INFERENCE-TIME OPTIMIZATION FOR EXPERIMENT-GROUNDED PROTEIN ENSEMBLE GENERATION

Advaith Maddipatla[1], **Anar Rzayev**[1], **Marco Pegoraro**[1], **Martin Pacesa**[2], **Paul Schanda**[1], **Ailie Marx**[3,4], **Sanketh Vedula**[1,5,6], **Alex Bronstein**[1,7]

[1]ISTA [2]University of Zurich [3]Tel Hai, Israel [4]Migal, Israel [5]Princeton University [6]Broad Institute of MIT and Harvard [7]Technion

ABSTRACT

Protein function relies on dynamic conformational ensembles, yet models like AlphaFold3 (AF3) often fail to produce ensembles that match experimental data. Recent experiment-guided approaches address this by steering the reverse diffusion process, but remain limited by fixed sampling horizons and sensitivity to initialization, often yielding thermodynamically implausible results. We introduce a general inference-time optimization framework that instead optimizes latent representations to maximize ensemble likelihood, removing dependence on diffusion length, reducing initialization bias, and enabling flexible incorporation of external constraints. Further, we present sampling schemes for generating Boltzmann-weighted ensembles by combining AF3 structural priors with force-field energies. Across X-ray and NMR benchmarks, our method improves diversity, physical energy, and agreement with experimental data, often surpassing deposited PDB structures. Finally, ipTM maximization shows that perturbing AF3 embeddings can artificially inflate model confidence, exposing a vulnerability in current design metrics and suggesting ways to reduce false discoveries in binder engineering.

1 INTRODUCTION

Proteins are dynamic systems that populate multiple conformational states and interact with diverse partners. Biological function is therefore often governed by conformational ensembles rather than a single structure (Furnham et al., 2006). Proteins further act through transient complexes with other proteins, making conformational heterogeneity central to molecular recognition. Capturing this heterogeneity is critical to design tasks that require generating and ranking plausible conformations.

Modern sequence-conditioned predictors (Jumper et al., 2021) provide strong structural priors. Diffusion-based models such as AlphaFold3 (AF3) (Abramson et al., 2024) can generate high-quality structures, but unguided sampling often under-represents rare functional states and may fail to fully match experimental observations, particularly for flexible loops (Alderson et al., 2023). This motivates methods that incorporate experimental measurements and design objectives during inference.

Recent work has explored guiding sequence-conditioned diffusion models with differentiable objectives derived from forward models, including NOE-based distance likelihoods for NMR, real-space density agreement for X-ray crystallography, and electrostatic potential agreement for cryo-EM (Maddipatla et al., 2025a; Raghu et al., 2025; Dhariwal & Nichol, 2021).

In parallel, protein design workflows rely on objectives that are not physical measurements. In binder and protein–protein interface design, BindCraft-like pipelines (Pacesa et al., 2025) backpropagate through structure predictors using metrics such as interface predicted Template Modeling (ipTM) and predicted Template Modeling (pTM) to guide optimization and rank candidate complexes.

Currently, these objectives are optimized via coordinate-space guidance during the denoising process. While effective, this tightly couples results to the denoising trajectory, making them sensitive to initialization, scheduling, and limited step budgets. Moreover, such guidance enforces agreement at the level of ensembles but does not specify how to form thermodynamically meaningful ensembles. Our framework addresses this by decoupling conditioning from the denoising schedule and incorporating energy-based reweighting to obtain thermodynamically consistent ensembles when desired.

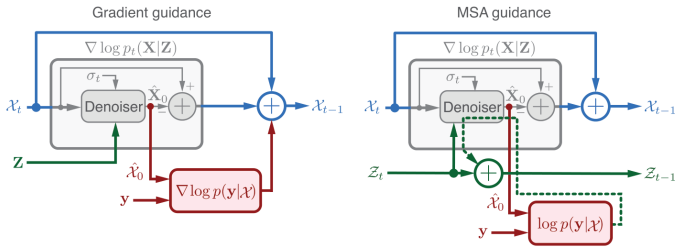

Figure 1: **Gradient guidance versus inference-time optimization in AlphaFold3.** (Left) In gradient guidance, the embedding $\mathbf{Z}$ is fixed, and experimental gradients $\nabla_{\boldsymbol{\mathcal{X}}} \log p(\mathbf{y}|\boldsymbol{\mathcal{X}})$ are applied to the coordinates $\boldsymbol{\mathcal{X}}$ during reverse diffusion. (Right) In inference-time optimization, a batch of embeddings $\boldsymbol{\mathcal{Z}}$ is updated using the experimental likelihood $\log p(\mathbf{y}|\boldsymbol{\mathcal{X}})$, while structures are denoised via reverse diffusion conditioned on the optimized embeddings. The dotted line shows the gradient flow.

2 Contributions and Main Results

2.1 Contributions

Inference-time optimization. We introduce a novel inference-time optimization (IT-Optimization) framework for sequence-conditioned diffusion models that treats AF3 as a learned structural prior. We condition sampling by updating Pairformer trunk embeddings using gradients of any differentiable objective, including experiment-derived likelihoods, confidence metrics, or their combinations. Unlike coordinate-space guidance along a fixed denoising trajectory, representation-space updates persist across sampling runs and bias subsequent trajectories toward target-satisfying structures. The method is agnostic to the diffusion schedule and composes with any modified sampling scheme, functioning as a meta-guidance layer that shapes the conditioning landscape before coordinate-level guidance. Empirically, representation-space conditioning consistently outperforms coordinate-space guidance and serves as an effective upstream step for downstream coordinate-level guidance (see Figs. 1, 2).

Energy-weighted sampling. We combine the AF3 prior with an external force-field prior via energy-based reweighting, enabling Boltzmann-weighted ensemble statistics rather than uniform weighting of accepted samples. This preserves agreement with the conditioning signal while biasing populations toward thermodynamically plausible regions, yielding ensembles with improved energetic profiles.

2.2 Main results

Algorithm 1 IT-Optimization

Input: Initial trunk embeddings $\boldsymbol{\mathcal{Z}}$; experimental observation $\mathbf{y}$; noise schedule $[\sigma_T, \ldots, \sigma_0]$; learning rate η_z; ensemble size n; outer iterations K; prior weight λ_p
Output: Optimized trunk embeddings $\boldsymbol{\mathcal{Z}}$
for $k = 1$ to K ▷ Outer loop
  $\boldsymbol{\mathcal{X}} \sim \sigma_T \cdot [\mathbf{N}^1, \ldots, \mathbf{N}^n]^T \quad \mathbf{N}^i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  for $t \in [T-1, \ldots, 0]$ ▷ Inner loop
    $\boldsymbol{\mathcal{Z}} \leftarrow \boldsymbol{\mathcal{Z}} + \nabla_{\boldsymbol{\mathcal{Z}}}(\eta_z \log p(\mathbf{y}|\hat{\boldsymbol{\mathcal{X}}}_0) + \lambda_p \log p(\boldsymbol{\mathcal{Z}}|\mathbf{a}))$
    $\boldsymbol{\mathcal{X}} \leftarrow \mathrm{ReverseSDE}(\boldsymbol{\mathcal{X}}; \boldsymbol{\mathcal{Z}}, t, \sigma_t)$ ▷ Eq 3
  end for
end for
return $\boldsymbol{\mathcal{Z}}$

NMR. We evaluate our approach on solution-state NMR structure determination, where protein conformational ensembles are inferred from nuclear Overhauser effect (NOE) measurements (Vögeli, 2014), which provide inter-proton distance restraints via cross-peaks. Using proteins from the NMRDB dataset (Klukowski et al., 2024), IT-optimization substantially reduces NOE restraint violations relative to NOE-guided AF3 (see Tabs. 3- 4 & Figs. 4-5, 8). To obtain thermodynamically plausible conformational ensembles, we apply energy-reweighted sampling (and its IT-Optimization analogue). This further reduces restraint violations while simultaneously producing ensembles with lower effective energies under the Amber99 (Wang et al., 2004) force field (see Tabs. 5-6 & Figs 7).

Crystallography. We further evaluate our method on X-ray crystallographic benchmarks comprising sequence segments that either exhibit alternative conformations (altlocs) or correspond to unrestrained short regions adopting context-dependent structures (protein-bound peptides). We observe that the proposed IT-Optimized AF3 framework consistently outperforms experiment-guided AF3 (Maddipatla et al., 2025a) across all evaluated

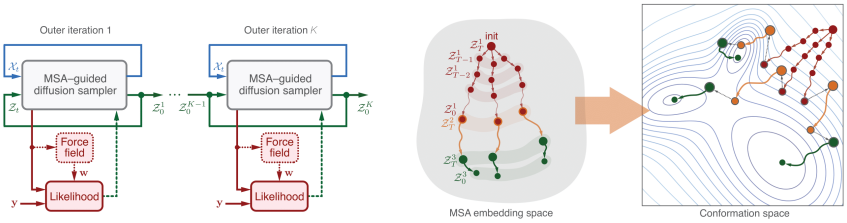

Figure 2: **Nested inference-time optimization of AlphaFold3 embeddings.** (Left) The outer loop runs K diffusion trajectories, each initialized from noise while carrying forward optimized embeddings $\mathcal{Z}$ from the previous diffusion process. Within each trajectory, experimental likelihood gradients update the embeddings, which condition subsequent denoising steps; optional force-field-based Boltzmann weights bias ensemble statistics toward thermodynamically plausible conformations. (Right) Successive trajectories refine embeddings in MSA space, inducing diverse ensembles that match experiments.

X-ray benchmarks, achieving lower R_{work} and R_{free} values, improved local density alignment (see Tabs. 7-9), and greater reproducibility across random seeds (Tab. 10). Notably, while current approaches require terminal restraints to model peptides and struggle to recover widely separated altlocs, IT-optimization enables restraint-free modeling and improves recovery of separated altlocs (Fig.3, 8).

ipTM guidance. We examine IT-Optimization of the interfacial predicted TM-score (ipTM) to evaluate its use as an objective for protein-protein complex prediction and to characterize the sensitivity of ipTM-directed perturbations in AF3 embedding space. Our benchmark spans: (i) complexes with low baseline ipTM (Wee & Wei, 2024), (ii) p53 N-terminal domain complexes with transiently ordered regions and structurally characterized partners (Oren, 1999), and (iii) de novo BindCraft protein-binder complexes. Here, we observe that ipTM-based IT-optimization can improve predictions where unguided AF3 performs poorly (See Fig. 6, 11), but the effect is not consistent across systems. Notably, ipTM can be driven to high-confidence regimes with very small embedding perturbations in the embedding space (0.01%) (See Fig 8). While this sometimes improves structural agreement, in other cases it yields high-confidence predictions with inaccurate interfaces (See Fig. 9, 10). These results suggest that ipTM and related confidence metrics should be used cautiously as optimization objectives, as confidence can increase without corresponding improvements in structural accuracy.

3 Conditional Ensemble Generation

Notation. We denote the amino acid sequence of a protein by $\mathbf{a}$ and its folded 3D structure by $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_m)$, where $\mathbf{x}_i \in \mathbb{R}^3$ denotes the Cartesian coordinate of the i-th atom and m is the total number of atoms in the structure. Furthermore, $\mathbf{Z} = (\mathbf{s}, \mathbf{z})$ is the conditioning variable produced by AF3's Pairformer module.[1] Here, $\mathbf{s}$ and $\mathbf{z}$ denote the single and pairwise representations, which together form learned continuous representations that integrate multiple sequence alignment (MSA) co-evolutionary information with Pairformer-learned pairwise positional and contextual features.

Problem statement. Given a protein sequence $\mathbf{a}$ and an experimental observation $\mathbf{y}$, our goal is to optimize a set of n conditioning variables $\mathcal{Z} = \{\mathbf{Z}^1, \ldots, \mathbf{Z}^n\}$ that induce an ensemble of corresponding structures $\mathcal{X} = \{\mathbf{X}^1, \ldots, \mathbf{X}^n\}$. In AF3, the conditioning variable is a deterministic transformation $\mathbf{Z}(\mathbf{a})$ of the sequence $\mathbf{a}$ produced by the Pairformer. We relax this relation into a separable prior $p(\mathcal{Z}|\mathbf{a}) = p(\mathbf{Z}^1|\mathbf{a}) \cdots p(\mathbf{Z}^n|\mathbf{a})$ preventing $\mathbf{Z}$ from drifting arbitrarily in the embedding space and penalizing off-manifold solutions. In what follows, we assume a simple Gaussian prior $p(\mathbf{Z}|\mathbf{a}) = \mathcal{N}(\mathbf{Z}(\mathbf{a}), \sigma^2)$. Invoking Bayes' theorem, we obtain the posterior

$$p(\mathcal{Z}|\mathbf{y}, \mathbf{a}) \propto p(\mathcal{Z}|\mathbf{a}) p(\mathbf{y}|\mathcal{Z}, \mathbf{a}) = p(\mathcal{Z}|\mathbf{a}) p(\mathbf{y}|\mathcal{Z}),$$

Marginalizing out $\mathcal{X}$ yields,

$$p(\mathbf{y}|\mathcal{Z}) = \int p(\mathbf{y}|\mathcal{X}) p(\mathcal{X}|\mathcal{Z}) d\mathcal{X} = \mathbb{E}_{\mathcal{X} \sim p(\mathcal{X}|\mathcal{Z})} p(\mathbf{y}|\mathcal{X}),$$

[1] "MSA embeddings" and "AF3 embeddings" are used interchangeably to denote outputs of the Pairformer.

where $p(\mathbf{y}|\boldsymbol{\mathcal{X}})$ is a generally inseparable likelihood that quantifies the agreement between the structural ensemble $\boldsymbol{\mathcal{X}}$ and the experimental observation $\mathbf{y}$. This data term is specified by forward models reflecting the physics of the experimental measurement. This leads to the exact MAP problem

$$\boldsymbol{\mathcal{Z}}^* = \arg\max_{\boldsymbol{\mathcal{Z}}} \log \mathbb{E}_{\boldsymbol{\mathcal{X}}\sim p(\boldsymbol{\mathcal{X}}|\boldsymbol{\mathcal{Z}})} p(\mathbf{y}|\boldsymbol{\mathcal{X}}) p(\boldsymbol{\mathcal{Z}}|\mathbf{a}). \tag{1}$$

To avoid the hardly-tractable log of the expectation, we invoke Jensen's inequality $\log \mathbb{E}p \geq \mathbb{E}\log p$ to obtain an evidence lower bound (ELBO)-type of an objective

$$\boldsymbol{\mathcal{Z}}^* = \arg\max_{\boldsymbol{\mathcal{Z}}} \mathbb{E}_{\boldsymbol{\mathcal{X}}\sim p(\boldsymbol{\mathcal{X}}|\boldsymbol{\mathcal{Z}})} \log p(\mathbf{y}|\boldsymbol{\mathcal{X}}) + \log p(\boldsymbol{\mathcal{Z}}|\mathbf{a}). \tag{2}$$

The maximization of this lower bound can be viewed as a form of variational inference with $\boldsymbol{\mathcal{Z}}$ acting as a variational parameter and the AF3 prior itself acting as a restricted family of priors, $q_{\boldsymbol{\mathcal{Z}}}(\boldsymbol{\mathcal{X}}) = p(\boldsymbol{\mathcal{X}}|\boldsymbol{\mathcal{Z}})$. The expectation is approximated by sampling from AF3.

4 Inference-time Optimization

AF3 (Abramson et al., 2024) generates protein structure ensembles using a diffusion-based generative model (Ho et al., 2020; Karras et al., 2022) defined over the coordinates of $\mathbf{X}$. The reverse diffusion process follows the stochastic differential equation (SDE):

$$d\mathbf{X} = -\left(\frac{1}{2}\mathbf{X} + \nabla_{\mathbf{X}} \log p_t(\mathbf{X}|\mathbf{Z})\right)\sigma_t dt + \sqrt{\sigma_t}\mathbf{N} \tag{3}$$

where σ_t denotes the diffusion noise schedule, $\mathbf{N} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is sampled from a standard normal distribution, and $\nabla_{\mathbf{X}} \log p_t(\mathbf{X}|\mathbf{Z})$ is the learned score function. Protein structures are generated by numerically integrating this SDE from an initial noise sample drawn from $\mathcal{N}(\mathbf{0}, \sigma_T\mathbf{I})$ at time $t = T$ to $t = 0$. In a typical guidance setting (Maddipatla et al., 2025b; 2024; 2025a), a non-i.i.d. ensemble $\boldsymbol{\mathcal{X}} = \{\mathbf{X}^i\}_{i=1}^n$ is generated by integrating an experimental likelihood term into the reverse SDE.

$$d\begin{bmatrix}\mathbf{X}^1\\ \vdots \\ \mathbf{X}^n\end{bmatrix} = -\left(\frac{1}{2}\begin{bmatrix}\mathbf{X}^1\\ \vdots \\ \mathbf{X}^n\end{bmatrix} + \begin{bmatrix}\nabla_{\mathbf{X}^1}\log p_t(\mathbf{X}^1|\mathbf{Z}^1)\\ \vdots \\ \nabla_{\mathbf{X}^n}\log p_t(\mathbf{X}^n|\mathbf{Z}^n)\end{bmatrix} + \eta\nabla_{\boldsymbol{\mathcal{X}}}\log p(\mathbf{y}|\hat{\boldsymbol{\mathcal{X}}}_0)\right)\sigma_t dt + \sqrt{\sigma_t}\begin{bmatrix}\mathbf{N}^1\\ \vdots \\ \mathbf{N}^n\end{bmatrix} \tag{4}$$

Here, $\nabla_{\boldsymbol{\mathcal{X}}} \log p(\mathbf{y}|\hat{\boldsymbol{\mathcal{X}}}_0)$ biases the SDE toward ensembles that maximize agreement with $\mathbf{y}$, $\hat{\boldsymbol{\mathcal{X}}}_0 = D_\theta(\boldsymbol{\mathcal{X}}; \boldsymbol{\mathcal{Z}}, t)$ denotes AF3's denoised structure prediction at diffusion step t, η controls the guidance strength, and $\mathbf{N}^i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ are independent innovation processes. Despite its flexibility, the reverse diffusion process is limited: it operates with a fixed number of steps, is sensitive to the initial noise realization, and its convergence rate is constrained by the noise schedule σ_t.

As a remedy, we propose a nested framework consisting of an outer loop and an inner loop to solve the optimization problem in Eq. 2 on a batch of embeddings $\boldsymbol{\mathcal{Z}} = \{\mathbf{Z}^i\}_{i=1}^n$ at inference-time.

Outer loop (Exploration). The outer loop performs a global optimization over $\boldsymbol{\mathcal{Z}}$. Due to the iterative nature of the diffusion process, we treat each outer iteration $k \in \{1, \ldots, K\}$, as a fresh sampling of the diffusion noise $\boldsymbol{\mathcal{X}}_T^k = \{\mathbf{X}_T^{k,i} \sim \mathcal{N}(\mathbf{0}, \sigma_T\mathbf{I})\}_{i=1}^n$ which initializes a new diffusion trajectory. This resampling promotes exploration across distinct diffusion paths and reduces sensitivity to starting noise. Crucially, this framework drives $\boldsymbol{\mathcal{Z}}$ to generalize across noise realizations and prevent overfitting to a specific diffusion trajectory. See Alg. 1 (lines 3-5) for details and Fig. 1,2.

Inner loop (Joint refinement). Within each outer loop, we run an inner loop that simulates the reverse diffusion process from $\boldsymbol{\mathcal{X}}_T^k$ to the denoised counterpart $\boldsymbol{\mathcal{X}}_0^k$. At each reverse diffusion step $t \in [T, \ldots, 0]$, we perform two coupled updates.

Embedding updates. We update $\boldsymbol{\mathcal{Z}}_t^k$ by doing gradient ascent on the experimental likelihood with respect to the embeddings:

$$\boldsymbol{\mathcal{Z}}_{t-1}^k = \boldsymbol{\mathcal{Z}}_t^k + \eta_z(\nabla_{\boldsymbol{\mathcal{Z}}_t^k}\log p(\mathbf{y}|\hat{\boldsymbol{\mathcal{X}}}_0) + \lambda_{\mathrm{p}}\nabla_{\boldsymbol{\mathcal{Z}}_t^k}\log p(\boldsymbol{\mathcal{Z}}_t^k|\mathbf{a})) \tag{5}$$

where η_z is the learning rate and λ_p is the prior weight. This update shifts the learned conditioning manifold toward regions of the embedding space that are more consistent with $\mathbf{y}$. Since the embeddings parameterize the structure generation process, this step indirectly biases subsequent structural samples toward experimentally faithful conformations.

Reverse SDE. The updated embeddings $\mathcal{Z}_{t-1}^k = \{\mathbf{Z}_{t-1}^{k,1}, \ldots, \mathbf{Z}_{t-1}^{k,n}\}$ are injected into the SDE, yielding the following reverse step for each ensemble member $\mathbf{X}^{k,i} \in \mathcal{X}^k$:

$$d\mathbf{X}^{k,i} = -\Big(\frac{1}{2}\mathbf{X}^{k,i} + \nabla_{\mathbf{X}^{k,i}} \log p_t(\mathbf{X}^{k,i}|\mathbf{Z}_{t-1}^{k,i})\Big)\sigma_t dt + \sqrt{\sigma_t}\mathbf{N}^i$$

Note that the score function is predicted by the model conditioned on the updated embeddings. With this alternating procedure, the conditioning variables act as a persistent memory of the experimental manifold across diffusion steps. Importantly, the embeddings obtained at the end of the inner loop $\mathcal{Z}_0^k$ are used to initialize the next outer iteration $\mathcal{Z}_T^{k+1}$, enabling cumulative refinement rather than restarting the optimization as shown in Fig. 8. See Alg. 1 (lines 6–8) and Figs. 1, 2.

Our framework thus serves as a meta-initialization; in our experiments, the embeddings $\mathcal{Z}_0^K$ obtained after the convergence of the outer loops are used as conditioning variables for the coordinate-space guidance procedure in Eq. 4. A detailed pseudocode can be found in Alg. 2, and additional implementation details are specified in the App. E.4.

Boltzmann re-weighting. While the log-likelihood in Eq.5 uniformly weights ensemble members, physical ensembles in solution follow a Boltzmann distribution (Barducci et al., 2011). To bias the ensemble toward thermodynamically plausible conformations, we refine the AF3-induced structural prior by tilting it with a differentiable energy function $E_\phi : \mathbb{R}^{m\times 3} \to \mathbb{R}$

$$\pi(\mathcal{X}|\mathcal{Z}) \propto p(\mathcal{X}|\mathcal{Z}) \cdot \exp(-\beta \sum_i E_\phi(\mathbf{X}^i)) \tag{6}$$

where $\beta = 1.68\ \mathrm{kcal}^{-1}\,\mathrm{mol}$ is the inverse temperature at physiological conditions ($300\,\mathrm{K}$). Here, the force field is treated as a thermodynamic prior that biases AF3 towards energetically favorable conformations. Because the normalizing constant of the tilted distribution π is intractable, we adopt a self-normalized importance sampling (SNIS) perspective (Skreta et al., 2025). Expectations under π are approximated by assigning Boltzmann weights w^i to samples drawn from the AF3 prior,

$$w^i = \frac{\exp(-\beta E_\phi(\mathbf{X}^i))}{\sum_{k=1}^n \exp(-\beta E_\phi(\mathbf{X}^k))}. \tag{7}$$

Under this reweighting, experimental observables are evaluated as Boltzmann-weighted ensemble averages, emphasizing low-energy conformations while preserving consistency with the AF3 prior. Additional implementation details, including temperature annealing and energy smoothing strategies, are provided in App. E.1.4.

5 Organization

Additional details about experiments and implementation are deferred to the Appendix. Appendix A provides an overview of the Appendix structure. Appendices B–D introduce the experimental log-likelihood terms, the case studies built around these likelihoods, and the baseline methods used for comparison. Appendices E–I present implementation details, evaluation metrics, related work, figures, algorithms, and tables with quantitative performance results.

6 Conclusion

In this work, we introduce an inference-time optimization (IT-Opt) framework that mitigates key limitations of traditional guidance and enables thermodynamically consistent ensemble generation. Across NMR and X-ray crystallography benchmarks, IT-Opt improves agreement with experimental observables and recovers challenging structural features. We further show that confidence metrics such as ipTM can be artificially inflated by small perturbations, revealing limitations of current confidence frameworks. Future work will extend IT-Opt to additional modalities, including cryo-EM.

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, 2024.

Jon Agirre, Mihaela Atanasova, Haroldas Bagdonas, Charles B Ballard, Arnaud Basle, James Beilsten-Edmands, Rafael J Borges, David G Brown, Joana Burgos-Maúrtinez, John M Berrisford, et al. The ccp4 suite: integrative software for macromolecular crystallography. *Acta Crystallographica Section D: Structural Biology*, 79(6):449–461, 2023.

Tara Akhound-Sadegh, Jarrid Rector-Brooks, Avishek Joey Bose, Sarthak Mittal, Pablo Lemos, Cheng-Hao Liu, Marcin Sendera, Siamak Ravanbakhsh, Gauthier Gidel, Yoshua Bengio, Nikolay Malkin, and Alexander Tong. Iterated denoising energy matching for sampling from boltzmann densities. In *International Conference on Machine Learning*, 2024.

T Reid Alderson, Iva Pritišanac, Desika Kolarić, Alan M Moses, and Julie D Forman-Kay. Systematic identification of conditionally folded intrinsically disordered regions by alphafold2. *Proceedings of the National Academy of Sciences*, 120(44):e2304302120, 2023.

Mois I. Aroyo (ed.). *International tables for crystallography volume A: Space-group symmetry*. Wiley, 6th edition, 2016.

Edgar N Baker and Roderick E Hubbard. Hydrogen bonding in globular proteins. *Progress in biophysics and molecular biology*, 44(2):97–179, 1984.

Alessandro Barducci, Massimiliano Bonomi, and Michele Parrinello. Metadynamics. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 1(5):826–843, 2011.

Patrick Bryant and Frank Noé. Improved protein complex prediction with alphafold-multimer by denoising the msa profile. *PLOS Computational Biology*, 20(7):e1012253, 2024.

Stephen K Burley, Helen M Berman, Gerard J Kleywegt, John L Markley, Haruki Nakamura, and Sameer Velankar. Protein data bank (pdb): the single global macromolecular structure archive. In *Protein Crystallography: Methods and Protocols*, volume 1607 of *Methods in Molecular Biology*, pp. 627–641. Springer, 2017. doi: 10.1007/978-1-4939-7000-1_26.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Jennifer A Doudna. Structural genomics of rna. *Nature Structural Biology*, 7(Suppl 11):954–956, 2000. doi: 10.1038/80729.

Roland L Dunbrack Jr. What's wrong with alphafold's iptm score and how to fix it. *bioRxiv*, 2025. doi: 10.1101/2025.02.10.637595.

Peter Eastman. Pdbfixer. `https://github.com/openmm/pdbfixer`, 2013.

Peter Eastman, Jason Swails, John D Chodera, Robert T McGibbon, Yutong Zhao, Kyle A Beauchamp, Lee-Ping Wang, Andrew C Simmonett, Matthew P Harrigan, Chaya D Stern, et al. Openmm 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS computational biology*, 13(7):e1005659, 2017.

Richard Evans, Michael O'Neill, Alexander Pritzel, Natasha Antropova, Andrew Senior, Tim Green, Augustin Žídek, Russ Bates, Sam Blackwell, Jason Yim, et al. Protein complex prediction with alphafold-multimer. *bioRxiv*, 2022. doi: 10.1101/2021.10.04.463034.

Alisia Fadini, Minhuan Li, Airlie J McCoy, Thomas C Terwilliger, Randy J Read, Doeke Hekstra, and Mohammed AlQuraishi. Alphafold as a prior: Experimental structure determination conditioned on a pretrained neural network. *bioRxiv*, 2025. doi: 10.1101/2025.02.18.638828.

Nicholas Furnham, Tom L Blundell, Mark A DePristo, and Thomas C Terwilliger. Is one solution good enough? *Nature structural & molecular biology*, 13(3):184–185, 2006.

Ronald J Gillespie. The vsepr model revisited. *Chemical Society Reviews*, 21(1):59–69, 1992.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

John B Ingraham, Max Baranov, Zak Costello, Karl W Barber, Wujie Wang, Ahmed Ismail, Vincent Frappier, Dana M Lord, Christopher Ng-Thow-Hing, Erik R Van Vlack, et al. Illuminating protein space with a programmable generative model. *Nature*, 623(7989):1070–1078, 2023.

Bowen Jing, Bonnie Berger, and Tommi Jaakkola. Alphafold meets flow matching for generating protein ensembles. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 22277–22303. PMLR, 2024.

Robbie P Joosten, Fei Long, Garib N Murshudov, and Anastassis Perrakis. The pdb_redo server for macromolecular structure model optimization. *IUCrJ*, 1(4):213–220, 2014.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Wolfgang Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, 32(5):922–923, 1976. doi: 10.1107/S0567739476001873.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Piotr Klukowski, Fred F Damberger, Frédéric H-T Allain, Hideo Iwai, Harindranath Kadavath, Theresa A Ramelot, Gaetano T Montelione, Roland Riek, and Peter Güntert. The 100-protein nmr spectra dataset: A resource for biomolecular nmr data analysis. *Scientific data*, 11(1):30, 2024.

Patrick Kunzmann, Jacob Marcel Anter, and Kay Hamacher. Adding hydrogen atoms to molecular models via fragment superimposition. *Algorithms for Molecular Biology*, 17(1):7, 2022. doi: 10.1186/s13015-022-00215-x.

Advaith Maddipatla, Nadav Sellam Bojan, Meital Bojan, Volodymyr Masalitin, Sanketh Vedula, Paul Schanda, Ailie Marx, and Alex M Bronstein. Experiment-guided alphafold3 resolves accurate protein ensembles. *bioRxiv*, 2025a. doi: 10.1101/2025.10.11.681796.

Advaith Maddipatla, Nadav Bojan Sellam, Meital Bojan, Sanketh Vedula, Paul Schanda, Ailie Marx, and Alex M Bronstein. Inverse problems with experiment-guided alphafold. In *Forty-second International Conference on Machine Learning*, 2025b.

Sai Advaith Maddipatla, Nadav Bojan Sellam, Sanketh Vedula, Ailie Marx, and Alex Bronstein. Generative modeling of protein ensembles guided by crystallographic electron densities. *Machine Learning for Structural Biology Workshop at NeurIPS 2024*, 2024.

Ian K McDonald and Janet M Thornton. Satisfying hydrogen bonding potential in proteins. *Journal of molecular biology*, 238(5):777–793, 1994.

Milot Mirdita, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger. Colabfold: making protein folding accessible to all. *Nature methods*, 19(6):679–682, 2022.

Garib N Murshudov, Pavol Skubák, Andrey A Lebedev, Navraj S Pannu, Roberto A Steiner, Robert A Nicholls, Martyn D Winn, Fei Long, and Alexei A Vagin. Refmac5 for the refinement of macromolecular crystal structures. *Acta Crystallographica Section D: Biological Crystallography*, 67(4): 355–367, 2011.

Frank Noé, Simon Olsson, Jonas Köhler, and Hao Wu. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science*, 365(6457):eaaw1147, 2019.

Moshe Oren. Regulation of the p53 tumor suppressor protein. *Journal of Biological Chemistry*, 274 (51):36031–36034, 1999.

Martin Pacesa, Lennart Nickel, Christian Schellhaas, Joseph Schmidt, Ekaterina Pyatova, Lucas Kissling, Patrick Barendse, Jagrity Choudhury, Srajan Kapoor, Ana Alcaraz-Serna, et al. One-shot design of functional protein binders with bindcraft. *Nature*, 646:483–492, 2025. doi: 10.1038/s41586-025-09429-6.

Saro Passaro, Gabriele Corso, Jeremy Wohlwend, Mateo Reveiz, Stephan Thaler, Vignesh Ram Somnath, Noah Getz, Tally Portnoi, Julien Roy, Hannes Stark, David Kwabi-Addo, Dominique Beaini, Tommi Jaakkola, and Regina Barzilay. Boltz-2: Towards accurate and efficient binding affinity prediction. *bioRxiv*, 2025. doi: 10.1101/2025.06.14.659707.

Edward Prince (ed.). *International Tables for Crystallography, Volume C: Mathematical, physical and chemical tables*. Kluwer Academic Publishers, 3rd edition, 2004.

Rishwanth Raghu, Axel Levy, Gordon Wetzstein, and Ellen D. Zhong. Multiscale guidance of protein structure prediction with heterogeneous cryo-em data. *Advances in neural information processing systems*, 2025.

James P Roney, Chenxi Ou, and Sergey Ovchinnikov. Protein diffusion models as statistical potentials. *bioRxiv*, pp. 2025.12.09.693073, 2025. doi: 10.64898/2025.12.09.693073.

Aviv A Rosenberg, Sanketh Vedula, Alex M Bronstein, and Ailie Marx. Seeing double: Molecular dynamics simulations reveal the stability of certain alternate protein conformations in crystal structures. *bioRxiv*, pp. 2024.08.31.610605, 2024. doi: 10.1101/2024.08.31.610605.

Marta Skreta, Tara Akhound-Sadegh, Viktor Ohanesian, Roberto Bondesan, Alán Aspuru-Guzik, Arnaud Doucet, Rob Brekelmans, Alexander Tong, and Kirill Neklyudov. Feynman-kac correctors in diffusion: Annealing, guidance, and product of experts. In *Forty-second International Conference on Machine Learning*, 2025.

Andrey Smelter, Morgan Astra, and Hunter NB Moseley. A fast and efficient python library for interfacing with the biological magnetic resonance data bank. *BMC bioinformatics*, 18(1):175, 2017. doi: 10.1186/s12859-017-1580-5.

MS Smyth and JHJ Martin. x ray crystallography. *Molecular Pathology*, 53(1):8–14, 2000. doi: 10.1136/mp.53.1.8.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. pmlr, 2015.

ByteDance AML AI4Science Team, Xinshi Chen, Yuxuan Zhang, Chan Lu, Wenzhi Ma, Jiaqi Guan, Chengyue Gong, Jincai Yang, Hanyu Zhang, Ke Zhang, Shenghao Wu, Kuangqi Zhou, Yanping Yang, Zhenyu Liu, Lan Wang, Bo Shi, Shaochen Shi, and Wenzhi Xiao. Protenix - advancing structure prediction through a comprehensive alphafold3 reproduction. *bioRxiv*, 2025. doi: 10.1101/2025.01.08.631967. URL https://www.biorxiv.org/content/early/2025/01/11/2025.01.08.631967.

Eldon L Ulrich, Kumaran Baskaran, Hesam Dashti, Yannis E Ioannidis, Miron Livny, Pedro R Romero, Dimitri Maziuk, Jonathan R Wedell, Hongyang Yao, Hamid R Eghbalnia, et al. Nmr-star: comprehensive ontology for representing, archiving and exchanging data from nuclear magnetic resonance spectroscopic experiments. *Journal of biomolecular NMR*, 73(1-2):5–9, 2019. doi: 10.1007/s10858-018-0220-3.

Julia K Varga, Sergey Ovchinnikov, and Ora Schueler-Furman. actifptm: a refined confidence metric of alphafold2 predictions involving flexible regions. *Bioinformatics*, 41(3):btaf107, 2025.

Beat Vögeli. The nuclear overhauser effect from a quantitative perspective. *Progress in nuclear magnetic resonance spectroscopy*, 78:1–46, 2014.

Junmei Wang, Romain M Wolf, James W Caldwell, Peter A Kollman, and David A Case. Development and testing of a general amber force field. *Journal of computational chemistry*, 25(9): 1157–1174, 2004.

Hannah K Wayment-Steele, Adedolapo Ojoawo, Renee Otten, Julia M Apitz, Warintra Pitsawong, Marc Hömberger, Sergey Ovchinnikov, Lucy Colwell, and Dorothee Kern. Predicting multiple conformations via sequence clustering and alphafold2. *Nature*, 625(7996):832–839, 2024.

Jeffrey Wee and Guo-Wei Wei. Evaluation of alphafold 3's protein-protein complexes for binding free energy predictions. *Journal of Physical Chemistry Letters*, 2024. doi: 10.1021/acs.jpclett.4c02110.

Marcin Wojdyr. Gemmi - library for structural biology [software]. `https://github.com/project-gemmi/gemmi`, 2021.

Silong Zhai, Huifeng Zhao, Jike Wang, Shaolong Lin, Tiantao Liu, Shukai Gu, Dejun Jiang, Huanxiang Liu, Yu Kang, Xiaojun Yao, et al. Peppcbench is a comprehensive benchmarking framework for protein–peptide complex structure prediction. *Journal of Chemical Information and Modeling*, 65(16):8497–8513, 2025.

A APPENDIX ORGANIZATION

The Appendix is organized as follows:

- Appendix B describes the data terms used in our method: NOE distance restraints (Section B.1); real-space crystallographic density maps (Section B.2); and the ipTM confidence score (Section B.3).
- Appendix C details the experiments performed in this work: electron density-based IT-optimization for alternative conformations and protein-bound peptides (Section C.1); NOE-based IT-optimization with and without energy reweighting (Section C.2); and ipTM-based IT-optimization (Section C.3).
- Appendix D describes the baselines used in our experiments.
- Appendix E provides implementation details for NMR, X-ray crystallography, and ipTM-based IT-optimization.
- Appendix F reviews related work on experimental guidance, MSA manipulation, and ipTM.
- Appendix G- I collects all figures, tables, and algorithms, respectively.

B DATA TERMS

In this paper, we consider three data terms $\log p(\mathbf{y}|\boldsymbol{\mathcal{X}})$: (i) NOE distance restraints (Section B.1); (ii) real-space crystallographic density maps (Section B.2); and (iii) the ipTM confidence score (Section B.3).

B.1 NUCLEAR OVERHAUSER EFFECT RESTRAINTS

Nuclear Overhauser Effect (NOE) distance restraints derived from NMR spectroscopy encode pairwise proximity constraints between non-bonded atoms $(i, j) \in \mathcal{P}$. The likelihood of observing an NOE effect between two atoms depends on their time-averaged distance, r, as $1/r^6$. However, as other parameters, such as local dynamics, play a role, determining precise distances, although possible, is not commonly done (Vögeli, 2014), and only lower and upper distance bounds, $[\underline{d}_{ij}, \bar{d}_{ij}]$, are used. Given an ensemble $\boldsymbol{\mathcal{X}}$ and the set of restraints $\mathbf{D} = \{(\underline{d}_{ij}, \bar{d}_{ij}) : (i, j) \in \mathcal{P}\}$, the log-likelihood is

$$\log p(\mathbf{D} \mid \boldsymbol{\mathcal{X}}) = \quad - \sum_{(i,j)\in\mathcal{P}} \left(\left[\underline{d}_{ij} - d_{ij}(\boldsymbol{\mathcal{X}})\right]_+^2 + \left[d_{ij}(\boldsymbol{\mathcal{X}}) - \bar{d}_{ij}\right]_+^2 \right) \tag{8}$$

where $[\cdot]_+ = \max(\cdot, 0)$ and $d_{ij}(\boldsymbol{\mathcal{X}}) = \frac{1}{n}\sum_{k=1}^{n} \|\mathbf{x}_i^{(k)} - \mathbf{x}_j^{(k)}\|$ is the ensemble-averaged interatomic distance.

B.2 CRYSTALLOGRAPHIC ELECTRON DENSITIES

The observed real-space electron density map $F_{\mathrm{o}} \colon \mathbb{R}^3 \to \mathbb{R}$ represents an ensemble over the molecular conformations present in a crystal lattice. In X-ray crystallography, 2D diffraction images aggregate scattering contributions from a large population of molecules, such that the resulting electron density reflects an average over conformational heterogeneity (Smyth & Martin, 2000). Typical crystallographic pipelines attempt to fit an ensemble $\boldsymbol{\mathcal{X}}$ that best explains F_{o} (Doudna, 2000). This objective can be formulated as maximizing the likelihood $\log p(F_{\mathrm{o}} \mid \boldsymbol{\mathcal{X}})$. We define the log-likelihood as the negative L_1 distance between the observed and ensemble-averaged calculated density maps

$$\log p(F_{\mathrm{o}}|\boldsymbol{\mathcal{X}}) = -\int_{\Omega} \left\| F_{\mathrm{o}}(\boldsymbol{\xi}) - \frac{1}{n}\sum_{k=1}^{n} F_{\mathrm{c}}(\boldsymbol{\xi}; \mathbf{X}^k) \right\|_1 d\boldsymbol{\xi} \tag{9}$$

where, the loss is computed over the asymmetric unit cell Ω and $F_{\mathrm{c}}(\boldsymbol{\xi}; \mathbf{X}^k)$ is the theoretical electron density computed from the atomic coordinates of structure $\mathbf{X}^k$ (See Appendix E.2 for details).

B.3 INTERFACE PREDICTED TEMPLATE MODELING SCORE

The interface predicted template modeling (ipTM) score is a learned confidence metric produced by AF3 that quantifies the reliability of predicted inter-chain interfaces in multimeric protein structures (Bryant & Noé, 2024). High ipTM prediction indicates strong model confidence in the relative placement and contacts between chains, while low ipTM values typically correspond to poorly formed or disordered interfaces. Unlike experimental observations, ipTM is not a physical measurement; instead, it is derived from a neural network trained to estimate structural accuracy by comparing predicted interfaces against known protein complexes (Jumper et al., 2021; Abramson et al., 2024; Passaro et al., 2025). In our framework, we treat ipTM as a surrogate likelihood that provides a differentiable signal for assessing interfacial consistency. The data term can be defined as,

$$\log p(\mathbf{y} \mid \boldsymbol{\mathcal{X}}) \propto \sum_{i=1}^{n} \mathrm{ipTM}(\mathbf{X}^i) \tag{10}$$

where $\mathbf{y}$ denotes the (unknown) true interface configuration. Unlike other likelihoods, these terms are separable. By maximizing this likelihood, we bias the ensemble toward interface geometries that are assigned higher confidence by the model, thereby promoting coherent inter-chain arrangements.

C EXPERIMENTS

C.1 ELECTRON DENSITY-BASED IT-OPTIMIZATION

Alternative Conformations. In loop regions, density maps often exhibit multimodal backbone density due to alternative conformations (altlocs). As shown by Rosenberg et al. (2024), AF3 typically collapses such regions to a single mode, failing to accurately capture conformational heterogeneity. While guidance-based methods can recover multimodal ensembles, they degrade when alternative states are well separated (Maddipatla et al., 2025a). We illustrate this limitation using the laminarinase catalytic domain (Figure 3), where chain A of `3AZY` (1.65 Å) contains a seven-residue altloc region. AF3 recovers only one mode, and guidance produces a bimodal ensemble but misfits the backbone for one mode. In contrast, IT-Optimization accurately captures both modes, including sidechain placement. Table 10 further shows that IT-Optimization yields significantly more consistent ensembles across 5 random seeds than guidance on the same proteins. Similar trends are observed across additional altloc-containing structures (Tables 7– 8).

Protein-bound peptides. Short peptide chains are challenging for AF3 due to their high conformational flexibility. Guidance-based approaches simplify this setting by fixing the peptide N-C termini and optimizing only internal residues (Maddipatla et al., 2025a). Here, we remove this constraint and fit the entire peptide directly into the density map using both guidance and IT-optimization, and without the need to anchor any atoms or residues. Figure 3 shows chain B of `6I42` (1.38Å), a 13-residue peptide from the α-synuclein-cyclophilin A complex. AF3 fails to reproduce the experimental structure, and guidance improves only the backbone placement but yields incorrect side-chain packing. In contrast, IT-Optimization generates ensembles with well-fitted backbone and side-chain arrangements. Moreover, as shown in Figure 8, iterative optimization enables continued refinement until convergence, whereas guidance is limited to a fixed number of iterations. Consistent improvements are observed across the full benchmark (Tables 7–8).

C.2 NOE-BASED IT-OPTIMIZATION

We benchmark the proposed framework on a set of 20 protein structures from the NMRDB dataset (Klukowski et al., 2024) (Table 1). As shown in Tables 4, 3, IT-Optimization with uniformly weighted NMR log-likelihood (Equation 8) consistently reduces both the number of violated NOE restraints and the median violation distance compared to guided and unguided baselines. Incorporating energy reweighting (Equation 7) using force-field-predicted energies further improves NOE satisfaction while promoting thermodynamically plausible ensembles (Tables 6, 5). As a representative example, we consider the solution-NMR ensemble of an uncharacterized Rhodospirillum rubrum protein (PDB `2K0M`, BMRB 15652). Figure 5 shows that IT-Optimization reduces restraint violations relative to competing methods, with additional gains when combined with energy reweighting. This also results

in improved structural agreement with the reference ensemble (Figure 4). Lastly, energy-weighted variants produce lower-energy ensembles than their uniformly weighted counterparts under the optimization force field (ProteinEBM (Roney et al., 2025)) and the known AMBER99 force field (Wang et al., 2004) (Figure 7).

C.3 ipTM-BASED IT-OPTIMIZATION

We systematically analyzed the effects of inference-time optimization of the ipTM objective on predicted protein complexes, with the goal of characterizing how such optimization interacts with the AF3 embedding space. Given the widespread use of ipTM as a confidence and ranking metric in large-scale complex screening and binder design, it is important to understand how perturbations in the AF3 embeddings influence ipTM values and the resulting structural outputs.

We conduct the following perturbation experiment. For each target complex, we perform inference-time optimization of the AF3 embeddings to increase the predicted interfacial confidence of the resulting structure, as measured by the ipTM score (and related weighted metrics). Because AF3's ipTM predictor depends jointly on the AF3 embeddings and the predicted complex structure, it is sensitive to perturbations in the embedding space. Optimization is performed by iteratively perturbing the AF3 embeddings and resampling structures using the AF3 diffusion model to increase ipTM. A maximum relative perturbation budget is enforced on the AF3 embeddings; upon reaching this limit, optimization is halted, and the resulting embeddings are used for structure sampling.

The evaluated complexes comprise three distinct classes. Firstly, complexes involving the tumor repressor p53 determined by either NMR and focusing on the transiently ordered N-terminal domain (Oren, 1999); `2K8F`, `2L14`, and `2LY4` and the X-ray crystal structure characterizing the core domain interactions; `1YCS`. Here, unguided AF3 predictions exhibit low baseline confidence in terms of both pTM and ipTM. These systems, therefore, represent challenging test cases for confidence-driven inference-time optimization. Secondly, we consider de novo designed protein–protein interfaces generated using BindCraft; (`9HAD`, `9HAF`). These targets typically exhibit higher baseline confidence and provide a complementary regime for assessing the sensitivity of ipTM optimization in designed binding scenarios. Lastly, we consider `8Q70` a natural heterodimer with domain swap.

Sensitivity of ipTM to embedding perturbations. Using the inference-time optimization procedure described above, we first examine how the ipTM-driven objective (Eq. 10) evolves under continued optimization. Across all evaluated targets, both ipTM and pTM components can be systematically increased within the imposed perturbation budget (Fig. 10). Notably, these increases are often achieved with only very small relative changes to the embedding ($\approx 0.01\%$), indicating that the confidence objective is highly sensitive to local directions in embedding space. However, the figure further shows that increases in confidence are not consistently accompanied by improvements in structural accuracy: while confidence scores generally increase monotonically with perturbation budget, RMSD to experiment is target-dependent and may remain unchanged or even worsen in some cases, particularly in the absence of MSA input. Together, these results suggest that ipTM is highly sensitive to local variations in the embedding space and can be driven to high-confidence regimes via small perturbations, but that such increases do not necessarily imply improved agreement with experiment.

ipTM-guidance improves complex predictions in some cases. Due to this sensitivity, ipTM optimization can sometimes coincide with improved interfacial geometry and experimental agreement. In the crystallographic complex `1YCS`, ipTM-based IT-Optimization recovers a helix-mediated contact missed by AF3 (Fig. 6A) and increases the recovery of native inter-chain hydrogen bonds, yielding a higher average contact recovery rate, reflecting a larger fraction of experimentally observed contacts recovered per sampled structure (Table 11). A similar effect is observed for the NMR complex `2YL4`, which consists of a globular target bound to the transiently ordered N-terminal domain of `p53`. While unguided AF3 samples both an exterior binding mode associated with higher ipTM scores and an interior binding mode supported by the experimental NMR ensemble, ipTM-based inference-time optimization consistently favors the interior binding configuration across random initializations (Fig. 6B). This shift is accompanied by a substantial enrichment in experimentally observed inter-chain hydrogen bonds within the peptide helical region, indicating that ipTM optimization in this case promotes structures that better reflect the experimentally supported binding mode (Table 12). In low-information settings, ipTM-based inference-time optimization generally does not correct major

structural errors. For the designed complex 9AHF without MSA input, AF3 mispredicts one chain (27% agreement). Optimization increases secondary-structure agreement to 59.2% (Fig. 11), but does not recover the correct fold, suggesting confidence-driven optimization can partially reduce errors but is not reliably predictive in low-information regimes.

D BASELINES

For the X-ray and NMR experiments, we benchmark ensembles generated using inference-time optimization against three baselines: the guided AlphaFold3 framework (Maddipatla et al., 2025a), experimentally determined PDB structures (Burley et al., 2017), and unguided sequence-to-structure generative models, including AlphaFold3 (Abramson et al., 2024) and AlphaFlow (Jing et al., 2024). For ipTM experiments, we use the corresponding PDB complex (when available) as a reference and compare our multimeric ensembles to AlphaFold3, as AlphaFlow does not support multimer modeling.

E IMPLEMENTATION DETAILS

E.1 NMR ADDITIONAL DETAILS

E.1.1 EXTRACTING AND PROCESSING NOE DISTANCE RESTRAINTS

We obtain interatomic distance bounds from NMR-STAR (Ulrich et al., 2019) formatted depositions by reading them with the `pynmrstar` package (Smelter et al., 2017). Entries classified as NOE-type distance constraints are retained, while other restraint categories are discarded. When a cross-peak cannot be uniquely assigned to a single atom pair, it is represented as an OR-group comprising multiple candidate pairs, where satisfying any one pair fulfills the restraint. The distance bounds ($\underline{d}_{ij}$ and $\bar{d}_{ij}$) are taken from the deposited file in the PDB (Burley et al., 2017). Occasionally, where a lower bound is absent, we substitute 0.0 Å. The resulting restraint set feeds directly to the NOE log-likelihood in Section B.1.

Differentiable hydrogen placement. AlphaFold3 operates exclusively on non-hydrogen atoms, yet NOE observables depend on the interatomic distance between hydrogen atoms. Rather than approximating each restraint at the level of the parent non-hydrogen atom, we reconstruct explicit hydrogen coordinates at every sampling step using a differentiable PyTorch re-implementation of the Hydride placement algorithm (Kunzmann et al., 2022). For each hydrogen-bearing non-hydrogen atom, the local bonding topology is matched to a library of reference fragments; a rigid-body superposition of the fragment's heavy atoms onto the current structure then determines the corresponding proton positions. Because both fragment lookup and superposition are differentiable with respect to non-hydrogen atom coordinates, reconstructed hydrogen positions vary smoothly during diffusion, enabling gradient-based guidance on physically meaningful interatomic distances.

E.1.2 RELAXATION

In order to make sure that the generated ensemble has no structural violations, we minimize its energy using an off-the-shelf harmonic force field. In this work, we use OpenMM's (Eastman et al., 2017) implementation of the AMBER99 (Wang et al., 2004). The energy is minimized for a maximum of 2000 iterations with an energy tolerance threshold of 2.39 kcal$/$mol and stiffness of 100.0 kcal / mol Å^2.

E.1.3 NMR EVALUATION METRICS

We quantify how well the relaxed predicted ensemble $\boldsymbol{\mathcal{X}}$ respects the experimentally derived NOE bounds using the two complementary statistics.

Percentage of Violated NOE Constraints. Consider a restraint list organized into M OR-groups $\{G_1, \ldots, G_M\}$, each containing one or more candidate atom pairs $\mathbf{r} = (i,\, j,\, \underline{d}_{ij},\, \bar{d}_{ij})$. For every

pair of atoms i, j in structure k of $\boldsymbol{\mathcal{X}}$, we first compute the weighted distance across the ensemble,

$$d_{ij}(\boldsymbol{\mathcal{X}};\, \mathbf{w}) = \sum_{k=1}^{|\boldsymbol{\mathcal{X}}|} w_k \left\| \mathbf{x}_i^k - \mathbf{x}_j^k \right\|_2, \tag{11}$$

where $\mathbf{w} = (w_1, \ldots, w_{|\boldsymbol{\mathcal{X}}|})$ are non-negative weights summing to one. Choosing $w_k = 1/|\boldsymbol{\mathcal{X}}|$ yields a uniform ensemble average; alternatively, Boltzmann weights derived from an energy model (Equation 7) emphasize low-energy conformers. We then quantify the extent to which this weighted distance violates the allowed bounds,

$$v_{ij} = \max\left\{ \underline{d}_{ij} - d_{ij}(\boldsymbol{\mathcal{X}};\, \mathbf{w}),\ d_{ij}(\boldsymbol{\mathcal{X}};\, \mathbf{w}) - \bar{d}_{ij},\ 0 \right\}. \tag{12}$$

Because an OR-group is satisfied whenever at least one candidate lies within the allowed bounds, the group-level violation is defined as $v_{G_m} = \min_{\mathbf{r} \in G_m} v_{ij}$. The overall violation rate is then

$$\text{Viol. }\% = 100 \times \frac{|\{m : v_{G_m} > 0\}|}{M}. \tag{13}$$

Median Violation Magnitude To characterize the typical magnitude of violations, we additionally report

$$\text{Viol. Å} = \text{median}\left(\{v_{G_m} : v_{G_m} > 0\}\right)\ [\text{Å}], \tag{14}$$

computed only over restraint groups that are violated. While the violation percentage quantifies how many restraints are unsatisfied, this metric captures the severity of those violations. Reporting both measures provides a more complete characterization: an ensemble may violate only a few restraints but by large margins, or violate many restraints by small amounts.

E.1.4 Boltzmann Reweighting for Energy-Guided Ensembles

Ensembles obtained via experimental guidance are constrained to match the measured data, but their relative populations are not thermodynamically informed. We therefore re-weight ensemble members during the diffusion process using an energy model to recover experimentally faithful, thermodynamic ensembles.

Self-normalized importance sampling perspective. We interpret energy reweighting through a self-normalized importance sampling (SNIS) lens. Let the proposal distribution be the AF3 ensemble prior $q(\boldsymbol{\mathcal{X}}) = p(\boldsymbol{\mathcal{X}} \mid \boldsymbol{\mathcal{Z}}, \mathbf{a})$ and define the target distribution as

$$\pi(\boldsymbol{\mathcal{X}}) \ \propto\ p(\boldsymbol{\mathcal{X}} \mid \boldsymbol{\mathcal{Z}}, \mathbf{a}) \exp\Big(-\beta \textstyle\sum_{i=1}^{n} E_\phi(\mathbf{X}^i)\Big), \tag{15}$$

which tilts the AF3 prior toward thermodynamically favorable conformations. Following the importance sampling framework for Boltzmann densities (Noé et al., 2019), we treat the unmodified prior $q(\boldsymbol{\mathcal{X}}) = p(\boldsymbol{\mathcal{X}} \mid \boldsymbol{\mathcal{Z}}, \mathbf{a})$ as the proposal distribution. The importance ratio for any sample $\boldsymbol{\mathcal{X}} \sim q$ is

$$\frac{\pi(\boldsymbol{\mathcal{X}})}{q(\boldsymbol{\mathcal{X}})} \ \propto\ \exp\Big(-\beta \textstyle\sum_{i=1}^{n} E_\phi(\mathbf{X}^i)\Big), \tag{16}$$

where the proposal density cancels entirely. The self-normalized importance sampling (SNIS) estimator for an observable f under π is $\mathbb{E}_\pi[f] \approx \sum_{k=1}^{n} w_k\, f(\mathbf{X}^k)$, with unnormalized weights $\tilde{w}_k \propto \exp(-\beta E_\phi(\mathbf{X}^k))$. After normalization, these recover the Boltzmann weights in Equation 7.

Given a differentiable potential $E_\phi(\mathbf{X})$, the canonical probability of a conformation is

$$\pi(\mathbf{X}) \propto \exp\big(-\beta\, E_\phi(\mathbf{X})\big), \tag{17}$$

where $\beta = (k_B T_{\text{therm}})^{-1}$ is the inverse temperature and $k_B = 0.001987\ \text{kcal}\,\text{mol}^{-1}\,\text{K}^{-1}$ is the Boltzmann constant. At physiological conditions ($T_{\text{therm}} = 300\,\text{K}$), $\beta \approx 1.68\ \text{kcal}^{-1}\,\text{mol}$. Setting $\mathbf{w}$ in Equation 11 to the normalized Boltzmann factors yields the energy-reweighted distance

$$d_{ij}^{w}(\boldsymbol{\mathcal{X}}) = \sum_{k=1}^{n} w_k \left\| \mathbf{x}_i^k - \mathbf{x}_j^k \right\|_2, \tag{18}$$

with

$$w_k = \frac{\exp\big(-\beta\, E_\phi(\mathbf{X}^k)\big)}{\sum_{j=1}^{n} \exp\big(-\beta\, E_\phi(\mathbf{X}^j)\big)}. \tag{19}$$

Substituting these weighted distances into the NOE log-likelihood (Equation 8) gives

$$\log p(D \mid \boldsymbol{\mathcal{X}};\, \mathbf{a}) = -\sum_{(i,j)\in D} \Big(\big[\underline{d}_{ij} - d^w_{ij}(\boldsymbol{\mathcal{X}})\big]_+^2 + \big[d^w_{ij}(\boldsymbol{\mathcal{X}}) - \bar{d}_{ij}\big]_+^2 \Big). \tag{20}$$

Low-energy conformers thereby exert a stronger pull on the ensemble distance, steering the generated structures toward regions of the energy surface that are both experimentally consistent and energetically favorable. In the limit $\beta \to 0$ (high temperature), all weights become uniform, and the unweighted formulation is recovered; as $\beta \to \infty$ (low temperature), the weights concentrate on the single lowest-energy member.

Annealing the Inverse Temperature. Applying a sharp energy bias from the beginning would restrict conformational exploration before the diffusion trajectory has resolved meaningful structural detail. We therefore anneal the inverse temperature β across diffusion steps according to a three-phase schedule:

$$\beta(t) = \begin{cases} \beta_{\text{low}}, & t \le t_1, \\ \beta_{\text{low}}(1-g) + \beta_{\text{high}} \cdot g, & t_1 \le t < t_2, \\ \beta_{\text{high}}, & t \ge t_2, \end{cases} \tag{21}$$

with the blending coefficient interpolated by a cubic Hermite spline,

$$g = 3\alpha^2 - 2\alpha^3, \quad \alpha = \frac{t - t_1}{t_2 - t_1}. \tag{22}$$

We set $\beta_{\text{high}} \approx 1.68\ \text{kcal}^{-1}\,\text{mol}$ to target the physically relevant thermodynamic regime, and $\beta_{\text{low}} \approx 0.029\ \text{kcal}^{-1}\,\text{mol}$ (corresponding to a high effective temperature) to effectively flatten the energy surface during early sampling. We use $t_1 = 100$ and $t_2 = 180$. During the initial phase ($t \le 100$), the small β_{low} allows the sampler to traverse a broad range of folds without energetic bias. Between steps 100 and 180, the schedule smoothly increases β toward β_{high}, progressively concentrating the Boltzmann weights on energetically plausible conformations at 300 K. Beyond step 180, the inverse temperature remains fixed for the remainder of the trajectory. For inference-time optimization, we apply the same annealing schedule over the outer optimization loops, where $t_1 = 2$, $t_2 = 10$, $\beta_{\text{low}} = 0.029$, and $\beta_{\text{high}} = 1.68$.

Exponential Moving Average (EMA) of energies. Force-field energies evaluated on partially denoised coordinates can be noisy, particularly at intermediate diffusion steps where local geometry is not yet fully resolved. Starting at step 160, we therefore replace the raw energy with an exponentially smoothed estimate:

$$E(\mathbf{X}^k_t)^{\text{EMA}} = \kappa \cdot E(\mathbf{X}^k_t) + (1-\kappa) \cdot E(\mathbf{X}^k_{t-1})^{\text{EMA}}, \tag{23}$$

Here, $\kappa = 0.3$. This stabilizes the energy gradient signal. A similar moving average mechanism is used for the inner loop of inference-time optimization.

Energy Biasing. To improve numerical stability and prevent the softmax from being dominated by a single low-energy outlier, we exploit the translation-invariance of the Boltzmann distribution and bias the distribution by subtracting the 10th percentile from the energies. Specifically, given ensemble energies $\{E_\phi(\mathbf{X}^1), \ldots, E_\phi(\mathbf{X}^n)\}$, we compute $E_{10} = \text{Percentile}_{10}\big(\{E_\phi(\mathbf{X}^k)\}_{k=1}^n\big)$ and define

$$\tilde{E}_\phi(\mathbf{X}^k) = \max\big(E_\phi(\mathbf{X}^k) - E_{10},\ 0\big).$$

The Boltzmann weights are computed as $w_k \propto \exp(-\beta\, \tilde{E}_\phi(\mathbf{X}^k))$. Conformers whose energy falls at or below the 10th percentile receive a shifted energy of zero and thus equal maximum weight; conformers above the threshold are penalized proportionally to their excess energy. This clamping prevents any single lowest-energy structure from dominating the reweighting, while the choice of the 10th percentile, rather than the minimum, provides additional robustness against occasional energy outliers.

Choice of Energy Model. The potential $E(\mathbf{X})$ is predicted by ProteinEBM (Roney et al., 2025), a differentiable energy-based model over protein conformations. Boltzmann reweighting requires committing to an explicit energy predictor that defines the thermodynamic prior. Any energy model is accurate only within limits imposed by its functional form, training data, and level of coarse-graining. We adopt ProteinEBM because its energy is a conservative potential obtained by parameterizing the diffusion score as $s_\theta = -\nabla_{\mathbf{X}} E_\theta$, which theoretically converges to $-\log p_{\text{data}}(\mathbf{X} \mid \mathbf{a})$ (Sohl-Dickstein et al., 2015). Moreover, ProteinEBM is finetuned on molecular dynamics trajectories at 300 K, aligning with the target thermodynamic regime for Boltzmann reweighting.

E.2 X-RAY ADDITIONAL DETAILS

E.2.1 FORWARD MODEL

The theoretical real-space electron density map $F_{\text{c}} : \mathbb{R}^3 \to \mathbb{R}$ corresponding to a protein structure with atomic coordinates $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots \mathbf{x}_n\}$ is defined at a Cartesian location $\boldsymbol{\xi} \in \mathbb{R}^3$ as

$$F_{\text{c}}(\boldsymbol{\xi}; \mathbf{X}) = \sum_{q=1}^{N_s} \sum_{i=1}^{m} \sum_{j=1}^{6} a_{i,j} \left(\frac{4\pi}{b_{i,j} + B_i^k} \right)^{1.5} \cdot \exp\left(-\frac{4\pi^2}{b_{i,j} + B_i^k} \|(\mathbf{R}_q \mathbf{x}_i^k + \mathbf{t}_q) - \boldsymbol{\xi}\|_2^2 \right)$$

where N_s is the number of crystallographic symmetry operations (Aroyo, 2016), $\mathbf{R}_q \in SO(3)$ and $\mathbf{t}_q \in \mathbb{R}^3$ are the rotation and translation associated with the symmetry operation q, respectively. The coefficients $a_{i,j}$ and $b_{i,j}$ are tabulated atomic form-factor parameters for each element (Prince, 2004), and B_k^i denotes the isotropic atomic displacement parameter (B-factor) for atom i.

E.2.2 RELAXATION

In order to make sure that the generated ensemble has no structural violations, we minimize its energy using an off-the-shelf harmonic force field. OpenMM's (Eastman et al., 2017) implementation of the AMBER99 (Wang et al., 2004). The energy is minimized for a maximum of 2000 iterations with an energy tolerance threshold of 2.39 kcal/mol and stiffness of 10.0 kcal / mol Å^2.

E.2.3 SELECTION ALGORITHM

After relaxation, we aim to report a non-redundant subset of samples $\boldsymbol{\mathcal{X}}_{\mathcal{I}} = \{\mathbf{X}^k : k \in \mathcal{I}\}$ that best explains experimental observation $\mathbf{y}$. To avoid overfitting to noise $\mathbf{y}$ and reduce redundancy, we adopt a matching pursuit strategy as done in Maddipatla et al. (2025a).

E.2.4 X-RAY METRICS

We use two metrics to evaluate how well our density-guided structures agree with experimental X-ray crystallographic data: cosine similarity and R values.

Cosine similarity. We report a single local score over the optimized residue range using cosine similarity between the observed (F_{o}) and calculated (F_{c}) electron densities. The metric is computed over voxels $\boldsymbol{\xi} \in \mathbb{R}^3$ within 2.5 Å of atoms in the selected residue range:

$$\text{Cosine Similarity} = \frac{\sum_{\boldsymbol{\xi}} F_{\text{o}}(\boldsymbol{\xi}) F_{\text{c}}(\boldsymbol{\xi})}{\sqrt{\sum_{\boldsymbol{\xi}} F_{\text{o}}^2(\boldsymbol{\xi})} \cdot \sqrt{\sum_{\boldsymbol{\xi}} F_{\text{c}}^2(\boldsymbol{\xi})}}$$

Values approaching 1 indicate strong global agreement between calculated and observed densities, reflecting a good overall fit to the experimental map.

R-factors. The crystallographic R-factor is a global score that quantifies the agreement between observed and calculated X-ray diffraction data, typically stored in MTZ files, by comparing observed and calculated structure factor amplitudes $|F_{\text{obs}}|$ and $|F_{\text{calc}}|$:

$$R_{\text{work}} = \frac{\sum_{\mathbf{h}} \big| |F_{\text{obs}}(\mathbf{h})| - |F_{\text{calc}}(\mathbf{h})| \big|}{\sum_{\mathbf{h}} |F_{\text{obs}}(\mathbf{h})|}$$

where $\mathbf{h} = (h, k, l)$ indexes reflections in reciprocal space. To mitigate overfitting, R_{free} is computed analogously over a held-out subset T of reflections:

$$R_{\text{free}} = \frac{\sum_{\mathbf{h} \in T} \big| |F_{\text{obs}}(\mathbf{h})| - |F_{\text{calc}}(\mathbf{h})| \big|}{\sum_{\mathbf{h} \in T} |F_{\text{obs}}(\mathbf{h})|}$$

Note that $|F_{\text{obs}}|$ and $|F_{\text{calc}}|$ are reciprocal-space amplitudes, whereas F_{o} and F_{c} denote real-space 3D electron density grids obtained via inverse Fourier transformation. We report both R_{work} and R_{free} values as computed by `REFMAC5` from the CCP4 suite (Murshudov et al., 2011; Agirre et al., 2023) after refinement.

E.2.5 Dataset & input preparation

In the X-ray crystallography workflow, the inputs are a PDB ID, chain identifier, and an amino acid subsequence; modeling is restricted to single protein chains. The corresponding PDB structure and MTZ file are retrieved from PDB-Redo (Joosten et al., 2014), and the target chain is extracted using Gemmi (v0.6.5) (Wojdyr, 2021). Only standard amino acid residues explicitly modeled in the PDB are retained, excluding waters, hydrogens, and non-standard residues. Selenium methionine (MSE) and S-hydroxycysteine (CSO) are converted to Methionine (MET) and Cystine (CYS), respectively. Alternate conformations (Rosenberg et al., 2024), if present, are split into separate PDBs and renumbered to one-based indexing for AF3 compatibility. Missing atoms are modeled using PDBFixer (v1.9.0) (Eastman, 2013) with OpenMM residue templates (Eastman et al., 2017), followed by AMBER99 relaxation (Wang et al., 2004) to resolve steric clashes; imputed atoms are assigned an isotropic B-factor of 100.00. An atom mask for the substructure conditioner is constructed from the provided amino acid subsequence to distinguish residues optimized by the density-based loss from those guided by the substructure conditioner.

E.3 ipTM additional details

E.3.1 Inter-chain Hydrogen Bond Analysis

Inter-chain hydrogen bonds were identified using geometric criteria adapted from established definitions (Baker & Hubbard, 1984; McDonald & Thornton, 1994). A hydrogen bond between a donor atom D and an acceptor atom A was considered only if the donor and acceptor belonged to different protein chains and satisfied the following conditions.

First, a distance criterion was applied, requiring the Euclidean distance between the donor and acceptor non-hydrogen atoms to satisfy $d(D, A) < d_{\text{max}} = 3.5$ Å, where $d(D, A)$ denotes the donor–acceptor distance.

When hydrogen atom coordinates were available, an additional angular constraint was enforced, requiring the donor–hydrogen–acceptor angle to satisfy $\theta_{D\text{-}H\cdots A} > \theta_{\text{min}} = 120°$. The angle $\theta_{D\text{-}H\cdots A}$ was computed at the hydrogen atom H as $\arccos\left(\frac{(\mathbf{r}_D - \mathbf{r}_H) \cdot (\mathbf{r}_A - \mathbf{r}_H)}{\|\mathbf{r}_D - \mathbf{r}_H\| \, \|\mathbf{r}_A - \mathbf{r}_H\|}\right)$, where $\mathbf{r}_D$, $\mathbf{r}_H$, and $\mathbf{r}_A$ denote the Cartesian coordinates of the donor, hydrogen, and acceptor atoms, respectively.

Hydrogen bond recovery was quantified as the fraction of the reference inter-chain hydrogen bonds that were correctly predicted, computed as $|H_{\text{ref}} \cap H_{\text{pred}}|/|H_{\text{ref}}|$, where H_{ref} and H_{pred} denote the sets of inter-chain hydrogen bonds identified in the reference and predicted structures, respectively.

E.3.2 ipTM Optimization Details

For ipTM-based inference-time optimization, we perform optimization of the AF3 embeddings to maximize the predicted interfacial confidence, measured using the combined objective $0.8 \times \text{ipTM} + 0.2 \times \text{pTM}$. Because AlphaFold3's ipTM predictor depends jointly on the Pairformer embeddings and the predicted complex structure, it is sensitive to perturbations in the embedding space. Optimization is carried out by iteratively perturbing the AF3 embeddings and resampling structures using the AlphaFold3 diffusion model.

Evaluation complexes. The evaluated complexes comprise three distinct classes:

- **NMR-determined complexes:** 2K8F, 2L14, 2LY4, and the crystallographic complex 1YCS. These systems exhibit low baseline confidence (both pTM and ipTM) under unguided AlphaFold3 predictions and represent challenging test cases for confidence-driven optimization.
- **Natural heterodimer with domain swapping:** 8Q70. This complex provides a structurally asymmetric interface and tests the sensitivity of optimization to non-symmetric binding geometries.
- **BindCraft-designed (Pacesa et al., 2025) complexes:** 9HAD, 9HAE, and 9HAF. These de novo designed protein–protein interfaces typically exhibit higher baseline confidence and provide a complementary regime for assessing ipTM optimization sensitivity.

E.4 Inference-time Optimization additional details

E.4.1 Model and hardware details

For all experiments, we used Protenix (v0.2.0) (Team et al., 2025), an open-source PyTorch reimplementation of AlphaFold3, except for the ipTM-based guidance experiments described in Section C.3, where we relied on the official JAX implementation (Abramson et al., 2024). All computations were carried out on NVIDIA H100 and L40 GPUs running Debian GNU/Linux 12.

Multiple sequence alignments (MSAs) were obtained using a wrapper around the ColabFold MMseqs2 API (Mirdita et al., 2022). This wrapper submits query sequences to a remote MMseqs2 server via HTTP POST requests, polls for job completion, and downloads and extracts the resulting alignments. The MSAs are provided in A3M format, which is directly compatible with AF3. To increase diversity in the IT-optimization initialization, we subsample the MSAs using AF-Cluster (Wayment-Steele et al., 2024). When the number of clusters produced by AF-Cluster is smaller than the ensemble size (as in the NMR experiments), we duplicate the corresponding embeddings across the batch.

E.4.2 Additional loss function

During guidance and IT-optimization, we incorporate additional loss terms into the log-likelihood to ensure that the generated structures remain physically valid.

Embedding Prior. As mentioned in Equations 2 and 5, to prevent the optimization from drifting toward degenerate embeddings that lie outside the manifold induced by the input sequence and its evolutionary context, we regularize the ensemble of embeddings $\boldsymbol{\mathcal{Z}} = \{\mathbf{Z}^k\}_{k=1}^n$ to stay close to their initialization $\boldsymbol{\mathcal{Z}}_0 = \{\mathbf{Z}_0^k\}_{k=1}^n$. Specifically, we introduce the following regularization term:

$$\log p(\boldsymbol{\mathcal{Z}}|\mathbf{a};\mathbf{w}) = -\lambda_{\mathrm{p}} \sum_{k=1}^{n} w^k \|\mathbf{Z}^k - \mathbf{Z}_0^k\|_2^2$$

where $\mathbf{w} = \{w^1, \ldots, w^n\}$ denotes a set of non-negative weights that can optionally be instantiated as Boltzmann weights.

Validity Likelihood. To discourage ensembles with unrealistic geometries, such as elongated covalent bonds or steric clashes, we introduce a validity log-likelihood regularizer, analogous to the violation loss used in AF2 (Jumper et al., 2021). Let $\boldsymbol{\mathcal{X}} = \{\mathbf{X}^k\}_{k=1}^n$ denote an ensemble of n protein conformations, where each structure $\mathbf{X}^k \in \mathbb{R}^{m\times 3}$ specifies the Cartesian coordinates of m atoms. We define a binary bond matrix $\mathbf{B} \in \{0,1\}^{m\times m}$ such that $B_{ij} = 1$ if atoms i and j are covalently bonded, and 0 otherwise. The bond length loss for $\mathbf{X}^k$ over bonded atom pairs ($B_{ij} = 1$) is given as,

$$\mathcal{L}_{\mathrm{bond}}(\mathbf{X}^k;\mathbf{a}) = \sum_{i=1}^{m} \sum_{j=i+1}^{m} B_{ij} \cdot (\max(0, |d_{ij}^k - d_{ij}^{\mathrm{ideal}}| - \delta_{\mathrm{bond}}))^2$$

where d_{ij}^{ideal} is the ideal bond length approximated as the sum of covalent radii of atoms i and j, $d_{ij}^k = \|\mathbf{x}_i^k - \mathbf{x}_j^k\|_2$ is the between atoms i and j in conformation k, and $\delta_{\mathrm{bond}} = 0.2\mathring{A}$ is a tolerance margin. Steric clashes between non-bonded atom pairs (both intra- and inter-residue) are penalized using a soft collision loss, defined as the maximum violation over neighboring atoms for each atom:

$$\mathcal{L}_{\mathrm{collision}}(\mathbf{X}^k;\mathbf{a}) = \sum_{i=1}^{m} \max_{j\neq i, B_{ij}=0} (\max(0, (d_{ij}^{\mathrm{ideal}} + p^{\mathrm{collision}}) - d_{ij}^k))$$

Here, $p^{\text{collision}} = 0.4\mathring{A}$ is a padding distance to prevent over-penalization of near-contact atoms. We additionally penalize bond-angle violations for triplets of bonded atoms. The bond-angle loss is defined as

$$\mathcal{L}_{\text{angle}}(\mathbf{X}^k; \mathbf{a}) = \sum_{j=1}^{m} \sum_{i \neq j} \sum_{k > i} B_{ij} B_{jk} \cdot (\max(0, |\theta_{ijk} - \theta_{ijk}^{\text{ideal}}| - \delta_{\text{angle}}))$$

where θ_{ijk} (in degrees) is calculated using the dot product of the bond vectors from central atom j to atoms i and k, $\theta_{ijk}^{\text{ideal}}$ (in degrees) is retrieved from the Valence Shell Electron Pair Repulsion (VSEPR) theory (Gillespie, 1992), and $\delta_{\text{angle}} = 12°$ is a tolerance margin. The resulting validity log-likelihood of the ensemble is given by

$$\log(\mathbf{B} \mid \boldsymbol{\mathcal{X}}, \mathbf{a}) = -\sum_{k=1}^{n} \lambda_{\text{bond}} \mathcal{L}_{\text{bond}}(\mathbf{X}^k; \mathbf{a}) + \lambda_{\text{collision}} \mathcal{L}_{\text{collision}}(\mathbf{X}^k; \mathbf{a}) + \lambda_{angle} \mathcal{L}(\mathbf{X}^k; \mathbf{a})$$

where $\lambda_{\text{bond}}, \lambda_{\text{collision}}$, and λ_{angle} are scaling factors that control the contribution of bond, collision, and bond angle terms. For ensembles guided using crystallographic density maps and ipTM, we use $\lambda_{\text{bond}} = \lambda_{\text{collision}} = \lambda_{\text{angle}} = 0.075$. For NOE-guided ensembles, we use $\lambda_{\text{bond}} = \lambda_{\text{collision}} = \lambda_{\text{angle}} = 0.25$.

Substructure Conditioner. For case studies involving altlocs in crystallographic targets, we optimize a specified subset of the protein while stabilizing the remaining regions by anchoring them to a set of reference atomic coordinates during the diffusion process. This conditioning strategy is analogous to the `SubstructureConditioner` used in Chroma (Ingraham et al., 2023). This anchor is *not* applied for peptide systems.

Let $Y = \{\mathbf{y}_i : i \in A\}$ denote a set of reference atom locations for atom indices $A \subseteq \{1 \dots m\}$ where m is the number of atoms in conformer $\mathbf{X}^k$. The log-likelihood is a quadratic penalty on the deviation from reference atom locations,

$$\log p(Y \mid \boldsymbol{\mathcal{X}}, \mathbf{a}) = -\lambda_{\text{sub}} \frac{1}{n} \sum_{k=1}^{n} \sum_{i \in A} \|\mathbf{x}_i^k - \mathbf{y}_i\|_2^2$$

Prior to evaluating this term, all ensemble members are rigidly aligned to the reference coordinates (restricted to the atom indices in A) using the Kabsch algorithm (Kabsch, 1976). For crystallographic refinement, we set $\lambda_{\text{sub}} = 0.1$, whereas this term is disabled for NOE-guided ensembles by setting $\lambda_{\text{sub}} = 0.0$.

E.4.3 Training Details & Hyperparameters

X-ray and NMR settings We optimize the conditioning variables $\boldsymbol{\mathcal{Z}}$ using Adam (Kingma & Ba, 2014) with the following default settings:

- Learning rate (η_z): 0.05
- Gradient clipping (max norm): 0.01
- Number of outer optimization loops K: 20
- Ensemble size n: 16
- Prior weight (λ_{p}): 10^{-4}

ipTM-specific settings. For ipTM-based experiments, we use a smaller ensemble and shorter optimization schedule:

- Learning rate (η_z): 0.1
- Gradient clipping (max norm): 1.0
- Outer optimization iterations K: 10
- Ensemble size n: 5
- Prior weight (λ_{p}): 10^{-4}

- Stopping criterion: maximum relative perturbation budget $|\mathcal{Z}_{\text{opt}} - \mathcal{Z}|/|\mathcal{Z}|$ on AF3 embeddings $\mathcal{Z}$.

For all experiments, we optimize $\mathcal{Z}$ up to inner diffusion iteration $t = 160$ at each outer iteration. We apply early stopping because, in the diffusion schedule of Karras et al. (2022), the final denoising steps become effectively deterministic refinement and no longer inject stochastic noise. Since our method relies on noise-driven sampling, further optimization provides limited benefit. After optimization, the learned embeddings are used as conditioning variables, and we run coordinate-space guidance as in Maddipatla et al. (2025a) using similar numerical tricks and hyperparameters.

F Related Work

Experimental guidance. Recent methods treat AF3 as a sequence-conditioned prior and incorporate experimental likelihoods from cryo-EM, NMR, and X-ray crystallography into the reverse SDE (Maddipatla et al., 2024; Raghu et al., 2025; Maddipatla et al., 2025b;a). Our method directly optimizes the Pairformer embeddings, thereby avoiding limitations imposed by the diffusion schedule and finite denoising steps. This approach further relates to enhanced-sampling methods like metadynamics (Barducci et al., 2011), as we utilize differentiable force fields to recover experimentally faithful, Boltzmann-distributed ensembles. Our Boltzmann reweighting draws on the broader paradigm of sampling from Boltzmann densities with learned generative models (Noé et al., 2019; Akhound-Sadegh et al., 2024), though it requires no additional training and instead performs post-hoc reweighting of AF3 samples at inference time.

MSA manipulation. Prior works have shown that manipulating the MSA can recover distinct conformational states (Wayment-Steele et al., 2024) or steer predictions toward desired structural targets (Bryant & Noé, 2024; Fadini et al., 2025) by optimizing AlphaFold2's MSA profile. However, these methods require expensive backpropagation through the Evoformer and assume a single structural target. We avoid these bottlenecks by optimizing a batch of embeddings directly. This allows for efficient inference-time steering and enables the generation of ensembles that capture the experimental heterogeneity inherent to proteins.

ipTM benchmarking and confidence. The interface predicted TM-score (ipTM), introduced with AF-Multimer, is a confidence metric designed to assess the accuracy of predicted protein–protein interfaces by focusing on inter-chain geometry (Evans et al., 2022). Benchmark studies show that ipTM correlates well with interface accuracy metrics such as DockQ across diverse heteromeric datasets (Zhai et al., 2025). However, ipTM does not capture all structural failure modes, particularly in flexible or disordered regions (Wee & Wei, 2024), motivating the development of refined interface confidence metrics such as actual interface pTM (ActifpTM) and interaction prediction score from aligned errors (ipSAE) (Varga et al., 2025; Dunbrack Jr, 2025).

G FIGURES

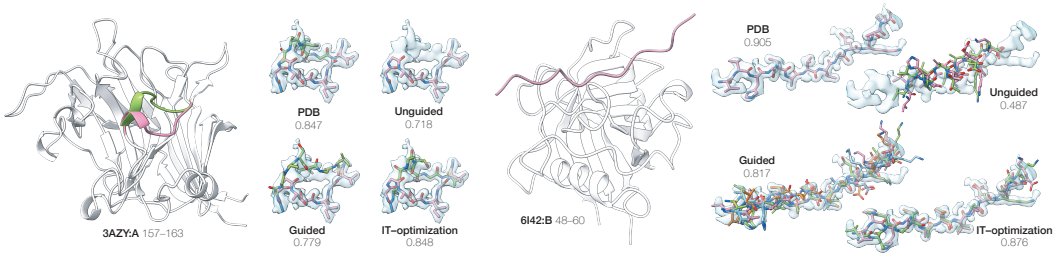

Figure 3: **Inference-time (IT) optimization improves structural accuracy over guided and unguided baselines.** (Top) `3AZY:A` (1.45Å) exhibits bimodal distribution at residues 157-163. Unguided AF3 predicts a single mode, while guidance produces a bimodal ensemble with a poorly fit backbone at one of the modes. IT-optimization recovers both modes with accurate density fit, matching the PDB. (Bottom) For `6I42:B` (1.38Å), AF3 mispredicts the bound 13-residue peptide. Guidance improves backbone placement but poorly predicts side-chains, whereas IT-optimization yields accurate backbone and side-chain agreement. Numbers beneath each ensemble indicate cosine similarity to F_{o}.

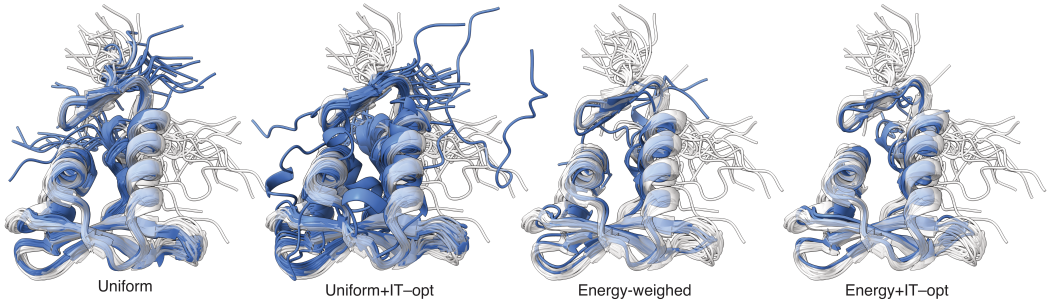

Figure 4: **Conformational ensembles for `2K0M` from Guidance (uniform), Inference-time Optimization (uniform), Energy-weighted Guidance, and Energy-weighted Inference-time Optimization (blue) overlaid with the NMR ensemble derived from the same NOESY data (white).** Uniform ensembles weight all sampled conformations equally, whereas energy-weighted ensembles display only conformations with non-negligible Boltzmann weight.

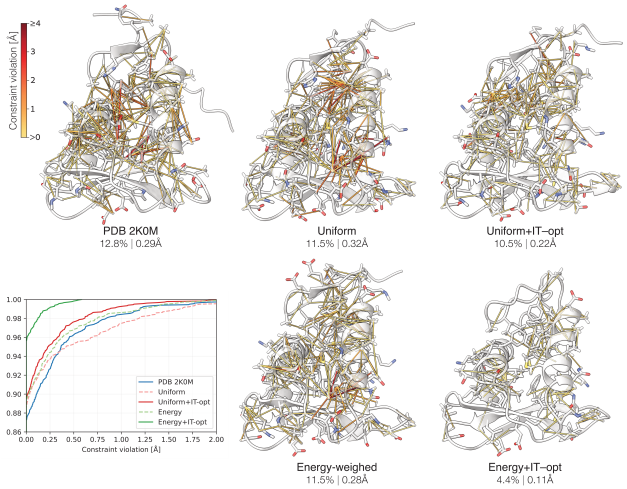

Figure 5: **Inference-time optimization, with and without energy weighting, reduces NOE restraint violations.** NOE constraint violations in `2K0M` and ensembles from PDB, Guidance (Uniform), Inference time Optimization (Uniform + IT-opt), Energy-weighted Guidance, and Energy-weighted IT-Opt. Violated constraints are shown as lines and colored by violation magnitude. The percentage of violated constraints and their median violation distance are reported below each structure. *Bottom left*: cumulative distribution of violation magnitudes across all five generated ensembles.

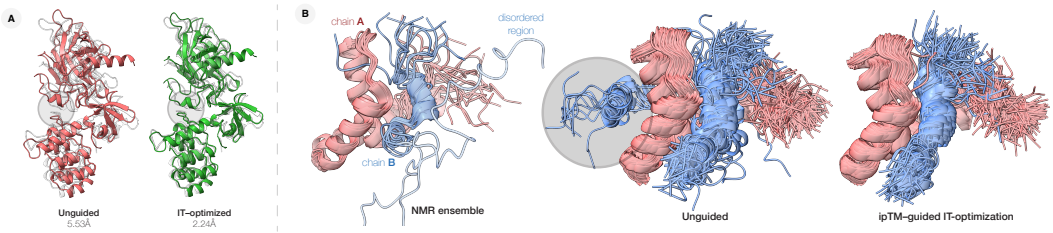

Figure 6: **Selected examples where ipTM-based inference-time optimization improves agreement with experiment.** (A) Complex `1YCS`. The experimental crystal structure is shown in white, the unguided AlphaFold3 prediction in red, and the ipTM-optimized prediction in green. In this case, the unguided AF3 prediction fails to form a specific helix-helix contact (left), whereas ipTM-based IT-Opt recovers this interaction (right), resulting in closer agreement with the experimental structure. (B) Complex `2YL4`, comprising a globular target (chain A) and a peptide derived from `p53`. The experimental NMR ensemble (left) supports an interior binding mode. Across 100 unguided AF3 samples (middle), two binding modes are observed: an exterior conformation associated with higher ipTM (0.55) and an interior conformation with lower ipTM (0.22). In this example, ipTM-based IT-Opt (right) consistently samples the interior binding mode across 100 random initializations, producing structures that better align with the experimental ensemble.

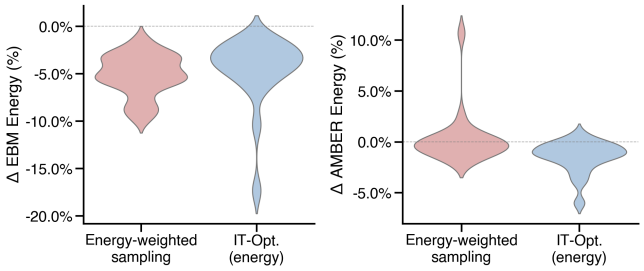

Figure 7: **Energy-weighted inference improves thermodynamic stability of the ensemble.** Violin plots show relative energy changes with respect to a uniformly weighted baseline (dashed line) for energy-weighted sampling and energy-weighted IT-optimization. Left: Energies derived using ProteinEBM (Roney et al., 2025). Right: AMBER99 (Wang et al., 2004). Negative values indicate lower energies and, therefore, more thermodynamically stable conformations. Across both energy functions, energy-weighted IT-optimization shifts the distribution toward lower energies and reduces high-energy outliers compared to energy-weighted sampling alone.

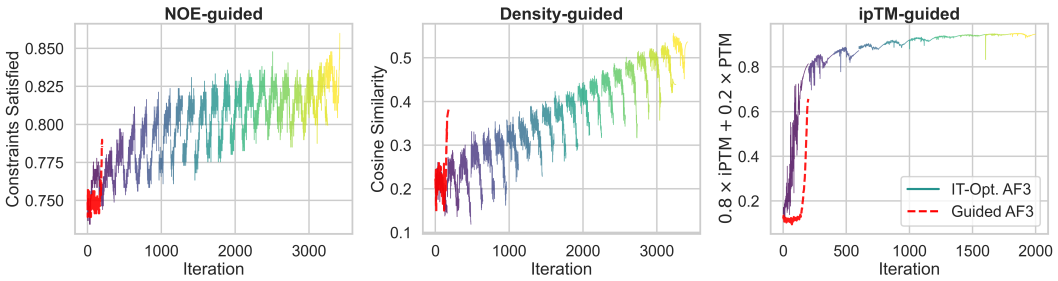

Figure 8: **Convergence behavior of guided versus inference-time optimized (IT-Opt) AlphaFold3 across experimental modalities.** Convergence curves are shown for NOE-based IT-Opt on an NMR structure (`1U0P`), density-based IT-Opt on an X-ray structure (`6I42`), and ipTM-based IT-Opt on a protein complex (`2L14`). In each panel, dashed red curves correspond to standard guided AlphaFold3, where guidance is applied during sampling and with a fixed budget of 200 diffusion iterations. Solid curves show inference-time optimization, where embedding updates are applied at each diffusion step (with early stopping at $t = 160$ in inner loop) and reused across successive diffusion rounds. In the final outer loop, (the last 200 iterations of IT-Opt), the optimized embeddings initialize coordinate-space guidance, yielding the late-stage performance improvement characteristic of traditional guidance.

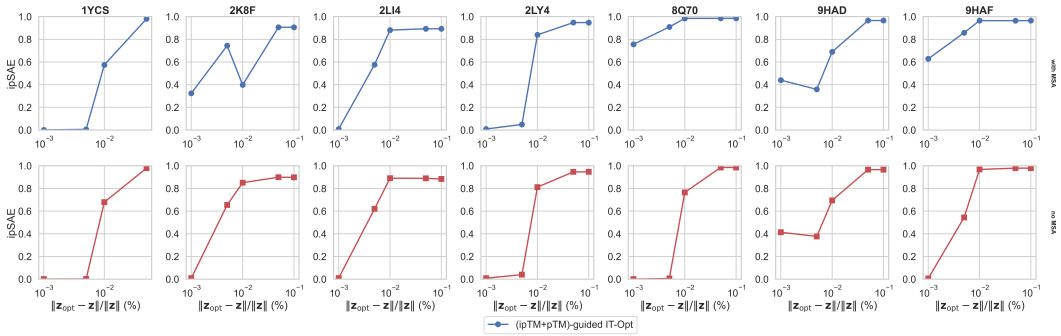

Figure 9: **Effect of inference-time optimization on Interaction Prediction Score from Aligned Errors (ipSAE) across biomolecular complexes.** The figure depicts the ipSAE values obtained as a result of (ipTM + pTM)-based inference-time optimization across seven biomolecular complexes, evaluated with and without MSA input. Curves show optimized solutions obtained under increasing relative perturbation budgets $\|\mathcal{Z}_{\text{opt}} - \mathcal{Z}|/\|\mathcal{Z}\|$ in AF3 embedding space. Across targets, increases in the optimized (ipTM + pTM) objective are accompanied by consistent increases in ipSAE, with ipSAE frequently approaching near-saturation values at moderate perturbation magnitudes. Similar trends are observed in both MSA-rich and MSA-poor settings, although sensitivity to perturbation varies across targets. These results indicate that, like ipTM, ipSAE is highly sensitive to small embedding-space perturbations and can be driven to high values through inference-time optimization.

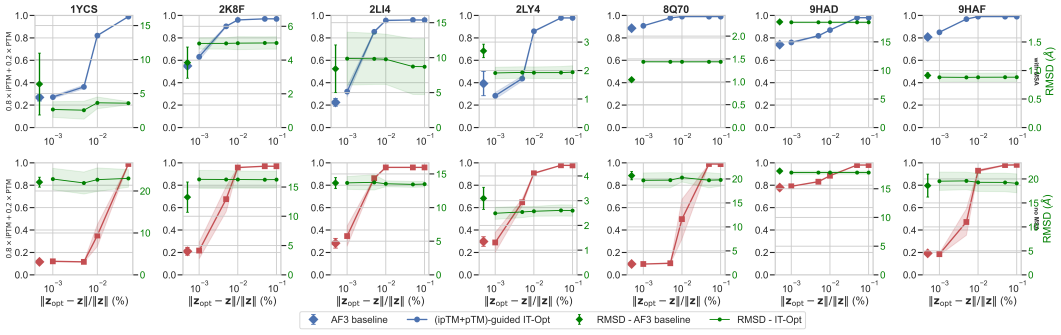

Figure 10: **Effect of inference-time optimization on ipTM across biomolecular complexes.** The figure shows the impact of inference-time optimization targeting $0.8 \times \text{ipTM} + 0.2 \times \text{pTM}$ across seven biomolecular complexes, evaluated with and without MSA input (top and bottom rows, respectively). Diamonds denote baseline AF3 predictions at zero perturbation, while curves show optimized solutions obtained under increasing relative perturbation budgets $\|\mathcal{Z}_{\text{opt}} - \mathcal{Z}\|/\|\mathcal{Z}\|$ in AF3 embedding space; shaded regions indicate variability across samples drawn given the optimized embedding. The left axis reports confidence (ipTM+pTM-based objective), while the right axis reports backbone RMSD to experiment. Across several targets, we observe that relatively small embedding perturbations can increase confidence scores. However, these increases are not uniformly accompanied by improvements in structural agreement with experiment, particularly in the no-MSA setting. These results suggest that while embedding-space optimization can effectively modulate confidence metrics, confidence improvements should be interpreted cautiously as indicators of experimental accuracy.

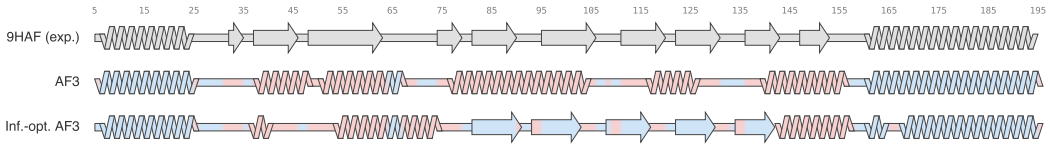

Figure 11: **Secondary-structure comparison for complex `9HAF` under ipTM-based inference-time optimization without MSA.** Schematic comparison of secondary-structure assignments along chain A of complex `9HAF`. The top row shows the experimentally determined secondary structure, while the middle and bottom rows correspond to the unguided AF3 prediction without MSA input and the IT-optimized AF3 prediction targeting the combined ipTM+pTM objective, respectively. Helices and β-strands are shown as ribbons and arrows, with residue-wise agreement (blue) and disagreement (red) relative to the experimental annotation indicated by color. In the unguided prediction, the structure is dominated by helical assignments and exhibits low secondary-structure agreement with the experimental annotation (41.9%). Following ipTM-based IT-optimization, large regions of misassigned helices are corrected, and extended β-strand regions are recovered, increasing secondary-structure agreement to 59.2%.

H TABLES

PDB ID	Seq Length	# NOEs
2K53	70	932
6SVC	35	1147
1U0P	27	198
1DEC	39	493
6SOW	58	1120
2KPB	26	267
5L82	37	485
2MRW	40	159
1D3Z	76	2727
2B5B	36	96
1RKL	36	218
2KNS	33	233
2RN7	108	353
2K0M	104	1834
2HEQ	84	485
2K57	55	982
1PQX	91	1490
1YEZ	68	1237
2KCD	120	1075
2HN8	38	719

Table 1: **NMR protein structures used for evaluation in our experiments**. For each structure, we report the amino acid sequence length and the number of experimentally derived Nuclear Overhauser Effect (NOE) distance restraints.

PDB ID	Residue Range	Region Sequence	Resolution (Å)
5D7F:P	1–10	GPWDSVARVL	1.30
6I42:B	48–60	VVHGVATVAEKTK	1.38
7ABT:B	1–8	PRPRPRPR	1.31
2DF6:C	180–197	PPVIAPRPEHTKSIYTRS	1.30
2DF6:D	180–191	PPVIAPRPEHTK	1.30
1DDV:B	1001–1006	TPPSPF	1.90
1CKB:B	1–8	PPPVPPRR	1.90
1CJR:A	1–15	KETAAAKFERQHMDS	2.30
5G51:A	291–295	SASDQ	1.45
6QQF:A	69–74	TPGSRN	1.95
3AZY:A	157–163	GGASIGV	1.65
2YNT:A	183–185	GNG	1.60
3V3S:B	246–249	AQER	1.90
4NPU:B	134–136	EEI	1.50
3QDA:A	55–63	SFSKDWSFY	1.57
4RMW:A	55–63	SFSKDWSFY	1.40

Table 2: **X-ray crystallography protein used for evaluation in our experiments**. For each structure, we report the chain ID, the residue range, the region sub-sequence, and the resolution of the density map.

PDB ID	PDB	AlphaFold3	AlphaFlow	Guidance	IT-Opt
2K53	29.6	41.1	40.5	28.1	**20.6**
6SVC	51.8	50.6	49.7	51.4	**44.4**
1U0P	57.5	60.4	53.1	50.2	**44.4**
1DEC	12.4	32.1	45.6	19.9	**14.4**
6SOW	32.0	43.3	43.7	30.0	**28.6**
2KPB	27.2	62.3	57.7	45.3	**40.7**
5L82	31.0	58.8	57.0	49.4	**44.0**
2MRW	23.3	38.4	33.9	**22.6**	24.5
1D3Z	24.9	23.8	24.0	13.5	**11.3**
2B5B	30.6	29.7	40.7	29.6	**28.7**
1RKL	53.0	57.2	56.7	54.0	**48.8**
2KNS	58.0	65.5	58.0	50.4	**49.6**
2RN7	11.4	9.4	11.4	5.3	**4.7**
2K0M	12.8	17.9	19.6	11.5	**10.5**
2HEQ	18.2	17.7	31.2	10.0	**9.5**
2K57	12.5	17.8	19.2	8.2	**7.9**
1PQX	1.2	5.1	5.2	3.3	**2.2**
1YEZ	12.7	13.8	17.0	7.5	**7.0**
2KCD	23.9	21.2	39.7	13.8	**10.5**
2HN8	40.5	54.0	49.6	41.6	**34.2**

Table 3: **Quantitative evaluation of NOE restraint violation percentage (%) for structures from Table 1**. The violation percentage (↓) denotes the fraction of NOE distance restraints that are not satisfied. Here, *Guidance* refers to ensembles derived using NOE-guided AlphaFold3, whereas *IT-Opt* refers to ensembles derived using NOE-based inference-time optimization. Entries highlighted in **green** correspond to ensembles that outperform all baseline methods, excluding the PDB, while entries highlighted in **blue** outperform all baselines, including the PDB.

PDB ID	PDB	AlphaFold3	AlphaFlow	Guidance	IT-Opt
2K53	0.35	0.53	0.47	0.20	**0.13**
6SVC	0.94	0.44	0.41	0.38	**0.33**
1U0P	0.13	0.35	0.34	0.22	**0.20**
1DEC	0.07	0.75	2.84	0.36	**0.23**
6SOW	0.51	0.69	0.71	0.42	**0.36**
2KPB	0.52	1.35	1.28	1.24	**0.63**
5L82	0.71	0.95	1.19	0.86	**0.62**
2MRW	0.10	0.14	0.12	0.30	**0.06**
1D3Z	0.39	0.34	0.36	0.18	**0.14**
2B5B	0.39	5.80	4.87	3.58	**0.73**
1RKL	0.36	0.63	0.59	0.42	**0.33**
2KNS	0.36	0.48	0.59	0.39	**0.30**
2RN7	0.44	0.28	0.33	0.08	**0.05**
2K0M	0.29	0.48	0.61	0.32	**0.22**
2HEQ	0.36	0.52	0.62	**0.12**	0.31
2K57	0.50	0.52	0.51	0.19	**0.06**
1PQX	0.05	0.71	0.82	0.43	**0.08**
1YEZ	0.46	0.41	0.46	0.24	**0.12**
2KCD	0.55	0.46	1.38	0.25	**0.09**
2HN8	0.08	0.34	0.32	0.24	**0.15**

Table 4: **Quantitative evaluation of NOE restraint distance violation magnitude (Å) for structures from Table 1**. The violation distance (↓) reports the median distance of unsatisfied NOE restraints. Here, *Guidance* refers to ensembles derived using NOE-guided AlphaFold3, whereas *IT-Opt* refers to ensembles derived using NOE-based inference-time optimization. Entries highlighted in **green** correspond to ensembles that outperform all baseline methods, excluding the PDB, while entries highlighted in **blue** outperform all baselines, including the PDB.

PDB ID	PDB	Guidance	Guidance (Energy)	IT-Opt	IT-Opt (Energy)
2K53	29.6	28.1	32.7	20.6	**19.0**
6SVC	51.8	51.4	42.3	44.4	**38.3**
1U0P	57.5	50.2	49.3	44.4	**41.1**
1DEC	12.4	19.9	18.7	14.4	**11.6**
6SOW	32.0	30.0	30.2	28.6	**20.0**
2KPB	27.2	45.3	35.7	40.7	**34.8**
5L82	31.0	49.4	49.2	44.0	**34.4**
2MRW	23.3	22.6	**17.6**	24.5	18.2
1D3Z	24.9	13.5	17.3	**11.3**	16.4
2B5B	30.6	29.6	30.6	28.7	**27.8**
1RKL	53.0	54.0	54.4	48.8	**42.3**
2KNS	58.0	50.4	47.1	49.6	**39.1**
2RN7	11.4	5.3	4.4	4.7	**4.4**
2K0M	12.8	11.5	10.5	10.5	**4.4**
2HEQ	18.2	10.0	6.1	9.5	**4.3**
2K57	12.5	8.2	11.1	7.9	**5.9**
1PQX	1.2	3.3	1.1	2.2	**1.2**
1YEZ	12.7	7.4	12.6	7.0	**4.5**
2KCD	23.9	13.8	12.8	10.5	**6.9**
2HN8	40.5	41.6	43.3	34.2	**30.3**

Table 5: **Quantitative evaluation of NOE restraint violation percentage (%) for structures from Table 1**. Here, *Guidance (Energy)* refers to ensembles derived using NOE-guided AlphaFold3 with Boltzmann-weighted ensemble statistics at 300 K, whereas *IT-Opt (Energy)* denotes ensembles obtained via NOE-based inference-time optimization with Boltzmann weighting at 300 K. Entries highlighted in **green** correspond to ensembles that outperform all baseline methods, excluding the PDB, while entries highlighted in **blue** outperform all baselines, including the PDB.

PDB ID	PDB	Guidance	Guidance (Energy)	IT-Opt	IT-Opt (Energy)
2K53	0.35	0.20	0.29	0.13	**0.18**
6SVC	0.94	0.38	0.38	**0.38**	0.42
1U0P	0.13	0.22	0.22	0.20	**0.15**
1DEC	0.07	0.36	0.58	0.23	**0.16**
6SOW	0.51	0.42	0.43	0.36	**0.27**
2KPB	0.52	1.24	0.56	0.63	**0.44**
5L82	0.71	0.86	1.07	0.62	**0.42**
2MRW	0.10	0.30	0.16	0.06	**0.05**
1D3Z	0.39	0.18	0.25	**0.14**	0.25
2B5B	0.39	3.58	4.12	**0.73**	2.04
1RKL	0.36	0.42	0.44	0.33	**0.23**
2KNS	0.36	0.39	0.43	0.30	**0.28**
2RN7	0.44	0.08	0.26	**0.05**	0.07
2K0M	0.29	0.32	0.28	0.22	**0.11**
2HEQ	0.36	0.12	0.28	0.31	**0.09**
2K57	0.50	0.19	0.29	0.06	**0.09**
1PQX	0.05	0.43	0.08	0.08	**0.05**
1YEZ	0.46	0.24	0.41	0.12	**0.10**
2KCD	0.55	0.25	0.46	0.09	**0.09**
2HN8	0.08	0.24	0.29	0.15	**0.14**

Table 6: **Quantitative evaluation of NOE restraint distance violation magnitude (Å) for structures from Table 1**. Here, *Guidance (Energy)* refers to ensembles derived using NOE-guided AlphaFold3 with Boltzmann-weighted ensemble statistics at 300 K, whereas *IT-Opt (Energy)* denotes ensembles obtained via NOE-based inference-time optimization with Boltzmann weighting at 300 K. Entries highlighted in **green** correspond to ensembles that outperform all baseline methods, excluding the PDB, while entries highlighted in **blue** outperform all baselines, including the PDB.

PDB ID	PDB	AlphaFold3	AlphaFlow	Guidance	IT-Opt
5D7F:P	0.696	0.435	0.523	0.638	**0.656**
6I42:B	0.905	0.487	0.664	0.817	**0.876**
7ABT:B	0.883	0.515	0.519	0.785	**0.812**
2DF6:C	0.852	0.419	0.679	0.825	**0.826**
2DF6:D	0.853	0.388	0.559	0.832	**0.878**
1DDV:B	0.939	0.561	0.853	0.893	**0.901**
1CKB:B	0.868	0.808	0.796	0.882	**0.891**
1CJR:A	0.928	0.672	0.861	0.893	**0.921**
5G51:A	0.884	0.598	0.502	0.877	**0.898**
6QQF:A	0.878	0.807	0.829	0.868	**0.869**
3AZY:A	0.847	0.717	0.721	0.780	**0.848**
2YNT:A	0.885	0.822	0.871	0.895	**0.905**
3V3S:B	0.767	0.645	0.731	0.755	**0.808**
4NPU:B	0.832	0.791	0.785	0.804	**0.828**
3QDA:A	0.893	0.594	0.584	0.889	**0.891**
4RMW:A	0.904	0.660	0.687	0.899	**0.900**

Table 7: **Quantitative evaluation of cosine similarity (↑) for structures from Table 2**. Here, *Guidance* refers to ensembles derived using electron density guided AlphaFold3 (Maddipatla et al., 2025b), whereas *IT-Opt* refers to ensembles derived using electron density-based inference-time optimization. Entries highlighted in **green** correspond to ensembles that outperform all baseline methods, excluding the PDB, while entries highlighted in **blue** outperform all baselines, including the PDB.

PDB ID	PDB	AlphaFold3	AlphaFlow	Guidance	IT-Opt
5D7F:P	0.211	0.237	0.223	0.220	**0.216**
6I42:B	0.169	0.213	0.222	0.195	**0.175**
7ABT:B	0.169	0.208	0.216	0.200	**0.179**
2DF6:C	0.174	0.254	0.207	0.188	**0.183**
2DF6:D	0.174	0.211	0.253	0.179	**0.177**
1DDV:B	0.210	0.263	0.243	0.233	**0.227**
1CKB:B	0.1767	0.196	0.204	0.186	**0.181**
1CJR:A	0.201	0.279	0.250	0.227	**0.208**
5G51:A	0.191	0.205	0.215	0.192	**0.191**
6QQF:A	0.159	0.163	0.163	0.161	**0.161**
3AZY:A	0.173	0.174	0.174	0.174	**0.173**
2YNT:A	0.158	0.160	0.159	0.158	**0.158**
3V3S:B	0.163	0.164	0.164	**0.162**	0.163
4NPU:B	0.190	0.194	0.194	0.192	**0.191**
3QDA:A	0.172	0.217	0.222	0.178	**0.174**
4RMW:A	0.187	0.227	0.228	0.191	**0.190**

Table 8: **Quantitative evaluation of R_{work} (↓) for structures from Table 2**. Here, *Guidance* refers to ensembles derived using Electron density guided AlphaFold3 (Maddipatla et al., 2025b), whereas *IT-Opt* refers to ensembles derived using electron density-based inference-time optimization. Entries highlighted in **green** correspond to ensembles that outperform all baseline methods, excluding the PDB, while entries highlighted in **blue** outperform all baselines, including the PDB.

PDB ID	PDB	AlphaFold3	AlphaFlow	Guidance	IT-Opt
5D7F:P	0.230	0.234	0.236	0.238	**0.233**
6I42:B	0.183	0.231	0.237	0.212	**0.192**
7ABT:B	0.193	0.228	0.234	0.219	**0.205**
2DF6:C	0.192	0.270	0.229	0.207	**0.204**
2DF6:D	0.192	0.231	0.269	0.198	**0.197**
1DDV:B	0.257	0.289	0.200	0.274	**0.268**
1CKB:B	0.234	0.259	0.257	0.243	**0.235**
1CJR:A	0.2634	0.346	0.319	0.289	**0.270**
5G51:A	0.211	0.223	0.231	0.215	**0.211**
6QQF:A	0.208	0.211	0.211	**0.205**	0.206
3AZY:A	0.194	0.195	0.196	0.195	**0.194**
2YNT:A	0.200	**0.191**	0.200	0.200	0.200
3V3S:B	0.202	0.202	**0.201**	0.202	0.202
4NPU:B	0.213	0.217	0.217	0.215	**0.214**
3QDA:A	0.207	0.256	0.261	0.214	**0.213**
4RMW:A	0.216	0.259	0.262	**0.218**	0.221

Table 9: **Quantitative evaluation of R_{free} (↓), for structures from Table 2**. Here, *Guidance* refers to ensembles derived using Electron density guided AlphaFold3 (Maddipatla et al., 2025b), whereas *IT-Opt* refers to ensembles derived using electron density-based inference-time optimization. Entries highlighted in **green** correspond to ensembles that outperform all baseline methods, excluding the PDB, while entries highlighted in **blue** outperform all baselines, including the PDB.

PDB	Method	Cosine Similarity (↑)	$\mathbf{R}_{\text{work}}$ (↓)	$\mathbf{R}_{\text{free}}$ (↓)
3AZY:A	Guidance	0.776 ± 0.0461	0.174 ± 0.0004	0.195 ± 0.0001
	IT-Opt	$\mathbf{0.842 \pm 0.0037}$	$\mathbf{0.173 \pm 0.0001}$	$\mathbf{0.194 \pm 0.0002}$
4NPU:B	Guidance	0.812 ± 0.0415	0.191 ± 0.0003	0.214 ± 0.0009
	IT-Opt	$\mathbf{0.822 \pm 0.0047}$	$\mathbf{0.191 \pm 0.0001}$	$\mathbf{0.213 \pm 0.0006}$
3QDA:A	Guidance	0.883 ± 0.0117	0.177 ± 0.0025	0.210 ± 0.0048
	IT-Opt	$\mathbf{0.894 \pm 0.0021}$	$\mathbf{0.174 \pm 0.0005}$	$\mathbf{0.209 \pm 0.0028}$
4RMW:A	Guidance	0.896 ± 0.0132	0.189 ± 0.0017	0.219 ± 0.0027
	IT-Opt	$\mathbf{0.905 \pm 0.0057}$	$\mathbf{0.188 \pm 0.0013}$	$\mathbf{0.217 \pm 0.0031}$

Table 10: **IT-Optimization exhibits improved stability compared to guidance.** Crystallographic metrics computed across five independent runs on four proteins for electron density-guided AlphaFold3 and electron density-based Inference-time Optimization (IT-Opt). IT-Opt consistently achieves improved agreement with experimental density, reflected by higher cosine similarity and lower R_{free} and R_{work}, while also exhibiting reduced variability in metrics across runs. **Bold** entries reflect these consistent improvements.

Donor	Acceptor	AlphaFold3	IT-Opt	Δ
A:ARG280:NH1	B:GLU493:O	0/5	3/5	+3
A:ARG248:NH1	B:GLU495:OE2	4/5	5/5	+1
B:ASN473:ND2	A:ASN247:O	0/5	1/5	+1
B:TRP498:NE1	A:SER241:O	4/5	5/5	+1
A:ARG248:NH2	B:ASP475:OD1	4/5	5/5	+1
A:ASN247:ND2	B:TYR469:OH	0/5	1/5	+1
A:HIS178:NE2	B:MET422:O	0/5	0/5	0
A:ARG248:NH1	B:ASP494:OD2	0/5	0/5	0
A:SER183:N	B:SER425:O	0/5	0/5	0
B:ASN473:ND2	A:ARG248:O	4/5	3/5	−1

Table 11: **Inter-chain hydrogen-bond recovery in the `1YCS` complex.** This table summarizes hydrogen-bond donor-acceptor pairs observed in the crystal structure of `1YCS` and compares their recovery across sampled predictions. The AlphaFold3 column reports the number of AlphaFold3 samples (out of 5) in which a given hydrogen bond is present, while IT-Opt reports the corresponding count for ipTM-based inference-time optimization. **Δ** denotes the difference (IT-Opt − AlphaFold3). Across all listed contacts, unguided AlphaFold3 recovers 16 out of 50 possible hydrogen-bond occurrences (average recovery rate 0.32), whereas ipTM-based inference-time optimization recovers 23 out of 50 (average recovery rate 0.46), corresponding to an absolute increase of 0.14, and a relative enrichment of 44%.

Donor	Acceptor	AlphaFold3	IT-Opt	Diff
A:LYS27:NZ	B:ASP57:OD2	8/100	21/100	+13
A:ARG23:NE	B:PHE54:O	11/100	22/100	+11
A:LYS27:NZ	B:GLU56:OE2	10/100	17/100	+7
A:ARG23:NE	B:TRP53:O	20/100	24/100	+4
A:ARG23:NE	B:GLU56:OE2	9/100	12/100	+3
A:GLY1:N	B:PRO60:O	0/100	1/100	+1
A:ARG9:NH1	B:TRP53:O	16/100	17/100	+1
A:LYS27:NZ	B:TRP53:O	0/100	1/100	+1
A:GLY1:N	B:GLY59:O	0/100	1/100	+1

Table 12: **Inter-chain hydrogen-bond recovery in the `2YL4` complex.** This table summarizes hydrogen-bond donor-acceptor pairs observed in the NMR-determined structure of `2YL4` at the helical region in the peptide (highlighted in Figure 6, Panel B), and compares their recovery across sampled predictions. The AlphaFold3 column reports the number of AlphaFold3 samples (out of 100) in which a given hydrogen bond is present, while IT-Opt reports the corresponding count for ipTM-based inference-time optimization. **Δ** denotes the difference (IT-Opt − AlphaFold3). Across the listed contacts, unguided AlphaFold3 recovers 74/900 hydrogen-bond occurrences (0.082), whereas ipTM-based inference-time optimization recovers 116/900 (0.129), corresponding to a relative enrichment of 57%.

I ALGORITHMS

Algorithm 2 Inference-time Optimization

Input: Sequence $\mathbf{a}$; input features $\{\mathbf{f}^*\}$, $\{\mathbf{s}_i^{\text{inputs}}\}$; initial trunk embeddings $\mathbf{Z}_{\text{init}}$; experimental observation $\mathbf{y}$; noise schedule $[\sigma_0, \sigma_1, \ldots, \sigma_T]$; noise factor $\gamma_0 = 0.8$, minimum noise factor $\gamma_{\text{min}} = 1.0$; noise scale λ; step scale κ; ensemble size n; outer iterations K; inner MSA steps $M = 1$; learning rate η_z; prior weight λ_p
Output: Optimized batch trunk embeddings $\boldsymbol{\mathcal{Z}}$

$\boldsymbol{\mathcal{Z}} = [\mathbf{Z}_{\text{init}}, \ldots \mathbf{Z}_{\text{init}}]$ ▷ Initialize from Pairformer
for $k = 1$ to K ▷ Outer optimization loops
  $\boldsymbol{\mathcal{X}}_l \sim \sigma_T \cdot [\mathbf{N}^1, \ldots, \mathbf{N}^n]^T$ $\mathbf{N}^i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \boldsymbol{\mathcal{X}}_l \in \mathbb{R}^{n \times m \times 3}$
  for $\sigma_\tau \in [\sigma_{T-1}, \ldots, \sigma_0]$ ▷ Reverse diffusion schedule
    $\boldsymbol{\mathcal{X}}_l \leftarrow \text{CentreRandomAugmentation}(\boldsymbol{\mathcal{X}}_l)$
    $\gamma \leftarrow \gamma_0$ if $\sigma_\tau > \gamma_{\text{min}}$ else 0
    $\hat{t} \leftarrow \dfrac{\sigma_{\tau-1}(\gamma + 1)}{\sqrt{t^2 - \sigma_\tau^2 / \sigma_{\tau-1}}}$
    $\boldsymbol{\xi}_l \leftarrow \lambda \sqrt{\hat{t}^2 - \sigma_\tau^2} \cdot [\mathbf{N}^1, \ldots, \mathbf{N}^n]^T$ $\mathbf{N}^i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
    $\boldsymbol{\mathcal{X}}_l^{\text{noisy}} \leftarrow \boldsymbol{\mathcal{X}}_l + \boldsymbol{\xi}_l$
    for $j = 1$ to M ▷ Inner embedding optimization
      $\hat{\boldsymbol{\mathcal{X}}}_0 \leftarrow \text{DiffusionModule}(\{\boldsymbol{\mathcal{X}}_l^{\text{noisy}}\}, \hat{t}, \{\mathbf{f}^*\}, \{\mathbf{s}_i^{\text{inputs}}\}, \boldsymbol{\mathcal{Z}})$
      $\boldsymbol{\mathcal{Z}} \leftarrow \boldsymbol{\mathcal{Z}} + \nabla_{\boldsymbol{\mathcal{Z}}}(\eta_z \log p(\mathbf{y} | \hat{\boldsymbol{\mathcal{X}}}_0) + \lambda_p \log p(\boldsymbol{\mathcal{Z}} | \mathbf{a}))$ ▷ Embedding update
    end for
    $\hat{\boldsymbol{\mathcal{X}}}_0 \leftarrow \text{DiffusionModule}(\{\boldsymbol{\mathcal{X}}_l^{\text{noisy}}\}, \hat{t}, \{\mathbf{f}^*\}, \{\mathbf{s}_i^{\text{inputs}}\}, \boldsymbol{\mathcal{Z}})$
    $\delta_l \leftarrow (\boldsymbol{\mathcal{X}}_l - \hat{\boldsymbol{\mathcal{X}}}_0)/\hat{t}$
    $dt \leftarrow \beta_\tau - \hat{t}$
    $\boldsymbol{\mathcal{X}}_l \leftarrow \boldsymbol{\mathcal{X}}_l^{\text{noisy}} + \kappa \cdot dt \cdot \delta_l$ ▷ Reverse diffusion step
  end for
end for
return $\boldsymbol{\mathcal{Z}}$

Algorithm 3 Inference-time Optimization with Boltzmann Reweighting

Input: Same as Algorithm 2, plus: energy function $E_\phi \colon \mathbb{R}^{m\times 3} \to \mathbb{R}$; inverse temperature $\beta = 1.68\ \text{kcal}^{-1}\text{mol}$
Output: Optimized batch trunk embeddings $\mathcal{Z}$
$\mathcal{Z} = [\mathbf{Z}_{\text{init}}, \dots \mathbf{Z}_{\text{init}}]$ ▷ Initialize from Pairformer
for $k = 1$ to K ▷ Outer optimization loops
  $\mathcal{X}_l \sim \sigma_T \cdot [\mathbf{N}^1, \dots, \mathbf{N}^n]^T$ $\mathbf{N}^i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \mathcal{X}_l \in \mathbb{R}^{n\times m\times 3}$
  for $\sigma_\tau \in [\sigma_{T-1}, \dots, \sigma_0]$ ▷ Reverse diffusion schedule
    $\mathcal{X}_l \leftarrow \text{CentreRandomAugmentation}(\mathcal{X}_l)$
    $\gamma \leftarrow \gamma_0$ if $\sigma_\tau > \gamma_{\text{min}}$ else 0
    $\hat{t} \leftarrow \dfrac{\sigma_{\tau-1}(\gamma+1)}{\sqrt{t^2 - \sigma_\tau^2/\sigma_{\tau-1}}}$
    $\boldsymbol{\xi}_l \leftarrow \lambda\sqrt{\hat{t}^2 - \sigma_\tau^2} \cdot [\mathbf{N}^1, \dots, \mathbf{N}^n]^T$ $\mathbf{N}^i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
    $\mathcal{X}_l^{\text{noisy}} \leftarrow \mathcal{X}_l + \boldsymbol{\xi}_l$
    for $j = 1$ to M ▷ Inner optimization loop with reweighting
      $\hat{\mathcal{X}}_0 \leftarrow \text{DiffusionModule}(\{\mathcal{X}_l^{\text{noisy}}\}, \hat{t}, \{\mathbf{f}^*\}, \{\mathbf{s}_i^{\text{inputs}}\}, \mathcal{Z})$
      $w^i \leftarrow \dfrac{\exp(-\beta E_\phi(\hat{\mathbf{X}}_0^i))}{\sum_{k=1}^n \exp(-\beta E_\phi(\hat{\mathbf{X}}_0^k))}$ for $i \in [n]$ ▷ Boltzmann weights
      $\mathcal{Z} \leftarrow \mathcal{Z} + \nabla_{\mathcal{Z}}(\eta_z \log p(\mathbf{y}|\hat{\mathcal{X}}_0; \mathbf{w}) + \lambda_p \log p(\mathcal{Z}|\mathbf{a}))$ ▷ Embedding update
    end for
    $\hat{\mathcal{X}}_0 \leftarrow \text{DiffusionModule}(\{\mathcal{X}_l^{\text{noisy}}\}, \hat{t}, \{\mathbf{f}^*\}, \{\mathbf{s}_i^{\text{inputs}}\}, \mathcal{Z})$
    $\delta_l \leftarrow (\mathcal{X}_l - \hat{\mathcal{X}}_0)/\hat{t}$
    $dt \leftarrow \beta_\tau - \hat{t}$
    $\mathcal{X}_l \leftarrow \mathcal{X}_l^{\text{noisy}} + \kappa \cdot dt \cdot \delta_l$ ▷ Reverse diffusion step
  end for
end for
return $\mathcal{Z}$