Generative Modeling of Solvated Biomolecules

MinGyu Choi [*1] **Derek Chen** [*1] **Gloria Ma** [*1] **Tommi S. Jaakkola** [1] **Regina Barzilay** [1]

Abstract

We introduce BoltzW, the first generative framework for solvated biomolecules. BoltzW models the joint distribution of proteins, nucleic acids, small molecules, and water. Explicit solvent modeling poses novel challenges, each concerning limited modeling capacity, data inconsistency, physical indistinguishability, and variable cardinality. We address these by selecting structured solvents, designing permutation-invariant training objectives, and formulating unified structure targets that bypass cardinality prediction. BoltzW improves biomolecular structure prediction with minimal finetuning from Boltz-2, while achieving state-of-the-art water placement compared to imputation methods that have access to the ground truth protein structure.

1 Introduction

Understanding how solvents interact with biomolecules is a central question in biochemistry, with implications for protein folding, ligand binding, and enzyme catalysis. Solvents mediate hydrogen bonding, stabilize secondary and tertiary structure, modulate electrostatics, and participate in binding thermodynamics and reaction mechanisms (Breiten et al., 2013; Samways et al., 2021; Wang et al., 2011; Zsidó & Hetényi, 2025). Most structure generation models, however, ignore solvents and generate biomolecular structures alone (Section 2.1), due to three key challenges: (i) *Modeling Capacity/Data Inconsistency*: Biomolecules are surrounded by thousands of solvents, making exhaustive modeling computationally prohibitive. Moreover, solvent positions are inconsistently resolved across crystal structures. (ii) *Permutation Invariance*: Solvents are physically indistinguishable, leaving no canonical correspondence between predicted and ground truth positions on which to anchor a stable training objective. (iii) *Cardinality*: The number of solvents is not determined *a priori* and depends circularly on the structure being predicted. The absence of solvation modeling introduces biases in structure prediction, which propagates to downstream biophysical analyses such as solvent imputation (Section 2.2).

[1]Massachusetts Institute of Technology. Correspondence to: MinGyu Choi <chemgyu@mit.edu>.

Accepted at the 2026 Workshop on Generative and Agentic AI for Biology (ICML 2026)

We advocate for modeling a true joint distribution, based on our probe experiments showing that the internal representation of current structure prediction model is not sufficient for solvent cardinality inference, even though hydration-related information is implicitly encoded in the learned representation (Section 3). We hypothesize that solvent configuration acts as a corrective signal, reducing biases in structural predictions while capturing physically accurate hydration patterns that can inform downstream biochemical analysis. To address the above challenges, we introduce three core techniques: (i) *Structured Solvent Selection* (Section 4.1): We find that limiting the generation target to structured solvents based on hydrogen-bond coordination and B-factor effectively reduces computational cost while yielding the greatest improvement in structure prediction. (ii) *Permutation-invariant Objectives* (Section 4.2): We design general permutation-invariant training objectives based on optimal transport matching. (iii) *Unified Cardinality Prediction* (Section 4.3): We bypass cardinality prediction by reformulating it as structure prediction. We introduce fake solvents at degenerate positions and learn their placement within a unified objective. Geometric decoding then removes predicted fake solvents, yielding a clean biomolecular structure.

With these advances, we propose BoltzW, the first joint generative framework for solvated biomolecules. We show that the joint generation improves biomolecular structure prediction with minimal finetuning from an existing model (Section 5.1). Furthermore, we also demonstrate that the performance gain is physically grounded, as BoltzW simultaneously outperforms water imputation methods despite operating without the ground truth biomolecular structure on which those baselines depend (Section 5.2).

Contributions. Our main results are summarized as follows:

- Probe analysis shows that current structure representation contains insufficient solvent information, motivating joint generative modeling (Section 3).
- We introduce BoltzW, the first joint generative framework; we introduce structured solvent selection criteria, permutation-invariant optimal transport objectives, and fake-water placement (Section 4).
- BoltzW achieved state-of-the-art performance on both structure and solvent prediction, with minimal finetuning and without ground-truth structure (Section 5).

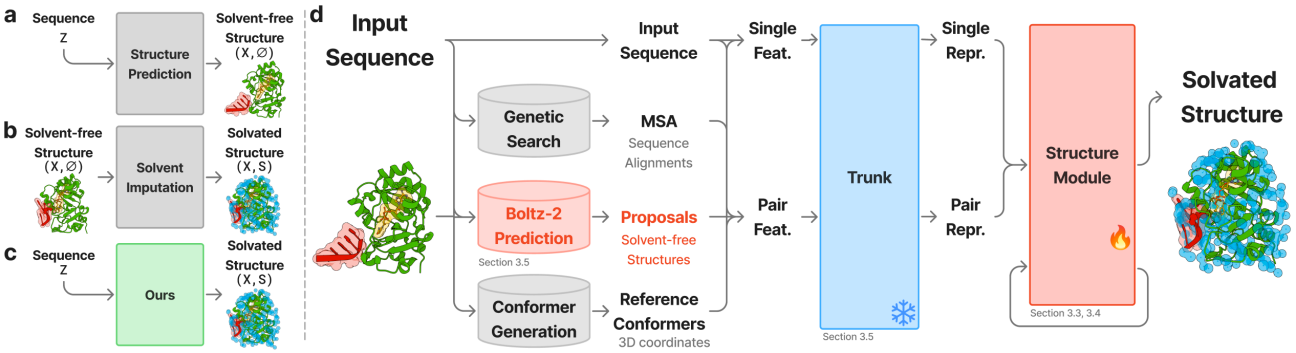

Figure 1. **Overview. a.** Structure prediction methods generate biomolecular structures given sequence. **b.** Solvent imputation methods predict solvent configurations given a fixed ground-truth biomolecular structure. **c.** BoltzW generates solvated biomolecular structures directly from the sequence. **d.** BoltzW improves biomolecular structure prediction with minimal finetuning from Boltz-2. BoltzW is optionally conditioned on Boltz-2 predicted structure, correcting biases while co-generating solvents.

2 Related Works

2.1 Structure Prediction

Generative structure prediction models are conditional diffusion frameworks that generate biomolecular conformations from sequence-level inputs (Abramson et al., 2024; Passaro et al., 2025; Stark et al., 2025). These models typically consist of two stages: representation learning and structure generation (see Figure 1d). Input sequences, together with evolutionary and structural context such as multiple sequence alignments (MSAs) and backbone templates, are processed by a trunk network that produces latent single and pair representations. Conditioned on these representations, a structure module iteratively denoises atomic coordinates to predict a clean conformation. The generative process is performed in Euclidean space. These models have substantially advanced structure prediction and enabled downstream studies of protein folding, ligand binding, and conformational dynamics (Jing et al., 2024). However, their latent representations are primarily optimized for recovering biomolecular geometry and do not explicitly model surrounding solvent environments, limiting their ability to capture solvent-coupled physical processes such as hydration-mediated conformational rearrangements and enzyme catalysis.

2.2 Solvent Imputation

If ground-truth biomolecular structure is accessible, one can *impute* solvents onto a fixed structure, corresponding to a factorized distribution (Equation 1). *Physics-based approaches* provide reliable estimates based on a physical foundation, including simulation-based methods (Hu & Lill, 2014; Ross et al., 2012; Sridhar et al., 2017; Yang et al., 2017), sampling-based methods such as grand-canonical Monte Carlo (Ross et al., 2015; Woo et al., 2004) and inhomogeneous solvation theory (Sindhikara & Hirata, 2013). However, they suffer from substantial computational overhead, slow convergence, and sensitivity to force-field parameterization, limiting their applicability in large-scale or learning-integrated pipelines (Hinz et al., 2025). *Data-driven alternatives* offer fast, amortized inference without the need for explicit simulation. GalaxyWater discretizes the protein environment onto voxel grids and predicts water occupancy (Park & Seok, 2022; Zamanos et al., 2024). More recently, SuperWater uses a diffusion framework to generate waters (Kuang et al., 2024). In practice, however, when using these models to hydrate a sample, only the *predicted* biomolecular structure is accessible, and it is biased due to the absence of explicit solvent modeling (Abramson et al., 2024; Geffner et al., 2025; Passaro et al., 2025; Stark et al., 2025). Errors from the biased biomolecular generator then propagate irreversibly into solvent imputation, yielding an incorrect sample distribution.

3 Probes Cannot Estimate Solvation

Setup and Notation. Let $X = (x_1, \ldots, x_N) \in \mathbb{R}^{3N}$ denote the ground-truth biomolecular structure as an ordered sequence of N atoms (from amino acids, nucleotides, or small molecules), where each x_i encodes the coordinates of the i-th atom. Let $\mathcal{S} = \{s_1, \ldots, s_M\} \subset \mathbb{R}^3$ denote an *unordered* set of M ground-truth solvent (water) coordinates. We write $(X, \mathcal{S})$ for a solvated biomolecular state.

3.1 Local Hydration Prediction under Known Structure

Although existing structure prediction models do not explicitly parameterize solvent configurations, they are trained on experimental structures whose conformations are shaped by solvent interactions. One may therefore hypothesize that hydration-related information is implicitly encoded in their learned representations. We examine whether these representations expose sufficient information for *post hoc* solvent inference, or whether explicit solvent supervision is required. To study this, we train a probe on the internal representations of a pretrained structure prediction model to predict solvent cardinality; failure indicates that hydration is not easily recoverable from the representation.

Sufficient Statistics Hypothesis. We formalize the hypothesis through a latent-variable model. Let Z denote a latent representation encoding the statistical and physical degrees of freedom sufficient for biomolecular structure prediction but not uniquely determined by the observed structure X alone (e.g., protonation states, electronic environments, and local rearrangements). In our setting, Z corresponds to the evolutionary representation of a structure prediction model (Abramson et al., 2024; Passaro et al., 2025) (Section 2.1). The conditional distribution over biomolecular structures and solvent configurations is then decomposed as

$$p(X, \mathcal{S} \mid Z) = p(X \mid Z)\, p(\mathcal{S} \mid X, Z). \tag{1}$$

This decomposition motivates the two-stage strategy adopted in prior solvent imputation methods: first generate a biomolecular structure $X \sim p(X \mid Z)$, then infer solvent configurations through a conditional imputation model $\mathcal{S} \sim p(\mathcal{S} \mid X, Z)$ (Kuang et al., 2024; Park & Seok, 2022). Implicitly, this assumes that (X, Z) retains sufficient information for downstream solvent inference.

However, since Z is optimized solely for structure prediction, it remains unclear whether hydration-related information is sufficiently preserved for downstream solvent inference. We therefore ask whether such information is recoverable from (X, Z) using downstream predictors.

Cardinality Prediction as a Solvent Proxy. Although solvent cardinality is only a coarse observable of hydration structure, it is induced by the underlying solvent configuration.

Remark 3.1. Let $p^\star(\mathcal{S} \mid X, Z)$ denote the physical distribution over solvent configurations conditioned on biomolecular structure X and latent representation Z. Any notion of solvent cardinality M can be expressed as a deterministic functional $M = g(\mathcal{S})$, where g aggregates solvent coordinates. The induced conditional distribution over solvent cardinality is therefore

$$p^\star(M \mid X, Z) = \int \delta(M - g(\mathcal{S}))\; p^\star(\mathcal{S} \mid X, Z)\, d\mathcal{S}. \tag{2}$$

Thus, while accurate prediction of M does not require reconstructing the full solvent configuration $\mathcal{S}$, it nevertheless requires representations that retain solvent-dependent structural information.

Motivated by this observation, we use solvent cardinality prediction as a proxy task for probing hydration information in existing structure representations. A representation that fails even on this low-dimensional task is unlikely to support accurate recovery of more detailed solvent organization.

Empirical Validation. We construct a controlled setting in which the ground-truth biomolecular structure X and evolutionary representation from the trunk Z are provided as input, with all modules frozen except a prediction head following the confidence module architecture (Abramson et al., 2024; Passaro et al., 2025). We evaluate solvent cardinality predictability under three tasks of increasing complexity: binary hydration, discrete cardinality, and fractional cardinality (full definitions in Appendix A). As shown in Table 3, the model achieves an AUPRC of 0.850 on binary hydration, but performance collapses on both cardinality tasks ($R^2 = -0.12$ and -0.05), falling below a simple mean predictor. This failure persists when restricting targets to geometrically constrained, reproducible waters (Section 4.1), suggesting that hydration-related information is not readily recoverable from representations learned without explicit solvent supervision.

3.2 Imputed/Ground-truth Solvent Discrimination

Section 3.1 shows that frozen structure-model's residue representations do not expose enough information for reliable solvent cardinality prediction. We next ask a complementary question: whether these representations nevertheless contain fine-grained local chemical information relevant to identifying physically realized hydration sites. Rather than predicting how many waters should occur near a residue from the residue representation, we train a binary probe on water representation to distinguish experimentally observed structured waters from geometrically plausible imputed waters.

We define positive waters as crystallographic waters that form hydrogen bonds with at least three distinct biomolecular residues, following the structured-water criterion described in Section 4.1. Negative examples are generated by a geometric imputation algorithm that enumerates candidate sites satisfying local hydrogen-bond constraints. Through sweeping over clash thresholds, the imputation procedure is able to recover nearly all ($\geq 90\%$) ground-truth structured waters (Figure 5). Algorithmic imputation also recovers a similar distribution over the number of hydrogen-bond partner atoms, donor/acceptor role, and residue types of interacting partners as real structured solvents (see Appendix B). These negatives are therefore not arbitrary decoys: they occupy locally plausible hydration sites under simple geometric rules. Thus, successful discrimination by our classifier probe cannot rely only on coarse hydrogen-bond count or residue identity. It must exploit finer-grained contextual information present in the frozen trunk embeddings.

The probe is implemented as an auxiliary head attached to the frozen trunk. The head receives detached trunk features together with a pairwise residue-center distogram embedding, which encodes pairwise distances between residue centers: C_α for amino acids and $C1'$ for nucleotides. The head does not receive ground-truth atom coordinates beyond this residue-level positional information. Any atom-level signal useful for classifying real versus imputed hydration sites must therefore be encoded implicitly in the trunk representation, rather than the coordinate injection.

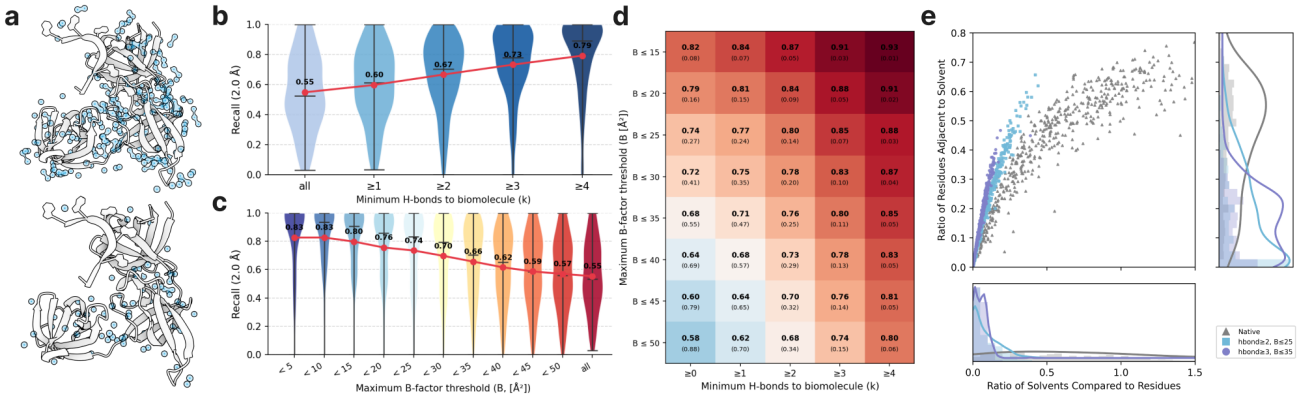

Figure 2. **Structured solvent selection. a.** Biomolecular structure before (upper) and after (lower) filtering. **b-c.** Reproducibility increases with hydrogen-bond coordination and decreases with B-factor, confirming that geometrically anchored and thermally localized waters are more consistently observed across replicates. **d.** Combining both criteria yields the highest reproducibility. Numbers denote recall, with the preserved ratio of waters in parentheses. **e.** Filtering substantially removes bulk solvents, resulting in a more compact structured subset, as shown by the linearity of total solvent ratio and solvent-interacting residue ratio.

With a positive class prevalence of 0.48, the full trunk probe achieves an AUPRC of 0.90, substantially above random chance. Adding 0.2 Å Gaussian noise to the imputed solvent coordinates leaves performance essentially unchanged, suggesting that the classifier is not simply memorizing deterministic artifacts of the imputation algorithm. In contrast, ablating the trunk embeddings and retaining only the token-level geometric injection reduces AUPRC to 0.59, indicating that residue-level coordinates alone contain limited signal for distinguishing real from imputed sites. Therefore, we hypothesize that frozen trunk embeddings with explicit solvent tokens encode substantial local hydration-relevant information, such as steric accessibility and atom-level exclusion constraints that helps distinguish real structured waters from plausible imputed decoys.

These results refine the conclusion of Section 3.1. Residue representations from current structure prediction models do not organize solvent information well enough to recover global cardinality directly, but with explicit solvent tokens they do encode local hydration-relevant chemistry that can distinguish plausible candidate sites from physically realized ones. This supports our larger claim: current representations contain latent solvent information, but explicit solvent supervision and joint biomolecule–solvent generation are needed to turn that information into accurate solvent prediction.

4 Joint Biomolecule–Solvent Generation

Having established that existing representations are insufficient for solvent inference, we now turn to the problem of joint generation. We address three key challenges identified in Section 1: structured solvent selection to define a tractable and consistent generation target (Section 4.1), permutation-invariant training objectives to handle the physical indistinguishability of solvent molecules (Section 4.2), and a fake solvent formulation to bypass explicit cardinality prediction (Section 4.3).

4.1 Defining the Solvent Generation Target

Unlike biomolecular structures, whose consistency is guaranteed by covalent geometry and mature structure determination pipelines (Liebschner et al., 2019; Murshudov et al., 2011), crystallographic waters depend on resolution, crystal packing, and refinement protocol (Liebschner et al., 2013; Nakasako, 2004; Wlodawer et al., 2024). A generative model trained naively on all crystallographic solvents would therefore be asked to reproduce a target that is itself inconsistent. Furthermore, not all solvents contribute equally to thermodynamic observables: bulk solvents can be captured implicitly through an averaged background (Feig & Brooks III, 2004), while a structured subset participates in stable hydrogen-bond networks and directly shapes the free-energy landscape (derivation at Appendix C) (Huggins, 2015; Mobley & Dill, 2009). Restricting the generation target to this structured subset therefore not only improves supervision quality but also aligns the modeling objective with the solvents that matter most for thermodynamics.

Structured Water Criterion. We restrict attention to a structured subset $\mathcal{S}_{\text{str}} \subset \mathcal{S}$, treating the remaining bulk solvents $\mathcal{S}_{\text{bulk}} = \mathcal{S} \setminus \mathcal{S}_{\text{str}}$ as background, where only $\mathcal{S}_{\text{str}}$ serves as the generation target. Throughout the remainder of this paper, we use $\mathcal{S}$ to refer to $\mathcal{S}_{\text{str}}$.

We identify $\mathcal{S}_{\text{str}}$ using two complementary criteria: hydrogen-bond coordination $\text{HB}(s, X)$, which anchors solvents to geometrically predictable regions (Laage et al., 2017; Liebschner et al., 2013), and crystallographic B-factor, which measures thermal localization and observational reliability (Klyshko

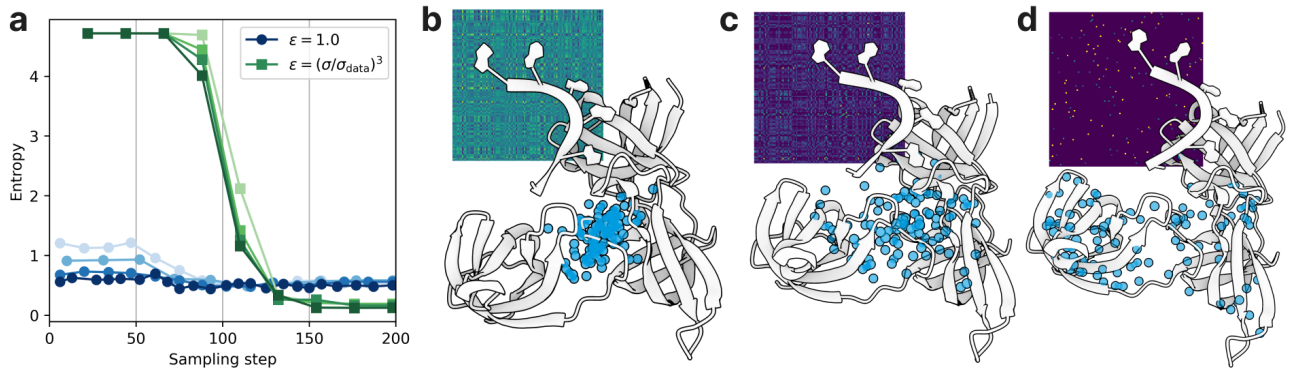

Figure 3. Entropic optimal transport. **a.** Mean row-wise entropy of the optimal transport plan P_S, over the training (darker means later checkpoints) and sampling steps. Annealed regularization provides sufficient entropy on high noise levels while sharper assignments on low noise level. **b-d.** Optimal transport plan and predicted structure at sampling steps 44, 110, and 176, respectively.

et al., 2023; Nakasako, 2004). We formalize this criterion as

$$\mathcal{S}_{\text{str}} = \left\{ s \in \mathcal{S} \;\middle|\; \text{HB}(s, X) \geq k,\ B(s) \leq \tau \right\}, \tag{3}$$

where $k \in \mathbb{Z}_{\geq 0}$ denotes the minimum hydrogen-bond coordination and $\tau \in \mathbb{R}_{>0}$ the maximum allowable B-factor. We use $k = 3$, $\tau = 35$ in practice.

Empirical Validation. We analyze cross-replicate reproducibility of water positions by grouping PDB entries with identical sequences and filtering for structural similarity (backbone RMSD $\leq$ 2 Å), resulting in 1,702 groups comprising 5,029 PDB structures, with group sizes ranging from 2 to 365. We observe clear correlations between water reproducibility and both hydrogen-bond coordination (Figure 2b) and B-factor (Figure 2c), with their combination providing improved selectivity (Figure 2d). Under the joint criterion ($\text{HB} \geq 3,\ B \leq 35$), we obtain a recall of 0.80 while retaining approximately 11% of ground-truth waters. The structured nature of the filtered subset is further supported by an approximately linear relationship between the total solvent-to-residue ratio and the fraction of residues interacting with solvent (Figure 2e).

4.2 Generative Modeling with Permutation Invariance

Biomolecules have a known correspondence to their generation target, allowing training with standard denoising diffusion objectives (Abramson et al., 2024; Passaro et al., 2025). Solvents, however, are physically indistinguishable – there is no one-to-one correspondence with the target. We therefore require permutation-invariant training objectives for a solvented system. Let $\hat{\mathbf{X}}$ and $\hat{\mathcal{S}}$ denote biomolecule and solvent coordinates predicted from the model. Similarly, let $\mathbf{X}$, $\mathcal{S}$ denote the corresponding ground-truth coordinates. All coordinates are in a common global reference frame, after rigid alignment between $\hat{\mathbf{X}}$ and $\mathbf{X}$. We also introduce ordered full-atom representations $\mathbf{Y} = (\mathbf{X}, \mathcal{S})$ and $\hat{\mathbf{Y}} = (\hat{\mathbf{X}}, \hat{\mathcal{S}})$ both in $\mathbb{R}^{3(N+M)}$, where the orders of $\mathcal{S}$ and $\hat{\mathcal{S}}$ are arbitrary and carry no physical meaning.

Entropic Optimal Transport Matching. To obtain a soft correspondence, we use entropically regularized optimal transport (Cuturi, 2013; Genevay et al., 2018). Let $K_{ij} = \|\hat{\mathbf{s}}_i - \mathbf{s}_j\|_2^2$ denote the pairwise solvent cost matrix. We define a solvent transport plan with regularizer $\epsilon \in \mathbb{R}_{\geq 0}$

$$P_S = \arg\min_{P \in \mathcal{U}} \Big(\sum_{i,j} P_{ij} K_{ij} + \varepsilon \sum_{i,j} P_{ij} \log P_{ij} \Big), \tag{4}$$

where $\mathcal{U}$ is the set of doubly stochastic matrices with uniform marginals, $P\mathbf{1} = \frac{1}{n}\mathbf{1}$, $P^\top \mathbf{1} = \frac{1}{n}\mathbf{1}$. We then extend this to full-atom transport plan $P = \text{diag}(I_N, P_S)$, which preserves the fixed ordering of biomolecular atoms while softly matching solvent atoms.

Permutation-invariant MSE. We define the permutation-invariant mean squared error as

$$\mathcal{L}_{\text{MSE-OT}}(\mathbf{Y}, \hat{\mathbf{Y}}) = \frac{1}{N+M} \sum_{i=1}^{N+M} \left\| Y_i - (\hat{Y}P)_i \right\|_2^2. \tag{5}$$

Permutation-invariant lDDT. We incorporate permutation-invariant local structural consistency. Let $D, \hat{D} \in \mathbb{R}^{(N+M)\times(N+M)}$ denote the pairwise distance matrices of Y and $\hat{Y}$, respectively. We first transport $\hat{D}$ using P, and then compare it against D:

$$\mathcal{L}(\mathbf{Y}, \hat{\mathbf{Y}}) = \frac{1}{\sum M_{ij}} \sum_{i,j} M_{ij} \odot \Big(D_{ij} - (P^\top \hat{D} P)_{ij} \Big). \tag{6}$$

where M is the standard lDDT locality mask evaluated on $\mathbf{Y}$.

Geometric Interpretation and Training Dynamics. Under Equation 5, each predicted solvent coordinate is attracted toward a transport barycenter $\bar{Y}_k(P) = \frac{\sum_i P_{ki} Y_i}{\sum_i P_{ki}}$ (similarly for Equation 6). Importantly, the sharpness of P is not fixed but evolves during training (Figure 3a, blue). In early stages, P exhibits high entropy, leading to diffuse barycenters that

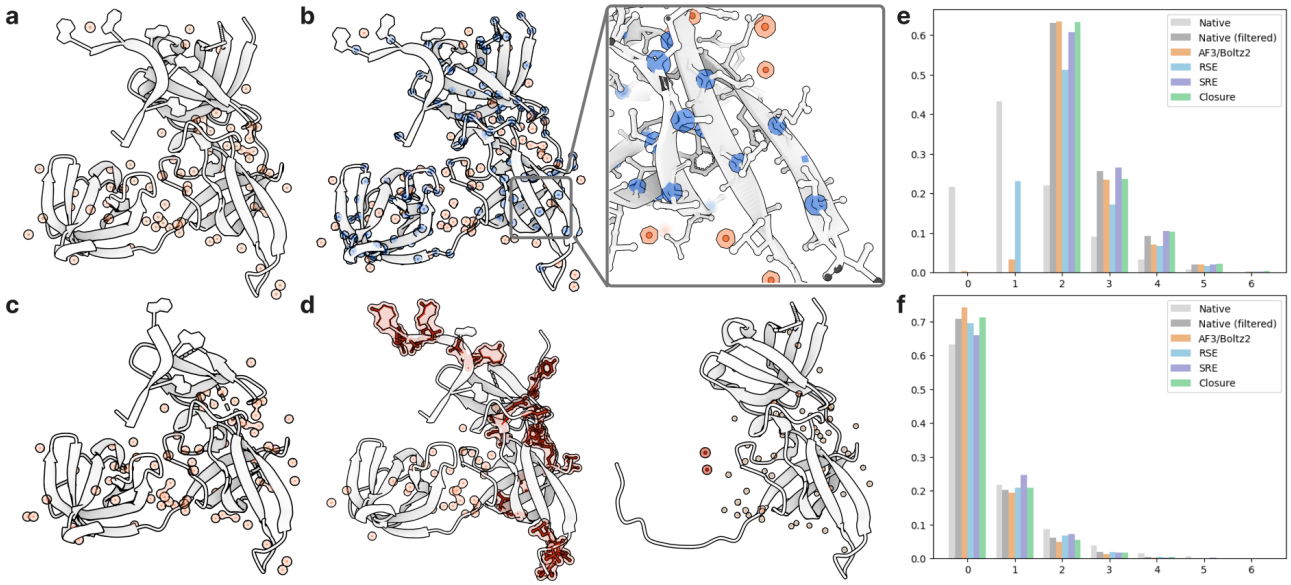

Figure 4. Fake solvent and hydrogen-bond closure cropping. **a.** Ground truth structure (PDBID: 7ZHH) with orange waters. **b.** Fake waters (blue) at C_α, C_1 positions (Section 4.3) **c.** Crop from hydrogen-bond closure (Section 4.4) **d.** Crops from (Abramson et al., 2024; Passaro et al., 2025), resulting orphan tokens. **e-f.** hydrogen-bond statistics, centered by waters (e) and biomolecules (f), respectively.

average over multiple ground-truth sites. As training progresses, P becomes peaked, reducing entropy and causing each prediction to specialize toward a single target.

Annealing. With fixed regularization (Figure 3a), the gap between entropies in the initial and final sampling steps is too small, implying that the regularization is not serving its function and is producing sharp correspondences from early timesteps. On the other hand, with too large an ϵ, the model does not receive a sufficiently sharp learning signal in the low-noise regime, resulting in poor fine-structure generation. To balance these two effects, we apply an annealing schedule for ϵ that depends on the noise level, i.e., $\epsilon = (\sigma/\sigma_{\text{data}})^3$, which provides a smoother transition (Figure 3).

Symmetry Breaking. A potential failure mode of OT-based matching is the existence of degenerate configurations. If multiple predicted points occupy exactly the same position it yields identical transport costs K and therefore identical assignments. In practice, this degeneracy is avoided by the stochastic noise inherent in the diffusion process: infinitesimal perturbations break such symmetries.

Alignment. Since the architecture is not SE(3)-invariant, $\hat{\mathbf{X}}$ and $\mathbf{X}$ must be aligned. In our setting, solvent correspondence is undefined prior to alignment: the cost matrix is only meaningful once a canonical frame is established. We therefore align exclusively on biomolecular atoms with solvent excluded, and then compute K in the aligned frame. This decoupling is valid provided that $\hat{\mathbf{X}}$ is a sufficiently accurate estimate of $\mathbf{X}$, so that the alignment defines a stable reference frame for solvent matching. We make this assumption reasonable by initializing the model from Boltz-2 (Passaro et al., 2025).

4.3 Handling Unknown Solvent Cardinality

A fundamental challenge in solvent generation is that the cardinality M is not known *a priori*: although Section 4.2 assumed M to be given, in realistic settings it must itself be inferred. Moreover, even when the ground-truth biomolecular structure X is known, predicting M remains non-trivial (Section 3). This motivates a novel strategy that avoids explicit cardinality prediction.

Fake Solvent Formulation. Inspired by the "fake atom" strategy in BoltzGen (Stark et al., 2025), we reformulate cardinality prediction as a byproduct of structure prediction. We initialize an overcomplete set of solvents, by sampling a total number of solvent tokens $\hat{M} \sim \mathcal{N}(\lambda N, \sigma N)$ where $\lambda = 0.25, \sigma = 0.025$ in practice (See Figure 2e). We place M solvent tokens at the true solvent positions and assign the remaining $\hat{M} - M$ tokens to degenerate "fake" locations, overlapping with reference atoms (C_α for amino acids and C_1 for nucleotides, see Figure 4b). The model is then trained to generate a configuration in which real solvents are separated while fake solvents remain collapsed, after which a final geometric decoding step removes overlapping atoms.

4.4 Other Details

Iterative Hydrogen-bond Closure Cropping. Standard cropping strategies produce orphan tokens (Figure 4d), miss-

Table 1. Joint generation of solvated system improves structure prediction. Generated solvents are excluded from both alignment and evaluation; all metrics are measured on the biomolecular coordinates. Global accuracy is reported as RMSD, while local structural consistency is measured using lDDT over heavy-atom neighborhoods, stratified by interaction type. (gt): BoltzW with ground-truth structure as proposal, ($\epsilon = 1$): BoltzW with fixed entropy regularizer.

		Global (RMSD ↓)			Local (lDDT ↑)					
Dataset	Model	Overall	Backbone	Sidechain	Prot.	DNA	Ligand	Prot.–Prot.	DNA–Prot.	Ligand–Prot.
Easy	Boltz-2	1.514	1.041	1.824	0.943	–	0.947	0.826	–	0.788
	BoltzW($\epsilon = 1$)	1.489	0.975	1.809	0.943	–	0.928	0.852	–	0.904
	BoltzW	1.187	0.648	1.515	0.944	–	0.914	0.890	–	0.902
	BoltzW(gt)	1.118	0.587	1.443	0.943	–	0.914	0.889	–	0.910
Local	Boltz-2	3.480	2.889	3.853	0.919	0.893	0.912	0.617	0.529	0.453
	BoltzW(ϵ=1)	2.493	1.627	2.943	0.930	0.833	0.942	0.742	0.542	0.395
	BoltzW	2.454	1.593	2.903	0.930	0.831	0.943	0.743	0.598	0.407
	BoltzW(gt)	2.271	1.466	2.703	0.930	0.823	0.932	0.758	0.543	0.428
Global	Boltz-2	8.481	8.159	8.722	0.872	0.907	0.730	0.422	0.597	0.430
	BoltzW(ϵ=1)	8.377	7.903	8.706	0.926	0.781	0.802	0.511	0.377	0.535
	BoltzW	8.736	8.278	9.049	0.927	0.776	0.761	0.511	0.332	0.459
	BoltzW(gt)	8.889	8.438	9.195	0.927	0.758	0.736	0.477	0.298	0.509

ing their hydrogen-bond partners. Let $V = V_{\mathrm{x}} \cup V_{\mathrm{s}}$ denote the full token set partitioned into biomolecular and solvent tokens, $E \subseteq V \times V$ the hydrogen-bond adjacency relation, and C_{x}, C_{s} the biomolecular and solvent subsets of a crop C. We introduce two complementary operators: Residue-to-Solvent Expansion (RSE), $\mathcal{E}_{\mathrm{x}\to\mathrm{s}}(C) := C \cup \{s \in V_{\mathrm{s}} : \exists x \in C_{\mathrm{x}},\ (s, x) \in E\}$, and Solvent-to-Residue Expansion (SRE), $\mathcal{E}_{\mathrm{s}\to\mathrm{x}}(C) := C \cup \{x \in V_{\mathrm{x}} : \exists s \in C_{\mathrm{s}},\ (x, s) \in E\}$, which eliminate orphan residues and solvents respectively. Since satisfying both simultaneously is intractable due to conflicting hydrogen-bond closure and sequence continuity constraints, we adopt an iterative scheme that alternates between RSE and SRE, progressing toward closure but terminating once a predefined token budget is reached. Our cropping algorithm best captures the native hydrogen-bond network (Figure 4c,e,f).

Proposal Conditioning. BoltzW conditions on a proposal structure from Boltz-2 (Passaro et al., 2025), provided via the template module (see Figure 1d). Template conditioning primarily constrains the backbone; the model refines the proposal by reorganizing side-chain rotamers and co-adapting solvents, correcting systematic biases in the solvent-free structure prediction model. To promote robustness and support diverse applications, we randomize the conditioning during training: with equal probability of $1/3$ each, the model receives a proposal structure, the ground truth structure, or no conditioning.

Trunk Freezing. We freeze the trunk and train only the structure module, as finetuning the full model exceeds typical academic compute budgets. This choice is also principled: solvent molecules carry no evolutionary information, so the frozen trunk representation requires no solvent-specific adaptation. Training is therefore focused on incorporating solvent effects into coordinate prediction given the frozen evolutionary context. The reduced memory requirements of a frozen trunk allows us to triple the crop size during training, which enables our model to reason over larger spatial contexts.

5 Experiments

We evaluate on two complementary tasks. First, we assess whether joint modeling improves structure-prediction beyond solvent-free modeling. Second, we evaluate the solvent prediction, examining whether improvements in structure prediction arise from physically grounded hydration. Training data consists of PDB structures (CC0 1.0) selected following (Stark et al., 2025). The model is trained on 32 A100 GPUs for 2 days with the AdamW optimizer.

5.1 Biomolecular Structure Prediction

Benchmark. To probe performance under varying proposal quality, we stratify the validation set into three regimes: *Easy*, where proposals are already close to the ground truth (all lDDTs > 0.8, RMSD < 5 Å, 72 structures); *Local* (any lDDT < 0.8, RMSD < 5 Å, 65 structures), where structures are globally aligned but contain local errors; and *Global* (RMSD ≥ 5 Å, 46 structures), where the proposal itself is significantly misaligned. We mainly compare against the proposal predicted structures from Boltz-2 (Passaro et al., 2025).

Evaluation Metrics. All reported metrics exclude solvent atoms from both alignment and evaluation. Global structural accuracy is evaluated using root-mean-square deviation (RMSD) after rigid alignment on biomolecular atoms. We report RMSD over all, backbone, and sidechain atoms. Local structural consistency is assessed using lDDT (Mariani et al., 2013). We compute lDDTs and stratify results by

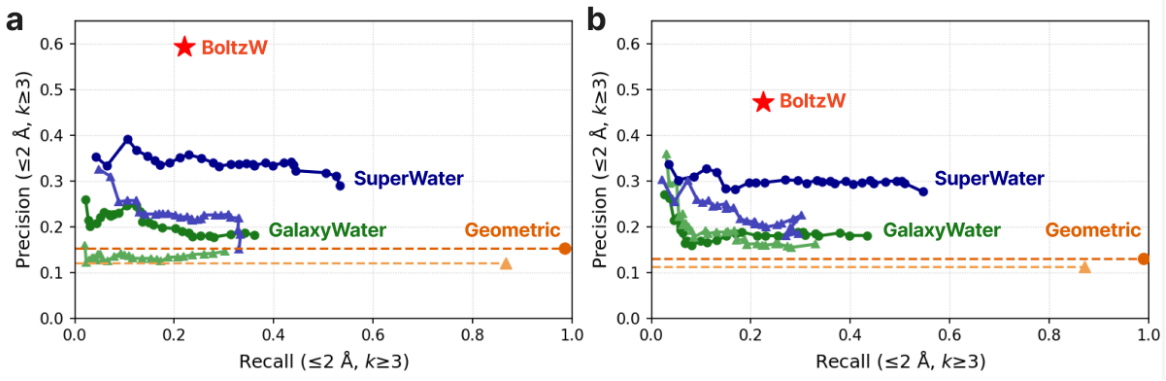

Figure 5. BoltzW achieves state-of-the-art solvent prediction. Precision-recall curves on **a.** the *Easy* task and **b.** the *Local* task, with waters filtered to those forming at least 3 hydrogen bonds in both ground truth and predicted structures. A distance threshold of 2 Å was used to determine occupancy. Points along each curve correspond to different score thresholds used to filter predicted solvents. Boltz-W and the geometric imputation method are shown as single points, as they do not produce a ranked list. Darker circle markers indicate imputation on ground truth structures; lighter triangle markers indicate imputation on Boltz-2 refolded structures.

Table 2. BoltzW analysis. Recall and precisions are measured with 2 Å distance threshold.

	Overall				HB=3		HB$\geq$ 4	
Dataset	Rec	Prec	Chamfer	Avg. HB	Frac(pred./gt.)	Rec	Frac(pred./gt.)	Rec
Easy	0.348	0.427	58.2	2.72	0.315 / 0.454	0.242	0.239 / 0.546	0.664
Local	0.284	0.341	38.7	2.64	0.353 / 0.525	0.375	0.244 / 0.475	0.576

interaction type, including intra-chain lDDTs (Prot., DNA, Ligand), and inter-chain interactions (Prot.–Prot., DNA–Prot., Ligand–Prot.).

Joint Generation Improves Structure Prediction. Across all regimes, BoltzW improves structural accuracy relative to Boltz-2, with gains concentrated in the *Easy* and *Local* settings. This is consistent with the role of solvent as a mediator of local interactions: explicit solvent primarily refines flexible degrees of freedom rather than global fold. In contrast, gains diminish in the *Global* regime, where large backbone errors limit the ability to recover correct hydration patterns. Overall, these results demonstrate that explicit solvent acts as an effective local correction signal for structure prediction.

Improvement Depends on Proposal Accuracy. To evaluate the importance of the proposal structure, we compare with conditioning on the ground-truth structure. As expected, better proposals yield better final structures: accurate templates produce reliable hydration patterns and strong refinement signals. However, ground truth conditioning was not helpful on *Global* dataset, implying that joint modeling and even strong conditioning barely reduces inherent bias in original model.

Regularization Annealing. We compare with simple regularization $\epsilon = 1$. As stated in Figure 3, too small regularization at the high noise level gives high-variance learning signal, harming the overall modeling capacity. Our annealing schedule provides smoother transition in correspondence matrices.

5.2 Solvent Prediction

Benchmark. To evaluate solvent prediction, structure prediction should be reliable to ensure reasonable solvent assignments; therefore we evaluate only on *Easy* and *Local*. We additionally remove structures with too few waters ($M < 5$) after solvent filtering (Section 4.1), resulting in 46 and 42 structures, respectively. We compare against two deep-learning solvent imputation baselines, GalaxyWater(Park & Seok, 2022) and SuperWater (Kuang et al., 2024) (see Section 2.2). We also compared with an algorithmic imputation method which is purely geometric: searching through positions that satisfy HB $\geq$ 3, and resolving clashes (inspired from (Kriegel & Muller, 2023)). Note that these baselines generate solvent positions conditioned on a fixed structure input. We therefore compare with each model imputing on ground-truth and Boltz-2 predicted structures.

Evaluation Metric. Evaluation is nontrivial because baselines used data under different solvent filtering criteria. To make evaluation comparable across methods, we apply the same hydrogen-bond-based filtering procedure to both ground-truth and predicted solvent sets ($k \geq 3$), and compute precision and recall only within the filtered subset. We use distance-based matching, after aligning biomolecular coordinates. We compare precision-recall curve across the models.

Results. As shown in Figure 5, BoltzW achieves the best precision-recall balance on both benchmarks, even without access to ground truth structures. Compared to solvent im-

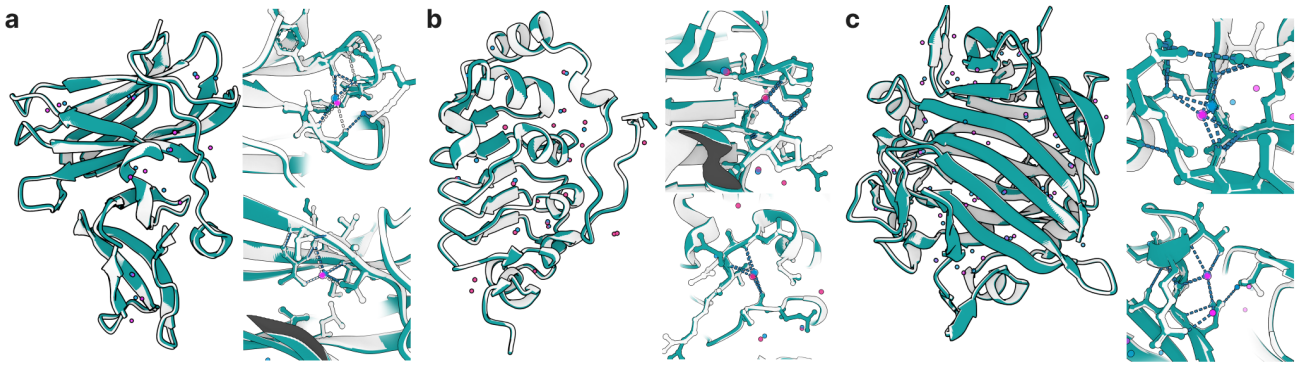

Figure 6. Qualitative results. **a.** 8EBB **b.** 8AXJ **c.** 8POF. Ground-truth and predicted structures are colored white and green. Ground truth and predicted waters are colored blue and magenta.

putation methods on Boltz-2 predicted structures, BoltzW achieves roughly triple the precision at a similar level of recall. The geometric algorithm's high recall but low precision suggests that identifying candidate water positions is relatively easy, but most of them do not correspond to true hydration sites. BoltzW is conservative in this regard — it sacrifices some recall for substantially better precision. While baseline methods perform consistently across both datasets, BoltzW shows lower performance on the *Local* dataset, reflecting the tighter coupling between solvent imputation and structure prediction in our model. These results suggest that the gains observed in structure prediction are not incidental, but arise from a more reliable internal model of solvent organization.

Analysis. Table 2 shows additional analysis. BoltzW recovers nearly 30% of structured waters overall, with recall rising to nearly 60% for highly structured waters (HB ≥ 4). This aligns with our intuition: fully surrounded cavities are likely to be occupied, and BoltzW successfully identifies most of these positions. However, the average hydrogen bond count below 3.0 indicates that roughly 40% of generated waters are dangling. This is partly a fine-structure generation challenge: many packages define hydrogen bonds with distance cutoffs of [2.5,3.5] Å, so a slight excess beyond 3.5 Å disqualifies an otherwise valid bond. Since BoltzW co-generates waters alongside their hydrogen-bond partners, jointly controlling all atomic positions is a combinatorial problem and remains as future work.

6 Conclusion

We introduced BoltzW, a generative model that jointly predicts solvated biomolecular structure. We first showed that existing structure prediction models struggle to recover even coarse solvent information such as cardinality, motivating the joint modeling approach taken by BoltzW (Section 3). We made explicit solvent generation tractable through three core technical contributions: (i) structured solvent selection via hydrogen-bond coordination and B-factor filtering, (ii) permutation-invariant training objectives derived from entropic optimal transport, and (iii) fake solvent-based cardinality prediction (Section 4). Empirically, BoltzW improves structure prediction, with explicit solvent modeling particularly beneficial for refining local interactions. BoltzW also outperforms solvent imputation baselines without access to ground-truth structures, especially in precision (Section 5).

Acknowledgements

We thank to Abhi Gupta, Hannes Stark and Julia Balla for their kind help on discussions and refining the paper. We also thank to Felix Faltings, Mateo Reveiz, and Peter Mikhael for their careful feedback.

Impact Statement

We hope BoltzW serves as an early step toward structure prediction models that treat the chemical environment as a first-class component of structure determination. We envision models that reason jointly over sequence, structure, and physical context (solvation, pH, membrane composition, and cofactor availability) ultimately accelerating drug discovery, enabling the rational design of enzymes and biologics.

References

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630 (8016):493–500, 2024.

Breiten, B., Lockett, M. R., Sherman, W., Fujita, S., Al-Sayah, M., Lange, H., Bowers, C. M., Heroux, A., Krilov, G., and Whitesides, G. M. Water networks contribute to enthalpy/entropy compensation in protein–ligand binding. *Journal of the American Chemical Society*, 135(41): 15579–15584, 2013.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Feig, M. and Brooks III, C. L. Recent advances in the development and application of implicit solvent models in biomolecule simulations. *Current opinion in structural biology*, 14(2):217–224, 2004.

Geffner, T., Didi, K., Zhang, Z., Reidenbach, D., Cao, Z., Yim, J., Geiger, M., Dallago, C., Kucukbenli, E., Vahdat, A., and Kreis, K. Proteina: Scaling flow-based protein structure generative models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=TVQLu34bdw.

Genevay, A., Peyré, G., and Cuturi, M. Learning generative models with sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pp. 1608–1617. PMLR, 2018.

Hinz, F. B., Masters, M. R., Nguyen, J. T., Mahmoud, A. H., and Lill, M. A. Accelerated hydration site localization and thermodynamic profiling. *Journal of Chemical Information and Modeling*, 65(6):2794–2805, 2025.

Hu, B. and Lill, M. Watsite: hydration site prediction program with pymol interface. *Journal of Computational Chemistry*, 35(16):1255–1260, 2014.

Huggins, D. J. Quantifying the entropy of binding for water molecules in protein cavities by computing correlations. *Biophysical journal*, 108(4):928–936, 2015.

Hui, K.-H., Liu, C., Zeng, X., Fu, C.-W., and Vahdat, A. Not-so-optimal transport flows for 3d point cloud generation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=62Ff8LDAJZ.

Jing, B., Berger, B., and Jaakkola, T. Alphafold meets flow matching for generating protein ensembles. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=rs8Sh2UASt.

Klyshko, E., Kim, J. S.-H., and Rauscher, S. Laws: Local alignment for water sites—tracking ordered water in simulations. *Biophysical Journal*, 122(14):2871–2883, 2023.

Kriegel, M. and Muller, Y. A. De novo prediction of explicit water molecule positions by a novel algorithm within the protein design software mumbo. *Scientific Reports*, 13(1): 16680, 2023.

Kuang, X., Su, Z., Liu, Y., Lin, X., Spencer-Smith, J., Derr, T., Wu, Y., and Meiler, J. Superwater: Predicting water molecule positions on protein structures by generative ai. *bioRxiv*, pp. 2024–11, 2024.

Laage, D., Elsaesser, T., and Hynes, J. T. Water dynamics in the hydration shells of biomolecules. *Chemical reviews*, 117(16):10694–10725, 2017.

Liebschner, D., Dauter, M., Brzuszkiewicz, A., and Dauter, Z. On the reproducibility of protein crystal structures: five atomic resolution structures of trypsin. *Biological Crystallography*, 69(8):1447–1462, 2013.

Liebschner, D., Afonine, P. V., Baker, M. L., Bunkóczi, G., Chen, V. B., Croll, T. I., Hintze, B., Hung, L.-W., Jain, S., McCoy, A. J., et al. Macromolecular structure determination using x-rays, neutrons and electrons: recent developments in phenix. *Biological Crystallography*, 75 (10):861–877, 2019.

Mariani, V., Biasini, M., Barbato, A., and Schwede, T. lddt: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, 29(21):2722–2728, 2013.

Mobley, D. L. and Dill, K. A. Binding of small-molecule ligands to proteins:"what you see" is not always "what you get". *Structure*, 17(4):489–498, 2009.

Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F., and Vagin, A. A. Refmac5 for the refinement of macromolecular crystal structures. *Biological crystallography*, 67(4): 355–367, 2011.

Nakasako, M. Water–protein interactions from high–resolution protein crystallography. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 359(1448):1191–1206, 2004.

Park, S. and Seok, C. 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.

Passaro, S., Corso, G., Wohlwend, J., Reveiz, M., Thaler, S., Somnath, V. R., Getz, N., Portnoi, T., Roy, J., Stark, H., et al. Boltz-2: Towards accurate and efficient binding affinity prediction. *BioRxiv*, 2025.

Ross, G. A., Morris, G. M., and Biggin, P. C. Rapid and accurate prediction and scoring of water molecules in protein binding sites. *PloS one*, 7(3):e32036, 2012.

Ross, G. A., Bodnarchuk, M. S., and Essex, J. W. Water sites, networks, and free energies with grand canonical monte carlo. *Journal of the American Chemical Society*, 137(47): 14930–14943, 2015.

Samways, M. L., Taylor, R. D., Bruce Macdonald, H. E., and Essex, J. W. Water molecules at protein–drug interfaces: computational prediction and analysis methods. *Chemical Society Reviews*, 50(16):9104–9120, 2021.

Sindhikara, D. J. and Hirata, F. Analysis of biomolecular solvation sites by 3d-rism theory. *The Journal of Physical Chemistry B*, 117(22):6718–6723, 2013.

Sridhar, A., Ross, G. A., and Biggin, P. C. Waterdock 2.0: Water placement prediction for holo-structures with a pymol plugin. *PloS one*, 12(2):e0172743, 2017.

Stark, H., Faltings, F., Choi, M., Xie, Y., Hur, E., O'Donnell, T. J., Bushuiev, A., Uçar, T., Passaro, S., Mao, W., et al. Boltzgen: Toward universal binder design. *bioRxiv*, pp. 2025–11, 2025.

Wang, L., Berne, B. J., and Friesner, R. A. Ligand binding to protein-binding pockets with wet and dry regions. *Proceedings of the National Academy of Sciences*, 108(4): 1326–1330, 2011.

Wlodawer, A., Dauter, Z., Rubach, P., Minor, W., Loch, J. I., Brzezinski, D., Gilski, M., and Jaskolski, M. Waterless structures in the protein data bank. *IUCrJ*, 11(6):966–976, 2024.

Woo, H.-J., Dinner, A. R., and Roux, B. Grand canonical monte carlo simulations of water in protein environments. *The Journal of chemical physics*, 121(13):6392–6400, 2004.

Yang, Y., Hu, B., and Lill, M. A. Watsite2. 0 with pymol plugin: Hydration site prediction and visualization. In *Protein Function Prediction: Methods and Protocols*, pp. 123–134. Springer, 2017.

Zamanos, A., Ioannakis, G., and Emiris, I. Z. 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.

Zsidó, B. Z. and Hetényi, C. Water in drug design: pitfalls and good practices. *Expert Opinion on Drug Discovery*, 20(6):745–764, 2025.

A Cardinality Predictions.

Setup. We study a controlled setting in which the ground-truth biomolecular structure X is given, and evaluate how well different aspects of solvent organization can be predicted from the Z that are trained to generate X alone. For each example, all solvent molecules are removed from the input, while the ground-truth biomolecular coordinates and features are retained. Let $\mathcal{R}_s \subseteq \{1, \ldots, N\} \subset X$ denote the set of tokens forming hydrogen bonds with solvent s. We supervise token-level hydration signals under three tasks of increasing complexity.

- **Binary cardinality.** Each token i is labeled according to whether at least one water forms a hydrogen bond with it:

$$y_i^{\text{bin}} = \mathbf{1}\{\exists s \in \mathcal{S} : i \in \mathcal{R}_s\}. \tag{7}$$

  This task captures the *support* of the hydration distribution, i.e., whether a residue is hydrated at all. It collapses all multiplicity and sharing information and can be interpreted as estimating a marginal event probability.

- **Discrete (degenerate) cardinality.** Each token predicts the number of interacting solvents, from zero to six interactions:

$$y_i^{\text{deg}} = |\{s \in S : i \in \mathcal{R}_s\}|. \tag{8}$$

  Compared to the binary task, this requires resolving multiplicity of interactions but still treats each token independently and does not encode how waters are shared across neighboring residues.

- **Fractional cardinality.** Assigns each water fractionally across all residues it interacts with. For each water s, we define a normalized contribution

$$\alpha_{s,i} = \begin{cases} \frac{1}{|\mathcal{R}_s|} & \text{if } i \in \mathcal{R}_s, \\ 0 & \text{otherwise,} \end{cases} \tag{9}$$

  and define the target as

$$y_i^{\text{frac}} = \sum_{s \in \mathcal{S}} \alpha_{s,i}. \tag{10}$$

  This construction yields a mass-conserving soft assignment satisfying $\sum_i y_i^{\text{frac}} = |\mathcal{S}|$. Unlike discrete counts, fractional cardinality encodes *shared solvent structure*: a single water simultaneously contributes to multiple residues, inducing coupling across tokens. This induces global dependencies across residues, requiring the model to capture structured interactions beyond local features.

Hierarchy of difficulty. These tasks form a natural hierarchy of increasing difficulty:

$$y_i^{\text{bin}} \prec y_i^{\text{deg}} \prec y_i^{\text{frac}}. \tag{11}$$

Binary prediction requires only detecting the *support* of the hydration distribution. Discrete cardinality additionally requires estimating *multiplicity* but remains a local marginal prediction. In contrast, fractional cardinality depends on the *incidence structure* of the bipartite water–residue interaction graph $\{(s, i) : i \in \mathcal{R}_s\}$, as each water contributes jointly to multiple residues. Consequently, fractional targets are globally coupled across tokens and cannot be decomposed into independent per-token predictions. This introduces substantially higher information requirements and makes the task strictly more challenging.

Table 3. Solvent predictability from structural representations under three tasks of increasing complexity. $R^2 < 0$ indicates performance below a mean predictor baseline.

Task	Metric	Baseline	Filtered
Binary hydration	AUPRC	0.25	0.85
Discrete cardinality	R^2	0.0	-0.12
Fractional cardinality	R^2	0.0	-0.05

Training and evaluation. All tasks share the same encoder and structural input but use separate prediction heads. Our architecture and training procedure follow the confidence modeling setup of Boltz-2, with all other modules frozen. The binary hydration task is trained as a binary classification and is evaluated primarily using AUPRC. The discrete and fractional cardinality tasks are trained to predict scalar-valued targets and are evaluated primarily using correlation-based metrics.

For validation, we use the same set of targets as in BoltzGen (Stark et al., 2025), but restrict evaluation to samples whose predicted biomolecular structures achieve less than 2 Å RMSD to the ground truth. This ensures that solvent prediction is assessed only on geometrically reliable biomolecular conformations.

B Geometric Solvent Imputation

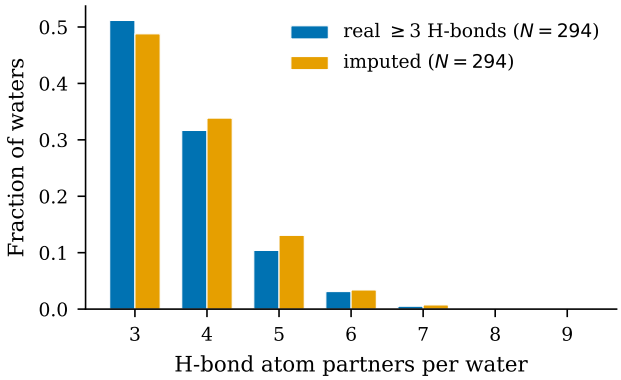

Figure 7. Algorithmic imputation recovers hydrogen-bond statistics of real waters.

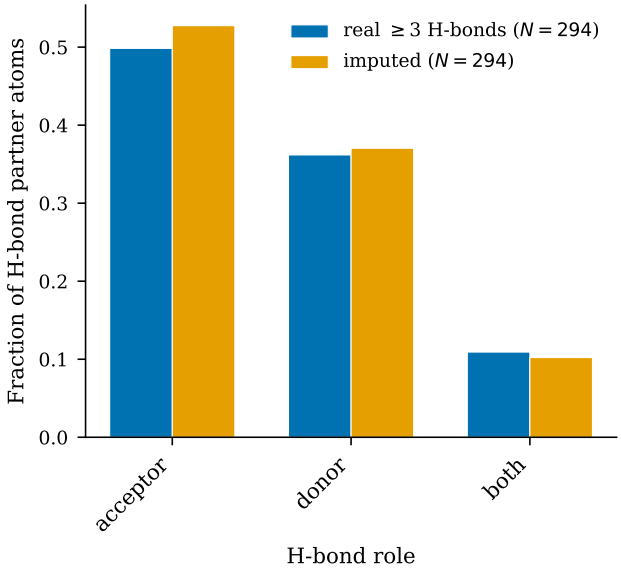

Figure 8. Algorithmic imputation recovers hydrogen-bond role distributions.

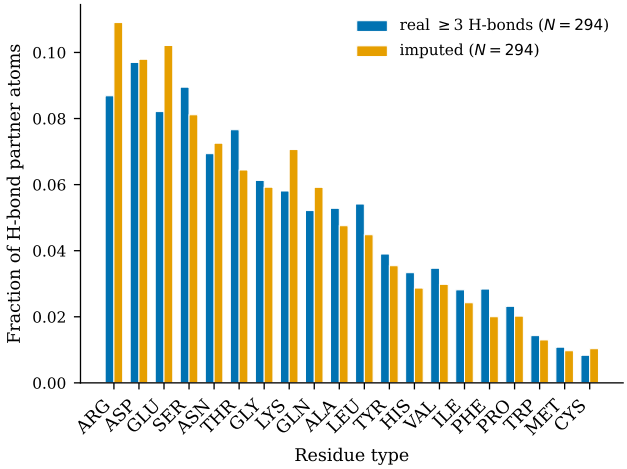

Figure 9. Algorithmic imputation recovers hydrogen-bond residue distribution.

C Thermodynamic Decompositions and Structured Water Criterion

Bulk-structured decomposition. We decompose the solvent into a bulk background and a structured foreground. Bulk solvent corresponds to rapidly exchanging, weakly localized waters whose contributions can be captured through an implicit potential of mean force. In contrast, structured waters occupy well-defined positions within hydrogen-bond networks, stabilize charged intermediates and metal ions, and mediate protein–ligand and proton-transfer interactions. These waters are tightly coupled to the local biomolecular geometry and must therefore be modeled explicitly.

Formally, for a fixed biomolecular transition from reference state X^R to transition state $X^\ddagger$, the activation free energy marginalizes over all solvent configurations:

$$\Delta G^\ddagger = -RT \log \frac{\int e^{-\beta E(X^\ddagger, S)}\, dS}{\int e^{-\beta E(X^R, S)}\, dS}. \tag{12}$$

Decomposing $S = (S_{\text{bulk}}, S_{\text{str.}})$ and noting that bulk solvent equilibrates rapidly and depends weakly on fine-grained biomolecular geometry, its contributions can be approximated as an averaged potential of mean force, yielding:

$$\Delta G^\ddagger \approx \Delta G^\ddagger_{\text{bulk}} + \Delta G^\ddagger_{\text{str.}}. \tag{13}$$

This decomposition establishes that only $S_{\text{str.}}$ requires explicit modeling. The remaining question is how to operationalize this decomposition in a way that is geometry-local, computationally tractable, and compatible with a generative modeling pipeline.

Derivation. This decomposition should be interpreted as an effective modeling approximation rather than an exact thermodynamic identity. Starting from the full solvent-marginalized activation free energy,

$$\Delta G^\ddagger = -RT \log \frac{\int e^{-\beta E(X^\ddagger, S_{\text{bulk}}, S_{\text{str}})}\, dS_{\text{bulk}}\, dS_{\text{str}}}{\int e^{-\beta E(X^R, S_{\text{bulk}}, S_{\text{str}})}\, dS_{\text{bulk}}\, dS_{\text{str}}},$$

we first integrate out the bulk solvent to define an effective free energy

$$F_{\text{eff}}(X, S_{\text{str}}) := -RT \log \int e^{-\beta E(X, S_{\text{bulk}}, S_{\text{str}})}\, dS_{\text{bulk}}.$$

We then approximate this effective energy as

$$F_{\text{eff}}(X, S_{\text{str}}) \approx G_{\text{bulk}}(X) + E_{\text{str}}(X, S_{\text{str}}),$$

which separates the contribution of rapidly equilibrating bulk solvent from the residual, geometry-sensitive contribution of structured waters. Under this approximation, the activation free energy decomposes as

$$\Delta G^\ddagger \approx \Delta G^\ddagger_{\text{bulk}} + \Delta G^\ddagger_{\text{str}}.$$

D Permutation Invariances

Intuition. Consider a simple example with three solvent molecules. Let the ground-truth configuration be $S^0 = \{A_0, B_0, C_0\}$, and suppose the model predicts $\hat{S} = \{A_0, C_0, B_0\}$, which is identical to S^0 up to permutation. Under index-wise supervision, this prediction incurs a large loss because $\hat{B}$ is compared to B_0 and $\hat{C}$ to C_0, despite the configuration being correct as a set. In contrast, a permutation-invariant objective identifies the optimal matching between predicted and ground-truth points and assigns zero loss to this configuration. This illustrates that index-based losses measure discrepancies in representation rather than in the underlying geometry. By resolving permutation ambiguity, our objective removes spurious error arising from arbitrary index assignments, yielding a consistent and low-variance training signal that allows the model to focus on learning the spatial organization of solvent.

Comparison with OT matching. Our formulation differs from approaches that apply optimal transport directly between intermediate noisy states S_t and the ground-truth configuration S^0, as done in transport-based or flow matching methods (Hui et al., 2025). In such approaches, the transport plan defines a displacement field that explicitly moves mass from S_t toward S^0, effectively learning a transport map or flow in Wasserstein space. In contrast, we apply optimal transport only between the predicted clean configuration $\hat{S}$ and S^0, using it solely to resolve permutation ambiguity in the supervision signal. As a result, our method does not learn an explicit optimal transport flow or enforce minimal transport trajectories during generation. Instead, it learns a mapping from noisy inputs to geometrically consistent solvent configurations that are correct up to permutation. This separation allows us to preserve the ground-truth configuration as the supervision target while benefiting from permutation-invariant alignment, without introducing the additional complexity or bias associated with learning transport dynamics.