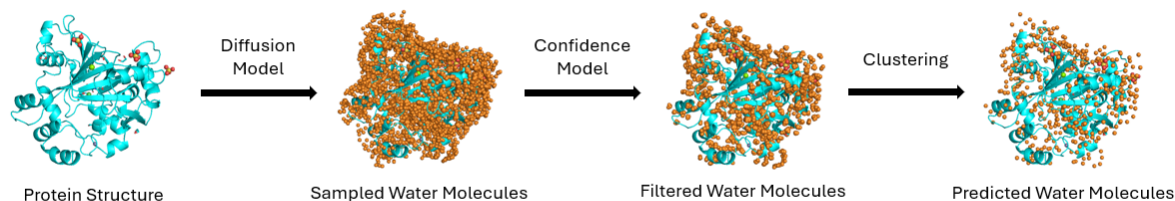
With advancements in deep learning, several 3D Convolutional Neural Network (CNN) methods have been developed to predict the placement of water molecules around proteins [20, 21, 22, 23]. These methods represent the 3D structure of a protein as a 3D image, with input channels corresponding to different atom types on 3D grid points. The models are trained to differentiate between water-occupied and unoccupied voxels near the protein surface, followed by a physics-based postprocessing step to refine predictions. These 3D-CNN methods have demonstrated superior water coverage compared to traditional physics-based approaches [20, 21]. However, despite these improvements, traditional CNN-based models still encounter significant challenges. Achieving sub-angstrom accuracy in water position prediction demands very fine grid spacing, which drastically increases computational costs [24]. Additionally, these CNN-based models are sensitive to the orientation of the input structure. Although techniques like local grids or data augmentation with random rotations can alleviate this issue, they do not completely solve the problem of rotational invariance [25, 26].

In recent years, diffusion models have emerged as powerful generative AI tools [27, 28], leading to significant advancements across various areas of bioinformatics. [29, 30, 31, 32, 33, 34, 35]. Building on this foundation, we introduce SuperWater, a novel generative AI approach that leverages a score-based diffusion model combined with equivariant graph neural networks to accurately predict the placement of water molecules around proteins. Rather than directly approximating the probability distribution of water molecules around the proteins, our model learns to estimate the gradient of water distributions. This learned gradient is then used to generate water positions from a normal distribution, which are further refined using a confidence model and a clustering algorithm to optimize accuracy. SuperWater surpasses existing methods, delivering state-of-the-art results in both the coverage of crystal water molecules and the precision of predicted positions. This approach holds promise for a wide range of applications, including structural studies, binding site predictions, multi-body docking, and water-aided drug design.

## 2 Methods

### 2.1 Overview

The SuperWater pipeline comprises three main stages. The process begins with the preprocessing of the protein structure and nearby small molecules or metal atoms into heterogeneous geometric graphs, formatted for model processing. These graphs are then fed into a score-based diffusion model to sample potential positions for water molecules around the protein structure. After the initial sampling, an equivariant graph neural network-based confidence model assigns scores to each candidate water position, filtering out low-confidence predictions. Finally, a clustering algorithm consolidates neighboring water molecule positions into a single representative point for each cluster.



**Fig. 1:** SuperWater Workflow.

Fig. 1 provides a visual illustration of the entire workflow. The primary steps of SuperWater include:

- A diffusion model that operates on geometric graphs of the protein structure, allowing for the sampling of water molecule placements around the protein.
- An equivariant graph neural network-based confidence model that assesses each sampled location, discarding low-confidence predictions.
- A post-processing step that enhances prediction accuracy by clustering the predicted positions.

### 2.2 Datasets and Preprocessing

We collected high-resolution protein structures (resolution better than 1.5 Å) from the PDBBank [36], resulting in an initial dataset of 23,189 PDB files. To analyze the water binding modes, we examined all protein pairs with sequence similarity greater than 90%, as illustrated in Fig. S1. Our analysis revealed that even

proteins with high sequence similarity (> 90%) exhibit distinct root-mean-square deviation (RMSD) values and variations in water distribution. This finding suggests that water molecules can adopt different binding modes or positions, even in highly similar protein sequences. Consistent with other studies that have clustered datasets based on binding modes rather than sequence similarity [37, 38, 39], we retained all selected PDB structures to capture the diversity of water-binding interactions and enhance model generalizability.

To further refine the dataset, we selected protein structures containing between 100 and 500 residues. Water molecules within a 4 Å radius of the protein were retained, as these fall within the second solvation shell and are involved in hydrogen bonding and hydrophobic interactions [40, 41, 42, 43, 44]. Each water molecule was represented solely by its oxygen atom [20]. For structures with multiple models, such as those with alternate residue conformations, only the first model was used. To ensure an adequate representation of water molecules, we retained only structures with a water-to-residue ratio greater than 0.6, yielding a final dataset of 17,092 PDB files. This dataset was randomly split into training, validation, and testing sets in an 8:1:1 ratio, resulting in 13,674 structures for training, 1,709 for validation, and 1,709 for testing.

## 2.3 Score-based Diffusion Model

In the diffusion model, each data point represents the 3D coordinates of a water molecule’s position on the protein structure. The generative model seeks to estimate the probability distribution  $p(\mathbf{x} | \mathbf{y})$ , where  $\mathbf{x}$  indicates water positions and  $\mathbf{y}$  denotes the protein structure [31]. Estimating  $p(\mathbf{x} | \mathbf{y})$  poses two significant challenges.

The first challenge arises from the intractability of directly computing the probability distribution, as it requires normalizing the distribution across the entire space of possible positions. To address this, rather than estimating  $p(\mathbf{x} | \mathbf{y})$  directly, the diffusion model estimates its gradient,  $\nabla_{\mathbf{x}} \log p(\mathbf{x} | \mathbf{y})$ , known as the score function,  $\mathbf{S}_{\theta}(\mathbf{x}|\mathbf{y})$  [28, 45]. The score function  $\mathbf{S}_{\theta}(\mathbf{x}|\mathbf{y})$  is a vector field that directs water molecules toward favorable configurations from their current positions in 3D space.

The second challenge arises from the lack of sufficient training data in certain regions of the protein. To address this, the true data distribution is ‘evolved’ into a known distribution, typically a normal distribution [27, 46]. Through this process, the model diffuses data (i.e., water molecules) across the entire three-dimensional space conditioned on the protein structure, effectively filling in knowledge gaps and learning from a broader range of information across the protein structure. The entire architecture and technical details are depicted in the Supporting Information.

### 2.3.1 Forward Diffusion and Training with SDE

The score-based diffusion model utilized in this study leverages stochastic differential equations (SDE) to generate the three-dimensional coordinates for water molecule positions around protein structures. The forward diffusion process transforms the data distribution  $p(\mathbf{x})$  into a Gaussian reference distribution  $p_T(\mathbf{x})$  through the gradual addition of noise. The dynamics of this process are governed by an SDE:

$$d\mathbf{x}_t = f(t, \mathbf{x}_t)dt + g(t)d\mathbf{w}_t, \quad (1)$$

where  $\mathbf{x}_t$  represents the noisy data at time  $t$ ,  $f(t, \mathbf{x}_t)$  is the drift coefficient,  $g(t)$  is the diffusion coefficient, and  $\mathbf{w}_t$  is a Wiener process. This formulation allows for a continuous transition from the real data distribution to a noise distribution, typically Gaussian.

The noise schedule is controlled by the variance  $\sigma_t$ , which evolves over time according to:

$$\sigma_t = \sigma_{\min}^{(1-t)} \cdot \sigma_{\max}^t, \quad (2)$$

where  $\sigma_{\min}$  and  $\sigma_{\max}$  are the minimum and maximum noise scales, respectively. The resulting SDE can then be expressed in terms of the time-evolving variance:

$$d\mathbf{x} = \sqrt{\frac{d\sigma^2(t)}{dt}}d\mathbf{w}. \quad (3)$$

To approximate  $p(\mathbf{x}_0|\mathbf{x}_t)$ , the reverse diffusion process is also modeled using an SDE:

$$d\mathbf{x}_t = [f(t, \mathbf{x}_t) - g(t)^2 \nabla_{\mathbf{x}} \log p_t(\mathbf{x}_t)] dt + g(t)d\tilde{\mathbf{w}}_t, \quad (4)$$

where  $\nabla_{\mathbf{x}} \log p_t(\mathbf{x}_t)$ , also known as the score function, guides the denoising process. The reverse-time SDE is parameterized by a neural network, which learns to predict the score function, effectively guiding the data back towards the original, noise-free distribution.

In the specific case of predicting the positions of water molecules around proteins, the score function can be represented as:

$$\nabla \log p_t(\Delta \mathbf{r} | \mathbf{y}) = -\frac{\Delta \mathbf{r}(t) - \mu(t)}{\sigma^2(t)}, \quad (5)$$

where  $\Delta \mathbf{r}(t)$  is the displacement at time  $t$ ,  $\mu(t)$  is the mean displacement, and  $\sigma^2(t)$  represents the variance at time  $t$ . This formulation helps guide the denoising process by indicating the direction and magnitude in which the noisy data should be adjusted.

During training, the model optimizes a loss function to learn the score function. The primary loss is defined as:

$$\mathcal{L}_{\text{score}} = \mathbb{E}_{t, \mathbf{x}_0, \epsilon} [\|\nabla_{\mathbf{x}} \log p_t(\mathbf{x}_t) - \mathbf{S}_{\theta}(\mathbf{x}_t, \mathbf{y}, t)\|^2], \quad (6)$$

where  $\mathbf{S}_{\theta}(\mathbf{x}_t, \mathbf{y}, t)$  is the score function predicted by the neural network parameterized by  $\theta$ . The current model was trained for 300 epochs using this loss function to optimize its performance by ensuring that it accurately learns to approximate the true gradient of the data distribution.

### 2.3.2 Reverse Diffusion and Sampling

In the reverse diffusion process, the objective is to transform noisy data back into the original data distribution by reversing the effects of the forward diffusion. This process is modeled using a stochastic differential equation (SDE), similar to the forward process but in reverse time. The reverse-time SDE is given by:

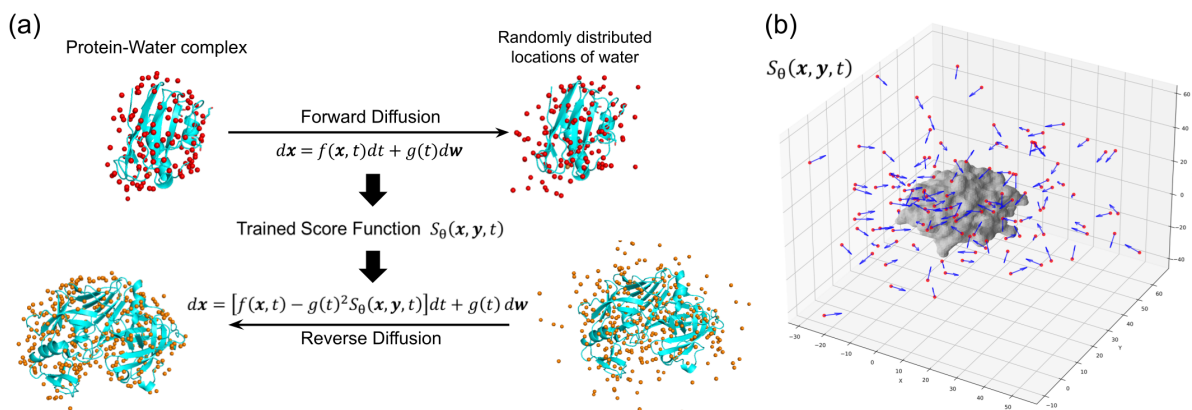
$$d\mathbf{x}_t = [f(t, \mathbf{x}_t) - g(t)^2 \mathbf{S}_{\theta}(\mathbf{x}_t, \mathbf{y}, t)] dt + g(t) d\bar{\mathbf{w}}_t, \quad (7)$$

where  $f(t, \mathbf{x}_t)$  represents the drift coefficient,  $g(t)$  is the diffusion coefficient, and  $\mathbf{S}_{\theta}(\mathbf{x}_t, \mathbf{y}, t)$  is the score function predicted by the model, which guides the denoising process. The term  $d\bar{\mathbf{w}}_t$  is a reverse Wiener process that models the stochastic component of the reverse diffusion.

The reverse diffusion process relies on the score function, which is learned during training. The score function  $\mathbf{S}_{\theta}(\mathbf{x}_t, \mathbf{y}, t)$  provides the direction and magnitude of the adjustments needed to progressively denoise the data, effectively reversing the noise added during the forward process. The model parameterized by  $\theta$  predicts the score function and guides the sampling process to recover realistic data.

To generate samples, the model starts from the final noisy state  $\mathbf{x}_T$  (which follows a Gaussian distribution) and iteratively applies the reverse SDE to gradually remove noise and restore the original structure. The step-wise application of the reverse-time SDE allows the model to produce realistic water molecule positions around the protein structure.

For each protein structure, the inference process of the diffusion model generates candidate water molecule positions, with the number of candidates being 15 times the number of protein residues. These candidates are initially distributed randomly in space based on the maximum noise level. The candidate positions are then guided by the learned score function,  $\mathbf{S}_{\theta}(\mathbf{x}_t, \mathbf{y}, t)$ , ultimately converging to their most favorable positions within the protein structure.



**Fig. 2:** Overview of the diffusion process in the protein-water complex. (a) The forward diffusion process adds noise to generate a random distribution of water molecule positions around the protein, while the reverse diffusion leverages the trained score function  $S_{\theta}(\mathbf{x}, \mathbf{y}, t)$  to guide these positions towards energetically favorable locations. (b) Visualization of the score function  $S_{\theta}(\mathbf{x}, \mathbf{y}, t)$  as a vector field, with the protein surface shown in gray and water molecules represented by red dots. The blue arrows indicate the direction of diffusion, illustrating how the score function directs water molecules toward optimal binding sites around the protein.

## 2.4 Confidence Model

The confidence model in SuperWater is based on SE(3)-equivariant convolutional networks, adapted from the previously developed diffusion model architecture [31]. The primary purpose of this model is to score and filter the water molecule positions sampled during the inference process. Specifically, the confidence model assigns a score to each sampled water molecule by evaluating its distance to the nearest crystal water molecule [20].

The training objective is to minimize the mean squared error (MSE) between the predicted confidence scores,  $p$ , and the normalized distances,  $N(d)$ :

$$\mathcal{L}_c = \text{MSE}(p, N(d)), \quad (8)$$

where the normalized distance  $N(d)$  is defined as:

$$N(d) = \left( \frac{2}{1 + e^{-\frac{s \cdot d}{\log(2)}}} - 1 \right)^2, \quad (9)$$

Here,  $d$  represents the distance between a sampled water molecule and its nearest crystal water molecule, and  $s$  is a scaling factor. The normalization function  $N(d)$  ensures that the distances fall within the  $[0, 1]$  range, facilitating smoother learning during training [20]. This normalization helps make the confidence scores interpretable and allows differentiation between high and low-confidence predictions.

Incorporating the normalized distances  $N(d)$  within the loss function allows the model to learn confidence scores that reflect the proximity of sampled water molecules to true crystal water positions. Higher confidence scores correspond to smaller distances, indicating greater accuracy. This is crucial for post-processing, as it enables effective filtering of low-confidence positions, ensuring optimal placement of water molecules in the protein structure.

## 2.5 Clustering Mechanism

The final step involves refining the predicted positions of water molecules. Initially, the predicted water molecule coordinates are filtered based on their confidence scores, retaining only those with confidence scores above a defined threshold. Next, the pairwise distance matrix of the water molecule coordinates is calculated to identify neighboring water molecules that are within the van der Waals radius of oxygen, which is 1.52 Å. For each water molecule, its confidence score is compared with those of its neighbors; if it has the highest confidence among its neighbors, it is designated as the core of a cluster, and its neighbors are added to that cluster.

For each cluster, the water molecule position is refined using a confidence-weighted average, ensuring that molecules with higher confidence have a greater influence on the final position of the cluster. If a cluster contains only one water molecule, the position of that molecule is directly taken as the cluster center. After the initial clustering, a second round of filtering is performed to prevent clashes between water molecules. The pairwise distances between cluster centroids are recalculated, and clusters with distances less than 1.52 Å are further screened to retain only the centroid with the highest confidence, while removing others in close proximity. The final output includes the refined water molecule positions along with their associated confidence scores, ensuring that the predicted water molecule locations are both accurate and unique.

## 2.6 Evaluation

To assess the accuracy of predicted water molecule positions, we utilized three primary metrics: precision, coverage (recall), and root mean square deviation (RMSD).

Precision is defined as the ratio of true positive water molecule positions (TP) to the total number of predicted positions, which includes both true positives and false positives (FP). True positives (TP) are predicted water molecule positions that correctly match the experimentally determined positions within a distance cutoff of 1 Å or 0.5 Å, while false positives (FP) are predicted positions that do not have any corresponding experimentally determined water molecule within the cutoff. Precision measures how many of the predicted positions are actually correct:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (10)$$

Coverage, also referred to as recall, represents the proportion of true water molecule positions that were correctly predicted by the model. Specifically, coverage is calculated after matching at most one crystal water to each predicted water within a distance cutoff of 1 Å or 0.5 Å. It is defined as the fraction of matched crystal water positions among all crystal water positions:



$$\text{Coverage (Recall)} = \frac{TP}{TP + FN} \quad (11)$$

RMSD quantifies the average distance between predicted water molecules and their closest experimentally determined water positions. It is calculated as:

$$\text{RMSD} = \sqrt{\frac{1}{N} \sum_{i=1}^N (\mathbf{x}_{\text{pred},i} - \mathbf{x}_{\text{true},i})^2} \quad (12)$$

where  $N$  is the number of predicted water molecules,  $\mathbf{x}_{\text{pred},i}$  represents the position of the  $i$ -th predicted water molecule, and  $\mathbf{x}_{\text{true},i}$  is the corresponding experimentally determined position. A lower RMSD value indicates that the predicted water positions are closer to the true positions, thereby reflecting higher accuracy.

Together, these metrics provide a comprehensive evaluation of the accuracy of predicted water molecule positions. Precision assesses how many of the predicted water molecules are correct, coverage measures how effectively all true water molecules are identified, and RMSD evaluates the positional accuracy of the predictions.

### 3 Results and Discussion

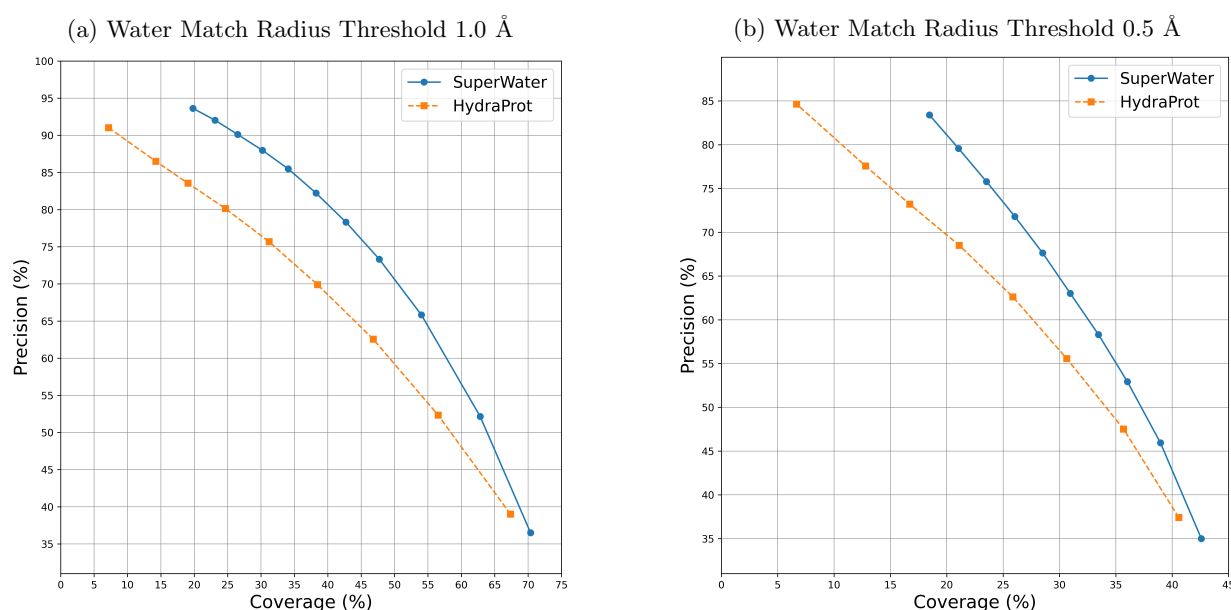
This section presents a comprehensive evaluation of SuperWater, comparing its performance with the state-of-the-art HydraProt method [20] for predicting water molecule positions around protein structures. Both methods were tested on the same dataset to ensure a fair comparison. In the inference phase of SuperWater, an initial set of water molecule positions was sampled randomly throughout the system, with the number of water molecules set to 15 times the number of protein residues to maximize coverage while balancing computational cost and accuracy. These initial positions were refined through a reverse diffusion process to adjust their translational degrees of freedom. Next, a confidence model was used to filter out low-confidence positions, retaining those most likely to correspond to true water molecule locations. Finally, a clustering algorithm was applied to further optimize and finalize the predicted water molecule positions. The dataset used for evaluation consisted of protein structures curated from the Protein Data Bank (PDB). Further details regarding dataset curation, preprocessing, and experimental conditions can be found in Methods.

#### 3.1 Comparison between SuperWater and HydraProt

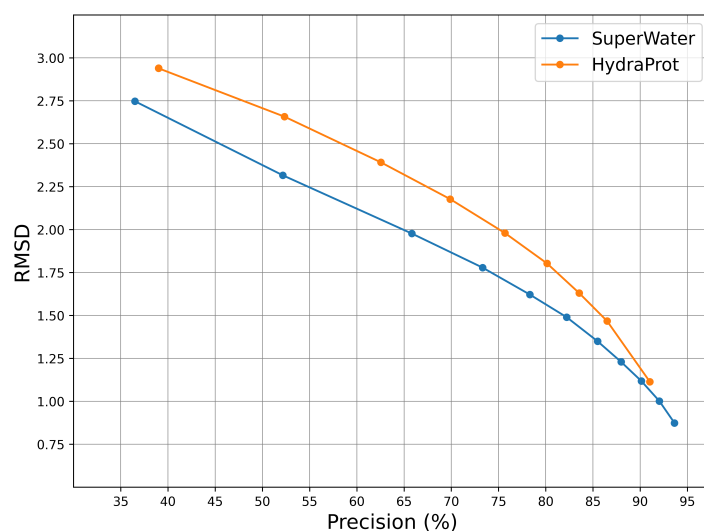
Fig. 3 illustrates the precision versus coverage curves for SuperWater and HydraProt across different thresholds. Predictive models typically exhibit a trade-off between precision and coverage, which aligns with our observations in these figures. The curves are generated by varying the prediction thresholds for both SuperWater and HydraProt.

Overall, SuperWater consistently demonstrates higher precision across a broader range of coverage levels compared to HydraProt, highlighting its superior predictive capabilities. In Fig. 3a, at a match radius threshold of 1.0 Å, SuperWater achieves 90% precision at approximately 27% coverage, whereas HydraProt reaches 90% precision at only about 8% coverage. This shows that SuperWater’s coverage is more than three times that of HydraProt at this level of precision. At a broader coverage of 54%, SuperWater maintains a precision of around 66%, significantly outperforming HydraProt, which drops to 55% precision. These results emphasize SuperWater’s ability to deliver high precision while achieving greater coverage, enhancing the reliability and scope of its predictions.

At a match radius threshold of 0.5 Å (Fig. 3b), SuperWater achieves approximately 70% precision at 27% coverage, while HydraProt reaches only about 61% precision at the same coverage level. As coverage increases to 40%, SuperWater maintains an precision of around 46%, whereas HydraProt’s precision drops to 37%. These findings underscore SuperWater’s superior performance across varying cutoff criteria.



**Fig. 3: Comparison of Precision and Coverage for SuperWater and HydraProt at Different Thresholds.** The graphs compare the performance of SuperWater (blue line) and HydraProt (orange line), showing how precision varies with coverage at different match radius thresholds: (a) 1.0 Å and (b) 0.5 Å.



**Fig. 4: Comparison of Root Mean Square Deviation (RMSD) between SuperWater (blue) and HydraProt (orange) Across Different precision Levels.** RMSD measures the average spatial deviation between all predicted water molecule positions and experimentally determined positions, serving as a metric for the overall spatial precision of predictions. Lower RMSD values indicate higher spatial accuracy in the predicted positions.

Beyond precision and coverage in predicting water-binding sites, it is also important to assess the spatial accuracy of these predictions. To achieve this, we calculated the RMSD between the predicted water positions and their corresponding nearest crystal water positions, as shown in Fig. 4. The graph reveals that SuperWater consistently yields lower RMSD values compared to HydraProt, demonstrating higher spatial accuracy across different precision thresholds. Further analysis of mean absolute deviation (MAD) between experimentally determined water positions and the correctly predicted positions (true positives) is provided in the supplementary material. At a probability threshold of  $cap = 0.5$ , SuperWater achieves a MAD of  $0.3 \pm 0.06$  Å, demonstrating the robustness of its predictions. However, in real-world applications, the actual positions of predicted water molecules are generally unknown. Therefore, we employ RMSD, calculated between all predicted positions and their nearest experimental counterparts, as a comprehensive metric to assess the model's overall spatial prediction performance.

## 3.2 Case Studies

To assess the performance and biological relevance of SuperWater, we present three case studies that showcase its capabilities in accurately predicting conserved water molecule positions across various scenarios: protein surfaces, protein-ligand binding sites, and protein-protein interaction interfaces, as depicted in Fig. 5, Fig. 6, and Fig. 7, respectively. The figures were generated using PyMOL[18], with experimental water molecules depicted in red, SuperWater-predicted water molecules in cyan, and HydraProt-predicted water molecules in yellow.

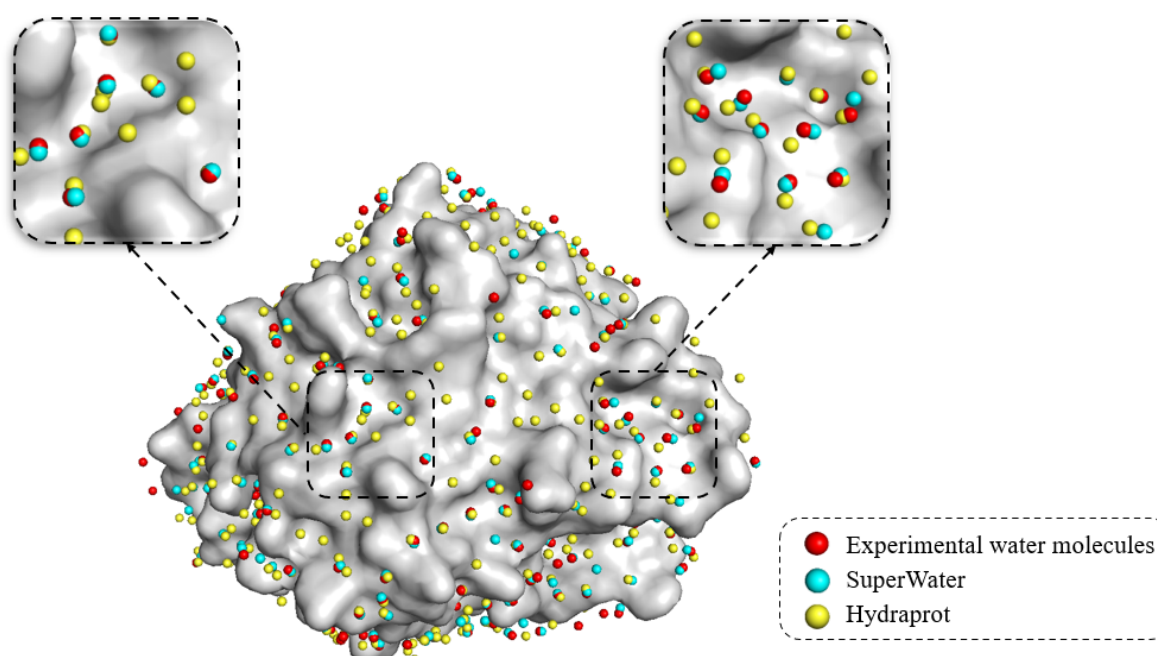
### 3.2.1 Protein Surfaces

The water molecules on proteins are necessary for their folding, stability, and functions [47]. Fig. 5 illustrates the conserved water molecule predictions on the surface of Carbonic Anhydrase II (PDB ID: 6OUH). The SuperWater-predicted water molecules (cyan) align closely with experimentally observed positions, with minimal false positives, indicating high precision. In contrast, HydraProt-predicted water molecules (yellow) are dispersed across a broader spatial region, resulting in reduced prediction precision due to numerous misplaced water molecules.

**Table 1:** Comparison of Computational Evaluation Metrics for SuperWater and HydraProt for 6OUH

Metrics	Recall		Precision		F1 Score		MAD (Å)	
	0.5	1.0	0.5	1.0	0.5	1.0	0.5	1.0
Models								
SuperWater cap 0.05	<b>0.761</b>	<b>0.851</b>	<b>0.620</b>	<b>0.693</b>	<b>0.683</b>	<b>0.764</b>	<b>0.241</b>	<b>0.288</b>
HydraProt cap 0.05	0.475	0.707	0.262	0.390	0.338	0.503	0.262	0.418

Table 1 summarizes the computational evaluation metrics for SuperWater and HydraProt on the 6OUH structure. SuperWater consistently outperforms HydraProt across recall, precision, F1 score, and mean absolute deviation (MAD) at thresholds of both 0.5 Å and 1.0 Å. The superior performance of SuperWater, particularly in terms of precision and F1 score, underscores its capability to accurately predict water molecule positions on protein surfaces.



**Fig. 5:** Visualization of experimental water molecules (red), SuperWater predicted water molecules (cyan), and HydraProt predicted water molecules (yellow) around Carbonic Anhydrase II (PDB ID: 6OUH).



### 3.2.2 Protein-Ligand Binding Sites

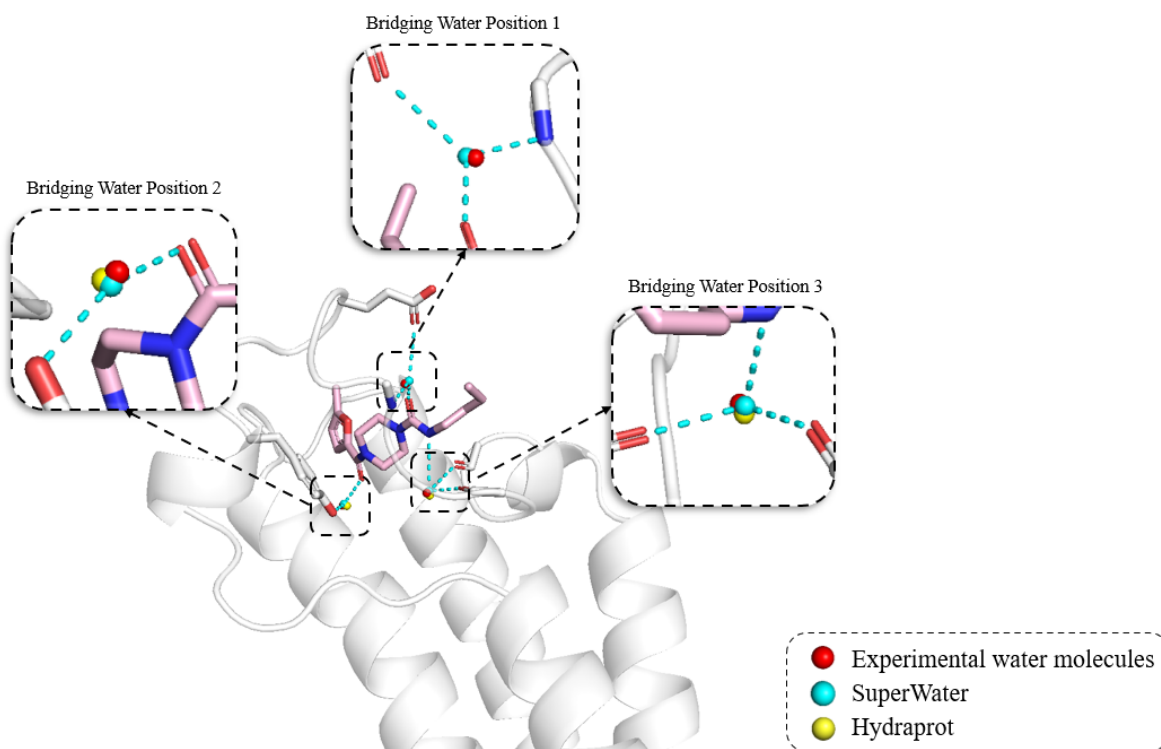
In this case study, we evaluate SuperWater’s ability to predict bridging water molecules within the binding pocket of the PHIP protein (PDB ID: 7FVP). Bridging water refers to water molecules that simultaneously form hydrogen bonds with both the protein and the ligand. These molecules play a crucial role in stabilizing the structure of the complex and facilitating the ligand recognition process [9].

Fig. 6 provides a detailed comparison of the experimental, SuperWater-predicted, and HydraProt-predicted water positions. In the binding pocket, three key bridging water molecules are highlighted, labeled as Bridging Water Position 1, Position 2, and Position 3. SuperWater accurately predicted all three bridging water positions, whereas HydraProt missed the crucial bridging water at Position 1.

**Table 2:** Comparison of Computational Evaluation Metrics for SuperWater and Hydraprot for 7FVP

Metrics	Recall		Precision		F1 Score		MAD (Å)	
Models	0.5	1.0	0.5	1.0	0.5	1.0	0.5	1.0
SuperWater cap 0.05	<b>0.705</b>	<b>0.855</b>	<b>0.581</b>	<b>0.705</b>	<b>0.637</b>	<b>0.773</b>	<b>0.231</b>	<b>0.313</b>
HydraProt cap 0.05	0.487	0.674	0.325	0.450	0.390	0.539	0.287	0.403

Table 2 summarizes the computational evaluation metrics for SuperWater and HydraProt for all experimental waters in the 7FVP structure. The results indicate that SuperWater consistently outperforms HydraProt in recall, precision, F1 score, and mean absolute deviation (MAD) at thresholds of 0.5 Å and 1.0 Å. Notably, SuperWater’s superior precision and F1 score highlight its capability to accurately predict water molecule positions that are essential for maintaining complex stability and supporting water-mediated drug design.



**Fig. 6:** Visualization of experimental water molecules (red), SuperWater-predicted water molecules (cyan), and HydraProt-predicted water molecules (yellow) around the PHIP protein (PDB ID: 7FVP). Three panels zoom into each of the three bridging water positions, which are labeled as Bridging Water Position 1, Bridging Water Position 2, and Bridging Water Position 3. Notably, SuperWater successfully captures all three bridging water positions, whereas HydraProt fails to identify the bridging water at Position 1.

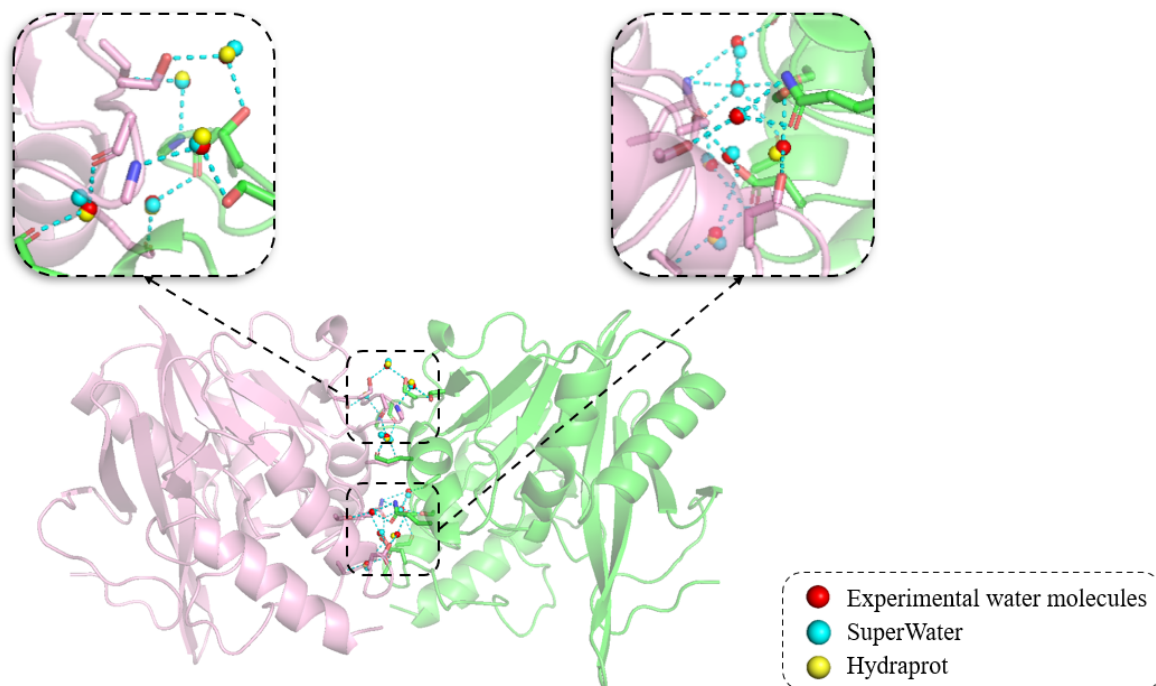
### 3.2.3 Protein-Protein Interaction Interfaces

Water molecules are also essential for mediating protein-protein interactions. Accurately predicting interfacial water molecules is essential for understanding protein interactions, which is vital for designing inhibitors or stabilizers for protein complexes. Fig. 7 highlights the bridging water molecules at the protein-protein interface of the NDM1-meropenem complex (PDB ID: 4EY2). The experimental structure reveals 12 bridging water molecules at this interface. SuperWater predicts 11 of these water molecules that form hydrogen bonds between the two protein chains with high precision, whereas HydraProt predicts only 8, with their positions falling within a 1 Å radius of the actual experimental water molecules. Further details regarding these bridging water molecules can be found in Supplementary Figs. S6 and S7.

**Table 3:** Comparison of Computational Evaluation Metrics for SuperWater and Hydraprot for 4EY2

Metrics	Recall		Precision		F1 Score		MAD (Å)	
	0.5	1.0	0.5	1.0	0.5	1.0	0.5	1.0
SuperWater cap 0.05	<b>0.496</b>	<b>0.706</b>	<b>0.361</b>	<b>0.516</b>	<b>0.418</b>	<b>0.596</b>	0.285	0.410
HydraProt cap 0.05	0.485	0.704	0.250	0.362	0.330	0.479	<b>0.265</b>	<b>0.398</b>

Table 3 summarizes the computational evaluation metrics for SuperWater and HydraProt for all experimental waters in the 4EY2 structure. The results indicate that SuperWater consistently outperforms HydraProt in terms of recall, precision, and F1 score at thresholds of 0.5 Å and 1.0 Å. However, HydraProt shows a slightly better mean absolute deviation (MAD) than SuperWater in this case. Accurately predicting interfacial water molecules is essential for understanding protein interactions, which plays a critical role in designing effective inhibitors or stabilizers for protein complexes.



**Fig. 7:** Visualization of experimental water molecules (red), SuperWater predicted water molecules (cyan), and HydraProt predicted water molecules (yellow) of protein-protein binding sites (PDB ID: 4EY2).

## 4 Conclusion

In this study, we presented SuperWater, a novel generative AI model designed for accurately predicting the positions of water molecules around protein structures. Our approach combines a score-based diffusion model with an equivariant graph neural network to generate candidate water placements, evaluate their confidence,

and refine the predictions. The performance of SuperWater was benchmarked against the state-of-the-art HydraProt model.

SuperWater demonstrated superior performance across multiple metrics, including precision, coverage, and overall spatial accuracy. Specifically, SuperWater outperformed HydraProt by achieving higher precision at broader coverage levels. For example, at a precision level of approximately 90%, SuperWater achieved nearly three times the coverage compared to HydraProt, highlighting its ability to identify a larger set of correct water positions without compromising precision. Furthermore, SuperWater consistently produced lower RMSD values, showcasing its spatial precision and robustness in predicting water locations. In all three case studies, SuperWater outperformed HydraProt in terms of recall, and precision, effectively capturing key water-mediated interactions in protein surfaces, protein-ligand binding, and protein-protein interfaces.

These results underscore SuperWater’s robustness and versatility, establishing it as a valuable tool for accurate water placement predictions on protein structures. Reliable water position prediction is crucial for understanding water-mediated biological processes, protein stability, and lead optimization in drug discovery. Although SuperWater is currently trained to predict water molecule positions, its framework can be readily adapted for diverse applications, including structural biology, metal-binding site prediction, multi-body docking, and water-mediated drug design.

**Data and Code Availability.** The processed datasets and cache files are available on Zenodo (<https://doi.org/10.5281/zenodo.14166655>). The source code for SuperWater is available on GitHub (<https://github.com/kuangxh9/SuperWater>).

**Supporting Information.** Supplementary information is available.

**Acknowledgements.** Z.S. thanks the support of the Vanderbilt Data Science Postdoctoral Fellowship. X.L. and X.K. are grateful for the research funding and support provided by the Vanderbilt Data Science Institute. Y.L. acknowledges the Nvidia hardware grant for accelerating the project development. X.L. also thanks the John R. Hall Professorship Endowment in Chemical Engineering for its support. J.L. expresses gratitude for the project opportunity provided by the Vanderbilt Data Science Institute. We sincerely thank Umang Chaudhry for facilitating access to these resources. The authors thank Tommi Jaakkola and Gabriele Corso at Massachusetts Institute of Technology for the help and guidance. We also acknowledge the computational resources (DGX A100) provided by the Vanderbilt Data Science Institute. J.M. is supported by a Humboldt Professorship of the Alexander von Humboldt Foundation. J.M. acknowledges funding by the Deutsche Forschungsgemeinschaft (DFG) through SFB1423 (421152132), SFB 1664 (514901783), TRR (514664767), and SPP 2363 (460865652). J.M. is supported by the Federal Ministry of Education and Research (BMBF) through the Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI), through the German Network for Bioinformatics Infrastructure (de.NBI), and through the German Academic Exchange Service (DAAD) via the School of Embedded Composite AI (SECAI 15766814). Work in the Meiler laboratory is further supported through the National Institute of Health (NIH) through R01 HL122010, R01 DA046138, R01 AG068623, U01 AI150739, R01 CA227833, R01 LM013434, S10 OD016216, S10 OD020154, S10 OD032234.

**Author contributions.** Z.S. conceptualized and designed the study and prepared the dataset. Z.S., X.K., Y.L., X.L., J.S., T.D., Y.W., and J.M. contributed to software development, model training, and analysis. Z.S., X.K., Y.L., J.S., T.D., Y.W., and J.M. contributed to the writing and review of the manuscript. J.S., J.M., and T.D. provided funding support.

## References

- [1] Lynn J Rothschild and Rocco L Mancinelli. Life in extreme environments. *Nature*, 409(6823):1092–1101, 2001.
- [2] Marie-Claire Bellissent-Funel, Ali Hassanali, Martina Havenith, Richard Henchman, Peter Pohl, Fabio Sterpone, David Van Der Spoel, Yao Xu, and Angel E Garcia. Water determines the structure and dynamics of proteins. *Chemical reviews*, 116(13):7673–7697, 2016.
- [3] Damien Laage, Thomas Elsaesser, and James T Hynes. Water dynamics in the hydration shells of biomolecules. *Chemical Reviews*, 117(16):10694–10725, 2017.
- [4] Matthew R Walsh, Carolyn A Koh, E Dendy Sloan, Amadeu K Sum, and David T Wu. Microsecond simulations of spontaneous methane hydrate nucleation and growth. *Science*, 326(5956):1095–1098, 2009.
- [5] Zhaoqian Su, Saman Alavi, John A Ripmeester, Gedaliah Wolosh, and Cristiano L Dias. Methane clathrate formation is catalyzed and kinetically inhibited by the same molecule: two facets of methanol. *The Journal of Physical Chemistry B*, 125(16):4162–4168, 2021.

- [6] Yaakov Levy and José N Onuchic. Water and proteins: A love-hate relationship. *Proceedings of the National Academy of Sciences*, 101(10):3325–3326, 2004.
- [7] AJ Venkatakrishnan, Anthony K Ma, Rasmus Fonseca, Naomi R Latorraca, Brendan Kelly, Robin M Betz, Chaitanya Asawa, Brian K Kobilka, and Ron O Dror. Diverse gpcrs exhibit conserved water networks for stabilization and activation. *Proceedings of the National Academy of Sciences*, 116(8):3288–3293, 2019.
- [8] Cristiano L Dias. Unifying microscopic mechanism for pressure and cold denaturations of proteins. *Physical review letters*, 109(4):048104, 2012.
- [9] Johannes Schiebel, Roberto Gaspari, Tobias Wulsdorf, Khang Ngo, Christian Sohn, Tobias E Schrader, Andrea Cavalli, Andreas Ostermann, Andreas Heine, and Gerhard Klebe. Intriguing role of water in protein-ligand binding studied by neutron crystallography on trypsin complexes. *Nature communications*, 9(1):3559, 2018.
- [10] Steve W Homans. Water, water everywhere—except where it matters? *Drug Discovery Today*, 12(13-14):534–539, 2007.
- [11] John F Darby, Adam P Hopkins, Seishi Shimizu, Shirley M Roberts, James A Brannigan, Johan P Turkenburg, Gavin H Thomas, Roderick E Hubbard, and Marcus Fischer. Water networks can determine the affinity of ligand binding to proteins. *Journal of the American Chemical Society*, 141(40):15818–15826, 2019.
- [12] Sergio E Wong and Felice C Lightstone. Accounting for water molecules in drug design. *Expert opinion on drug discovery*, 6(1):65–74, 2011.
- [13] Agnieszka A Kaczor, Agata Zieba, and Dariusz Matosiuk. The application of watermap-guided structure-based virtual screening in novel drug discovery. *Expert Opinion on Drug Discovery*, 19(1):73–83, 2024.
- [14] Oliviero Carugo and Domenico Bordo. How many water molecules can be detected by protein crystallography? *Acta Crystallographica Section D: Biological Crystallography*, 55(2):479–483, 1999.
- [15] Tom Young, Robert Abel, Byungchan Kim, Bruce J Berne, and Richard A Friesner. Motifs for molecular recognition exploiting hydrophobic enclosure in protein–ligand binding. *Proceedings of the National Academy of Sciences*, 104(3):808–813, 2007.
- [16] Zheng Li and Themis Lazaridis. Computing the thermodynamic contributions of interfacial water. *Computational Drug Discovery and Design*, pages 393–404, 2012.
- [17] Crystal N Nguyen, Tom Kurtzman Young, and Michael K Gilson. Grid inhomogeneous solvation theory: hydration structure and thermodynamics of the miniature receptor cucurbit [7] uril. *The Journal of chemical physics*, 137(4), 2012.
- [18] Bingjie Hu and Markus A Lill. Watsite: Hydration site prediction program with pymol interface, 2014.
- [19] Shannon T Smith, Laura Shub, and Jens Meiler. Placwaters: Real-time, explicit interface water sampling during rosetta ligand docking. *Plos one*, 17(5):e0269072, 2022.
- [20] Andreas Zamanos, George Ioannakis, and Ioannis Z Emiris. Hydraprot: A new deep learning tool for fast and accurate prediction of water molecule positions for protein structures. *Journal of Chemical Information and Modeling*, 64(7):2594–2611, 2024.
- [21] Sangwoo Park and Chaok Seok. Galaxywater-cnn: Prediction of water positions on the protein structure by a 3d-convolutional neural network. *Journal of Chemical Information and Modeling*, 62(13):3157–3168, 2022.
- [22] Ahmadreza Ghanbarpour, Amr H Mahmoud, and Markus A Lill. Instantaneous generation of protein hydration properties from static structures. *Communications Chemistry*, 3(1):188, 2020.
- [23] Kochi Sato, Mao Oide, and Masayoshi Nakasako. Prediction of hydrophilic and hydrophobic hydration structure of protein by neural network optimized using experimental data. *Scientific reports*, 13(1):2183, 2023.
- [24] Zhijian Liu, Haotian Tang, Yujun Lin, and Song Han. Point-voxel cnn for efficient 3d deep learning. *Advances in neural information processing systems*, 32, 2019.
- [25] Yang Zhang, Wenbing Huang, Zhewei Wei, Ye Yuan, and Zhaohan Ding. Equipocket: an e (3)-equivariant geometric graph neural network for ligand binding site prediction. *arXiv preprint arXiv:2302.12177*, 2023.
- [26] Sangwoo Park. Water position prediction with se (3)-graph neural network. *bioRxiv*, pages 2024–03, 2024.
- [27] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.
- [28] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.
- [29] Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach,



- Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De novo design of protein structure and function with rdiffusion. *Nature*, 620(7976):1089–1100, 2023.
- [30] Emiel Hoogetboom, Victor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d. In *International conference on machine learning*, pages 8867–8887. PMLR, 2022.
- [31] Gabriele Corso, Hannes Stärk, Bowen Jing, Regina Barzilay, and Tommi Jaakkola. Diffdock: Diffusion steps, twists, and turns for molecular docking. *arXiv preprint arXiv:2210.01776*, 2022.
- [32] Mohamed Amine Ketata, Cedrik Laue, Ruslan Mammadov, Hannes Stärk, Menghua Wu, Gabriele Corso, Céline Marquet, Regina Barzilay, and Tommi S Jaakkola. Diffdock-pp: Rigid protein-protein docking with diffusion models. *arXiv preprint arXiv:2304.03889*, 2023.
- [33] Karsten Kreis, Tim Dockhorn, Zihao Li, and Ellen Zhong. Latent space diffusion models of cryo-em structures. *arXiv preprint arXiv:2211.14169*, 2022.
- [34] Dominik JE Waibel, Ernst Röell, Bastian Rieck, Raja Giryes, and Carsten Marr. A diffusion model predicts 3d shapes from 2d microscopy images. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2023.
- [35] Zhiye Guo, Jian Liu, Yanli Wang, Mengrui Chen, Duolin Wang, Dong Xu, and Jianlin Cheng. Diffusion models in bioinformatics and computational biology. *Nature reviews bioengineering*, 2(2):136–154, 2024.
- [36] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1):235–242, 2000.
- [37] Pedro M Martins, Lucianna H Santos, Diego Mariano, Felipe C Queiroz, Luana L Bastos, Isabela de S Gomes, Pedro HC Fischer, Rafael EO Rocha, Sabrina A Silveira, Leonardo HF de Lima, et al. Propedia: a database for protein–peptide identification based on a hybrid clustering algorithm. *BMC bioinformatics*, 22:1–20, 2021.
- [38] Vincent Frappier, Madeleine Duran, and Amy E Keating. Pixeldb: Protein–peptide complexes annotated with structural conservation of the peptide binding mode. *Protein Science*, 27(1):276–285, 2018.
- [39] Sam M Ireland and Andrew CR Martin. Zincbind—the database of zinc binding sites. *Database*, 2019:baz006, 2019.
- [40] Pin Huang, Haoming Xing, Xun Zou, Qi Han, Ke Liu, Xiangyan Sun, Junqiu Wu, and Jie Fan. Accurate prediction of hydration sites of proteins using energy model with atom embedding. *Frontiers in Molecular Biosciences*, 8:756075, 2021.
- [41] George A Jeffrey and Wolfram Saenger. *Hydrogen bonding in biological structures*. Springer Science & Business Media, 2012.
- [42] GA Jeffrey. An introduction to hydrogen bonding, 1997.
- [43] Zhaoqian Su and Cristiano L Dias. Individual and combined effects of urea and trimethylamine n-oxide (tmao) on protein structures. *Journal of Molecular Liquids*, 293:111443, 2019.
- [44] Xianfeng Chen, Irene Weber, and Robert W Harrison. Hydration water and bulk water in proteins have distinct properties in radial distributions calculated from 105 atomic resolution crystal structures. *The Journal of Physical Chemistry B*, 112(38):12073–12080, 2008.
- [45] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33:12438–12448, 2020.
- [46] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021.
- [47] Masayoshi Nakasako, Masafumi Odaka, Masafumi Yohda, Naoshi Dohmae, Koji Takio, Nobuo Kamiya, and Isao Endo. Tertiary and quaternary structures of photoreactive fe-type nitrile hydratase from rhodococcus sp. n-771: roles of hydration water molecules in stabilizing the structures and the structural origin of the substrate specificity of the enzyme. *Biochemistry*, 38(31):9887–9898, 1999.