Generative modeling of protein ensembles guided by crystallographic electron densities

Sai Advaith Maddipatla*
Technion, Israel

Nadav Bojan Sellam*
Technion, Israel

Sanketh Vedula
Technion, Israel

Ailie Marx
Migal – Galilee Research Institute, Israel

Alex Bronstein
Technion, Israel and IST Austria

Abstract

Proteins are dynamic, adopting ensembles of conformations. The nature of this conformational heterogenity is imprinted in the raw electron density measurements obtained from X-ray crystallography experiments. Fitting an ensemble of protein structures to these measurements is a challenging, ill-posed inverse problem. We propose a non-i.i.d. ensemble guidance approach to solve this problem using existing protein structure generative models and demonstrate that it accurately recovers complicated multi-modal alternate protein backbone conformations observed in certain single crystal measurements.

1 Introduction

Proteins are dynamic molecules, transitioning between states during biological processes and otherwise thermodynamically sampling conformations within these states. Although models generated by X-ray crystallography typically depict a single conformation, this is actually an ensemble measurement. Protein crystals are an enormous array of molecules and the electron density reconstructed from the diffraction captures variability between the atom locations in this array. As flexibility in the protein chain increases, the electron density becomes increasingly spread out. Since it is difficult to recognise and model the specific conformations giving rise to the average density, variability around the best-fit model is typically reported only indirectly, in the form of the B factor. However, where detectable, crystallographers model atoms in multiple alternate locations (commonly termed, altlocs). Alternately located segments of the protein backbone have remained under-recognised, since most visualisation platforms (e.g., pymol and chimeraX) and programs using structural models as inputs (e.g., gromacs) ignore altlocs altogether or resolve them with simple heuristics [4]. A recent work [11] created a comprehensive catalogue of altlocs extracted from PDB structures, suggesting that this dataset should find use in efforts towards predicting multiple structures from a single sequence. Interestingly, the authors showed that for a set of well-separated and stable altlocs, even if structural ensemble predictors recognise the region as being flexible, they fail to capture the experimentally determined conformations or even the bimodality of the distribution of backbone conformations.

Representing proteins as ensembles of model structures, better capturing their dynamic nature, is a key goal in structural biology [15] and protein structure prediction. We suggest an generative approach that directly produces samples that are faithful to the experimental data. We showcase the proposed approach on electron densities resolved from X-ray crystallography, focusing on altloc regions having at least two clearly distinct conformations.

---

*Equal contribution. Correspondence Email: {maddipatlas, n.sellam}@campus.technion.ac.il

Contributions. We formulate the problem of reconstructing a protein structure ensemble that obeys given experimental measurements as an inverse problem, in which the forward model describes the formation of the electron density given an ensemble. We use a pre-trained diffusion model to embody the prior joint probability of the atom coordinates given the amino acid sequence, and guide it to sample from the conditional probability given the observed density. We develop a differentiable forward model and a non-i.i.d. score guidance method that uses the entire ensemble sampled from the model in contrast to the standard score guidance that uses individual samples. We demonstrate that the proposed methodology faithfully recovers alternate conformations from crystallographic electron densities that are not reproduced correctly by unconditional sampling.

2 Methods

Notation. We denote by $\mathbf{a}$ the amino-acid sequence, $\mathbf{x}$ the backbone coordinates of a protein structure, and let $\boldsymbol{\chi}$ be the sidechain dihedral angles from which the side chain atom coordinates can be straightforwardly computed and denoted by $\{\mathbf{y}_i(\mathbf{x}, \boldsymbol{\chi}, \mathbf{a})\}$. Note that the dependence on the sequence $\mathbf{a}$ is implicit as the amino acid identities are required to calculate the atom locations from $\mathbf{x}$ and $\boldsymbol{\chi}$. We denote by $p(\mathbf{x})$ the distribution of the random variable $\mathbf{x}$, and by $F_{\mathrm{o}} : \mathbb{R}^3 \to \mathbb{R}$ and $F_{\mathrm{c}} : \mathbb{R}^3 \to \mathbb{R}$ the observed and calculated electron densities, respectively.

Problem setting. Given an experimentally observed electron density of a protein structure F_{o} and the amino-acid sequence $\mathbf{a}$, our goal is to recover the posterior distribution $p(\mathbf{x}, \boldsymbol{\chi}|\mathbf{a}, F_{\mathrm{o}})$ of protein structure that explains F_{o} by sampling a *non-i.i.d.* ensemble $\mathcal{X} = \{(\mathbf{x}^m, \boldsymbol{\chi}^m)\}$ from it. Following the standard Bayes' theorem, the posterior of the ensemble can be factorized as $p(\mathcal{X}|\mathbf{a}, F_{\mathrm{o}}) \propto p(F_{\mathrm{o}}|\mathcal{X}, \mathbf{a}) \cdot p(\mathcal{X}|\mathbf{a}) = p(F_{\mathrm{o}}|\mathcal{X}, \mathbf{a}) \cdot \prod_k p(\mathbf{x}^m, \boldsymbol{\chi}^m|\mathbf{a})$, where $p(F_{\mathrm{o}}|\mathcal{X}, \mathbf{a})$ is the *likelihood* of observing F_{o} given a structure ensemble $\mathcal{X}$ and $p(\mathbf{x}, \boldsymbol{\chi}|\mathbf{a})$ is the *prior* distribution of the protein conformations given the sequence. It is crucial to observe that the samples in the ensemble are not independent. While the prior can be split into independent terms $p(\mathbf{x}^m, \boldsymbol{\chi}^m|\mathbf{a})$, the likelihood generally depends on all the samples at once and is, therefore, inseparable. Our approach is summarized in Figure 1.

Protein structure prior. We employ `Chroma` [5] to model the distribution of feasible protein structures. While other protein structure generative models exist [1, 6], we chose `Chroma` because it is a score-based generative models, and it effectively models $p(\mathbf{x}, \boldsymbol{\chi}, \mathbf{a})$. We also experimented with distributional graphormer (DiG)[17], however, we found that the score modeled by DiG is unstable, and it models only the backbone atoms, and does not model backbone oxygens and sidechains, which contribute to the majority of the electron density.
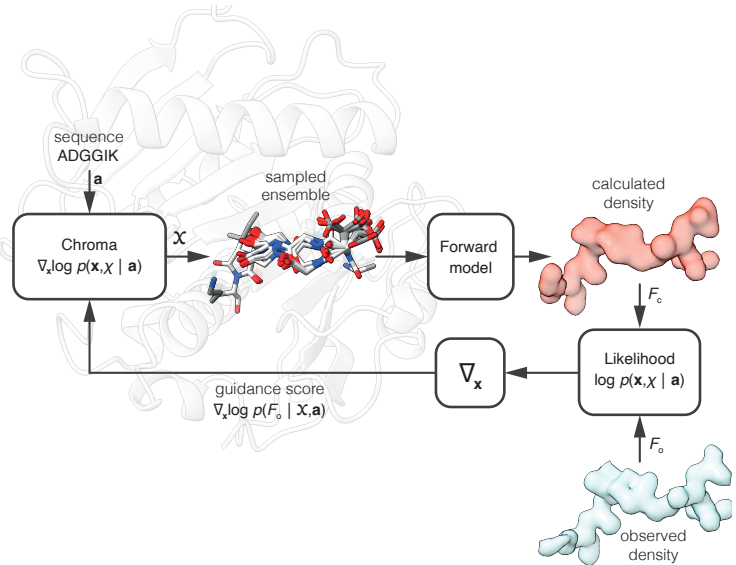

Figure 1: **The proposed density-guided protein ensemble generation method.** The diffusion model is used to sample a non-i.i.d. ensemble from which the likelihood of the observed electron density is calculated and used to guide the sampling.

`Chroma` models the all-atom representation of the protein structure by the following factorized form, $p(\mathbf{x}, \boldsymbol{\chi}, \mathbf{a}) = p(\mathbf{x})p(\mathbf{a}, \boldsymbol{\chi}|\mathbf{x})$, where the distribution of backbone coordinates, $p(\mathbf{x})$, is modeled via a score-based diffusion model, and $p(\mathbf{a}, \boldsymbol{\chi}|\mathbf{x})$ is modeled as an autoregressive model. `Chroma`'s forward diffusion process is modeled as a variance-preserving SDE [13], whose

backward SDE is given by,

$$dx = \left[-\frac{1}{2}\mathbf{x} - \mathbf{P}\mathbf{P}^\top \nabla_\mathbf{x} \log p_t(\mathbf{x}) \right] \beta_t dt + \sqrt{\beta_t} \mathbf{P}\mathbf{n}, \qquad (1)$$

where $\mathbf{P}$ is a fixed preconditioning matrix, β_t is the noise schedule, and $\mathbf{n}$ is isotropic Gaussian noise. The score function $\nabla_\mathbf{x} \log p_t(\mathbf{x})$ is modeled as an equivariant graph attention network. Sampling occurs by first drawing backbone coordinates by integrating Eq. 1 from $[T, 0]$. Modeling $\mathbf{x}_{\text{all}}$ given $\mathbf{x}$ is presented in the sequel. Because `Chroma` models the backbone coordinates $\mathbf{x}$ independently to the sequence, conditioning $\mathbf{x}_{\text{all}}$ on the sequence $\mathbf{a}$ can be done trivially by substitution. We discuss the details of sidechain packing in Section A.1 in the Appendix.

Modeling the likelihood. Assuming $p(F_\text{o}|\mathcal{X}, \mathbf{a})$ is Gaussian, the log-likelihood of F_o is given by,

$$\log p(F_\text{o}|\mathcal{X}, \mathbf{a}) = - \left\| F_\text{o} - \frac{1}{|\mathcal{X}|} \sum_m F_\text{c} \left(\mathbf{x}^m, \boldsymbol{\chi}^m, \mathbf{a} \right) \right\|_2^2, \qquad (2)$$

where $\| \cdot \|_2$ is the L_2 norm in the space of electron densities, and the individual F_c terms are given by the kernel density estimates,

$$F_\text{c}(\boldsymbol{\xi}) = \sum_{k=1}^{N_s} \sum_{i=1}^{A} \sum_{j=1}^{5} a_{ij} \cdot \left(\frac{4\pi}{b_{ij} + B} \right)^{\frac{3}{2}} \cdot \exp\left(-\frac{4\pi^2}{b_{ij} + B} \cdot \| \mathbf{R}_k \mathbf{y}_i(\mathbf{x}^m, \boldsymbol{\chi}^m, \mathbf{a}) + \mathbf{t}_k - \boldsymbol{\xi} \|_2^2 \right), \quad (3)$$

where N_s is the number of symmetry operations, A is the number of atoms in the asymmetric unit, $\mathbf{R}_k$ is the rotation matrix of the k-th symmetry operation, $\mathbf{t}_k$ is the translation vector of the k-th symmetry operation, $\mathbf{y}_i$ is the location of the i-th atom, a_{ij} and b_{ij} are tabulated form factors [10], B is the B-factor, and $\boldsymbol{\xi}$ is the point in space at which density is calculated. In standard molecular replacement and structure refinement pipelines, the B-factor is used to model the experimental electron density as a mixture of Gaussians. In our approach, however, we essentially sample directly from the joint distribution of the atom locations which allows more complex density models. The B-factor is, therefore, more akin to the bandwidth parameter in kernel density estimation (KDE) techniques. We apply a uniform B-factor across all atoms that should be inversely proportional to the ensemble size. Ideally, the optimization of the exact value should be part of the sampling procedure that we intend to develop in the future; here, we adopted a somewhat naive approach by taking the average of the individual B-factors for each atom in the given protein structure.

Sampling from the posterior. To sample a non-independent ensemble from the posterior, we define the joint diffusion variable $\mathbf{X} = (\mathbf{x}^1, \ldots, \mathbf{x}^{|\mathcal{X}|})$ concatenating all backbone coordinates of the ensemble, and plug Eq. 2 into the backward SDE equation [3, 16],

$$d\mathbf{X} = \left[-\frac{1}{2}\mathbf{X} - \mathbf{P}\mathbf{P}^\top (\nabla_\mathbf{X} \log p_t(\mathbf{X}) + \eta \cdot \nabla_\mathbf{X} \log p(F_\text{o}|\mathbf{X}, \boldsymbol{\chi}(\mathbf{X}), \mathbf{a})) \right] \beta_t dt + \sqrt{\beta_t} \mathbf{P}\mathbf{n}. \quad (4)$$

Note that the diffusion is only applied to the backbone coordinates, from which the sidechain dihedrals $\boldsymbol{\chi}(\mathbf{X}) = (\boldsymbol{\chi}^1, \ldots, \boldsymbol{\chi}^{|\mathcal{X}|})$ are calculated using the differentiable sidechain packer. While the unconditional (prior) score term, $\nabla_\mathbf{X} \log p_t(\mathbf{X})$, is separable (i.e., block diagonal), the guidance term is inseparable as mentioned above. The hyperparameter η is used to scale the guidance score directing the diffusion model to generate samples that are better aligned with the observed density, thereby increasing the likelihood of Eq. 2.

Filtering samples using matching pursuit. We observed that when sampling large ensembles using non-i.i.d. guidance, certain samples in the ensemble tend to overfit to noise in the density map, reducing the overall ensemble quality. To mitigate this, we adopt a matching pursuit-based approach [8] to filter out low-quality samples and greedily select a subset of the ensemble, $\mathcal{X}_\mathcal{I} = \{ (\mathbf{x}^m, \boldsymbol{\chi}^m) : m \in \mathcal{I} \}$, that best aligns with F_o. Starting with $\mathcal{I} = \emptyset$, each iteration seeks to maximize $\log p(F_\text{o}|\mathcal{X}_{\mathcal{I} \cup \{m\}}, \mathbf{a})$ over all $m \notin \mathcal{I}$. The optimal element m is then added to the support set $\mathcal{I}$ and the process is repeated until the likelihood no longer increases. The procedure is summarized as Algorithm A2 in Appendix A.
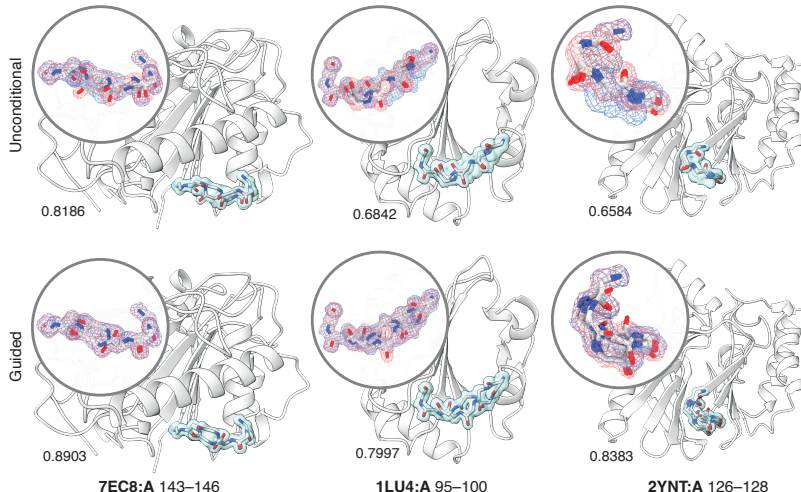
Figure 2: **Density-guided** `Chroma` **fits the density better than unguided** `Chroma`**, and recovers two known alternative locations accurately.** Full protein structure is displayed as white cartoons. The sampled ensembles in the region of interest are depicted as sticks and overlaid on the experimental density 1σ-isosurface. The inserts show the agreement of F_c (red) to the observed density F_o (blue) visualized as 1σ-isomeshes. Cosine similarities F_c and F_o are reported below each panel. Density guidance produces consistently better density alignment and correctly captures the multi-modal nature of the observed density.

3 Results

Experimental setup. To evaluate the efficacy of our methods, we primarily aim to assess the alignment between the density computed from sampled conformers, F_c, and the experimentally observed densities, F_o. Additionally, we seek to investigate the ability of our approach to generate accurate conformers in the structural space. We, therefore, focus on regions where the density exhibits multi-modal behavior and aim to recover alternate locations (altlocs) consistent with those modeled in the original PDB structures. To facilitate these investigations, we selected crystallographic protein structures from the Protein Data Bank (PDB) [2] that exhibit discernible separation between backbone and sidechain atoms at the altloc residues using the analysis in [11]. Note that all these examples present highly intricate multi-modal backbone distributions that are poorly predicted by existing ensemble sampling techniques. We further restricted our attention to high resolution structures (2.5 Å or better) to ensure high-quality electron density maps. Since the electron density maps available in the PDB are mean-centered and lack an absolute scale, we converted them to physical units ($e^-/\text{Å}^3$) following the method described in [7]. A comprehensive list of proteins used in our experiments is reported in Table A3.

For all experiments, we will use the publicly available versions of `Chroma`'s diffusion-based backbone sampler and `Chroma`'s GNN-based sidechain packer for χ-angle prediction as our prior [5]. Since the backbone sampler is not sequence-conditioned, applying density guidance across the entire protein would exacerbate the ill-posed nature of the problem. To mitigate this, we employ `Chroma`'s `SubstructureConditioner` to constrain atoms outside the target residues to their ground truth locations. All visualizations were rendered using ChimeraX [9].

Evaluation criteria. The results were evaluated using a batch sampling approach. For each batch, we identified a subset of samples that provided the optimal match to the observed electron density. The similarity between the observed electron density and the mean calculated density of the selected subset was quantified using cosine similarity $\langle F_o, F_c \rangle / (\|F_o\| \cdot \|F_c\|)$ in the space of electron densities. Furthermore, we computed an alternate location (altloc) proximity score for each sample in the batch. The score was calculated as $\texttt{bimodality\_score}(\mathbf{x}) = (1 - \min(r(\mathbf{x}), 1/r(\mathbf{x}))) \cdot \text{sign}(r(\mathbf{x}) - 1/r(\mathbf{x}))$, where $r(\mathbf{x}) = \|\mathbf{x} - \mathbf{x}_A\|/\|\mathbf{x} - \mathbf{x}_B\|$, and $\mathbf{x}_A$ and $\mathbf{x}_B$ are the backbone atom locations of altlocs A and B, respectively. In this formulation, $\text{score}(\mathbf{x})$ approaches -1 as $\mathbf{x}$ converges to altloc A, and approaches $+1$ as $\mathbf{x}$ converges to altloc B.
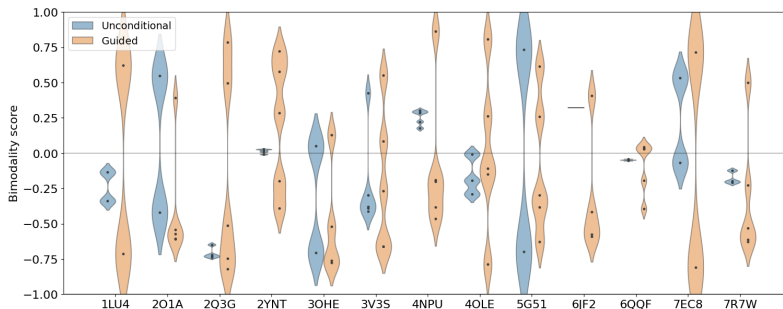
Figure 3: **Density-guided** `Chroma` **accurately captures the bimodal distribution of the backbone conformations, while unconditional sampling consistently fails to represent it.** Negative scores represent proximity to the modeled altloc A, while positive scores correspond to altloc B.

Density alignment. To evaluate the proposed method's effectiveness in aligning with F_o, we compare the samples generated using density-guided diffusion with those from an unconditional diffusion process. Using the procedure outlined in Algorithm A1 we sample a batch of 16 density-guided conformers. Likewise, a batch of 16 unconditional conformers are sampled using Eq. 1. Both batches are filtered (Algorithm A2) to retain at most 5 samples that best fit the observed F_o map. From these best-performing conformers, we construct the F_c map and compute the correlation with F_o using cosine similarity. This evaluation is repeated across all proteins listed in Table $A3$. The quantitative results comparing unconditional sampling and density-guided sampling are presented in Figure $A3$. Due to variations in resolution, the results are not comparable across proteins. However, we note that while `Chroma` provides a strong prior for protein structures, it struggles to accurately capture experimental density. This reinforces the finding that other protein ensemble sampling programs cannot correctly reproduce the observed densities in such intricately flexible regions [11]. Additionally, Figure 2 visually showcases, on three structures, how guided sampling improves the alignment between F_c and F_o.

Bimodality and structural alignment. We evaluate the impact of aligning with the experimental density on the structural space. Specifically, we focus on regions of proteins with alternate conformations (altlocs) where the density exhibits a bimodal distribution. By conditioning our diffusion process on F_o, we aim to capture this bimodality and generate samples consistent with the originally modeled altlocs in the PDB. From the aforementioned experiment, we utilize the filtered samples from both density-guided and unconditional ensembles. Each sample is assigned proximity score to one of the altloc (A or B). The resulting score distributions are visualized in Figure 3. Guided sampling consistently achieves bi- and multi-modal behavior with proximities to both modeled altlocs (positive and negative modes in the plot), while the unconditional counterpart often fails to correctly represent this behavior. Additionally, our method performs well in regions with unimodal density and no alternative conformations, as detailed in Section A.3 and visualized in Figures A1 and A2 in the Appendix.

4   Conclusion

In this work, we presented a novel density-guided generative modeling approach for reconstructing protein structure ensembles from crystallographic electron density maps. By formulating the task as an inverse problem, we leveraged a pre-trained diffusion model as a flexible prior over protein backbone conformations, introducing a non-i.i.d. score guidance technique that optimizes the ensemble as a whole, rather than individual structures thereof. Our method outperforms unconditional sampling, particularly in multimodal regions, recovering altlocs consistent with PDB data and aligning more closely with the observed electron densities. Looking forward, our method opens several promising directions for future exploration, such as extending to larger proteins, enhancing the modeling of B-factors, improving sidechain packing accuracy, and extending to the forward models of other experimental modalities such as cryoEM/ET. We believe that our work also provides an important step toward more accurate and data-driven modeling of protein dynamics.

References

[1] Avishek Joey Bose et al. "SE(3)-Stochastic Flow Matching for Protein Backbone Generation". In: *The International Conference on Learning Representations (ICLR)*. 2024.

[2] Stephen K Burley et al. "RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences". In: *Nucleic acids research* 49.D1 (2021), pp. D437–D451.

[3] Prafulla Dhariwal and Alexander Nichol. "Diffusion models beat gans on image synthesis". In: *Advances in neural information processing systems* 34 (2021), pp. 8780–8794.

[4] Torben Gutermuth et al. "Modeling with alternate locations in X-ray protein structures". In: *Journal of Chemical Information and Modeling* 63.8 (2023), pp. 2573–2585.

[5] John B. Ingraham et al. "Illuminating protein space with a programmable generative model". In: *Nature* (2023). DOI: 10.1038/s41586-023-06728-8.

[6] Bowen Jing, Bonnie Berger, and Tommi Jaakkola. "AlphaFold Meets Flow Matching for Generating Protein Ensembles". In: *Forty-first International Conference on Machine Learning*. 2024.

[7] P. Therese Lang et al. "Protein structural ensembles are revealed by redefining X-ray electron density noise". In: *Proceedings of the National Academy of Sciences* 111.1 (2014), pp. 237–242. DOI: 10.1073/pnas.1302823110. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.1302823110. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1302823110.

[8] Stéphane G Mallat and Zhifeng Zhang. "Matching pursuits with time-frequency dictionaries". In: *IEEE Transactions on signal processing* 41.12 (1993), pp. 3397–3415.

[9] Eric F Pettersen et al. "UCSF ChimeraX: Structure visualization for researchers, educators, and developers". In: *Protein science* 30.1 (2021), pp. 70–82.

[10] Edward Prince. *International Tables for Crystallography, Volume C: Mathematical, physical and chemical tables*. Springer Science & Business Media, 2004.

[11] Aviv A. Rosenberg et al. "Seeing Double: Molecular dynamics simulations reveal the stability of certain alternate protein conformations in crystal structures". In: *bioRxiv* (2024). DOI: 10.1101/2024.08.31.610605. eprint: https://www.biorxiv.org/content/early/2024/08/31/2024.08.31.610605.full.pdf. URL: https://www.biorxiv.org/content/early/2024/08/31/2024.08.31.610605.

[12] Maxim V Shapovalov and Roland L Dunbrack. "A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions". In: *Structure* 19.6 (2011), pp. 844–858.

[13] Yang Song and Stefano Ermon. "Generative modeling by estimating gradients of the data distribution". In: *Advances in neural information processing systems* 32 (2019).

[14] David A Van Dyk and Taeyoung Park. "Partially collapsed Gibbs samplers: Theory and methods". In: *Journal of the American Statistical Association* 103.482 (2008), pp. 790–796.

[15] Stephanie A. Wankowicz et al. "Uncovering Protein Ensembles: Automated Multiconformer Model Building for X-ray Crystallography and Cryo-EM". In: *bioRxiv* (2023). DOI: 10.1101/2023.06.28.546963. eprint: https://www.biorxiv.org/content/early/2023/06/29/2023.06.28.546963.full.pdf. URL: https://www.biorxiv.org/content/early/2023/06/29/2023.06.28.546963.

[16] Tomer Weiss et al. "Guided diffusion for inverse molecular design". In: *Nature Computational Science* 3.10 (Oct. 2023), pp. 873–882. ISSN: 2662-8457. DOI: 10.1038/s43588-023-00532-0. URL: https://doi.org/10.1038/s43588-023-00532-0.

[17] Shuxin Zheng et al. "Predicting equilibrium distributions for molecular systems with deep learning". In: *Nature Machine Intelligence* (2024), pp. 1–10.

A   Appendix and Supplemental Material

A.1   Side-chain packing

As previously mentioned in Section 2, the diffusion model's prior is defined over the backbone coordinates of the protein structure. However, majority of the electron density is concentrated near the sidechain atoms. To account for this, we leverage `Chroma`'s sidechain packer to sample the χ dihedral angles and impute the sidechain atoms using the sampled angles and the rotamer library [12]. Since the density map is generated from a noise-free protein structure, we denoise the backbone atoms $\mathbf{x}$ at each iteration using the diffusion model before computing the χ angles. The denoised variable $\hat{\mathbf{x}}_0$ serves as an estimate of the noiseless backbones $\mathbf{x}$.

The χ sampler is an autoregressive sidechain decoder that models $l_\theta(\boldsymbol{\chi}|\mathbf{x}, \mathbf{a}) \approx \log p(\boldsymbol{\chi}|\mathbf{x})$, where θ denote the model parameters. The autoregressive decomposition of the likelihood is given by

$$l_\theta(\boldsymbol{\chi}|\mathbf{x}, \mathbf{a}) = \sum_i l_\theta(\chi_i|\chi_{i-1}, \chi_{i-2}, \ldots \chi_1, \mathbf{a}, \mathbf{x}).$$

Once the χ angles are sampled, the combination with the backbone coordinates $\mathbf{x}$ fully describes the protein structure.

Unfortunately, sampling χ angles for every amino acid at each timestep of the diffusion process introduces a significant computational bottleneck, which can severely impede the sampling procedure. To mitigate this, we restrict χ dihedral angle sampling to the amino acids being optimized, along with a window of size $W = 5$ amino acids before and after the target region. Incorporating this window provides the model with greater structural context, thereby enhancing the accuracy of the angle predictions. For non-target residues, where the structure is anchored by the `SubstructureConditioner`, χ angles are computed using the ground truth coordinates from the PDB file.

Additionally, performing χ dihderal angle sampling and backpropagating through the χ sampling and the diffusion models at every iteration results in an excessively long and computationally expensive backpropagation graph. Given that atomic coordinates generally exhibit minimal change between iterations, this repeated sampling and backpropagation can become redundant. Hence, we implement a Gibbs-sampling-based optimization strategy [14].

Specifically, we sample the dihedral angles only once every T_χ iterations and then detach them from the backpropagation graph. During the intermediate iterations, we freeze the backbone coordinates $\mathbf{x}$ and optimize the angles S_χ times based on Eq. 2. This approach stabilizes the backpropagtaion process while ensuring that the sidechains conform to the observed density F_o.

The specifics of sidechain packing are presented in Algorithm A1.

A.2   Hyperparameters

In our experiment, we employed a guidance scale of 9.0 and clipped the gradient norms to a maximum of 32.5. The χ diheral angels were resampled every 40 diffusion steps (T_χ). For each diffusion step, 25 SGD optimization steps (S_χ) were performed on the χ with a step size of 0.001. We generated a batch of 16 samples with non-i.i.d guidance. Notably, all the results presented in this paper were obtained using these hyperparameters without any additional tuning, highlighting the simplicity and robustness of our density-guidance approach.

A.3   Unimodal densities

As a control experiment, we evaluated our method on protein 7EC8, specifically on chain B, which exhibits unimodal behavior. We sampled two regions: residues 205-208, which are characterized by a low B-factor, and residues 143-146, which display a higher B-factor. Although both regions are unimodal, the increased B-factor in residues 143-146 led to more spread-out and blurred density, resulting in slight variability in the samples. The cosine similarity results for both regions are shown in Table A1, demonstrating that our guided method consistently outperformed the unconditional approach. A visual comparison of the results is presented in Figures A1 and A2.

Residue Range	Guided	Unconditional
205–208	0.9240	0.7886
143–146	0.9316	0.7814

Table A1: Cosine similarity results for two sampled regions in chain B of protein 7EC8.

7EC8:A 143-146
dual conformation

7EC8:B 143-146
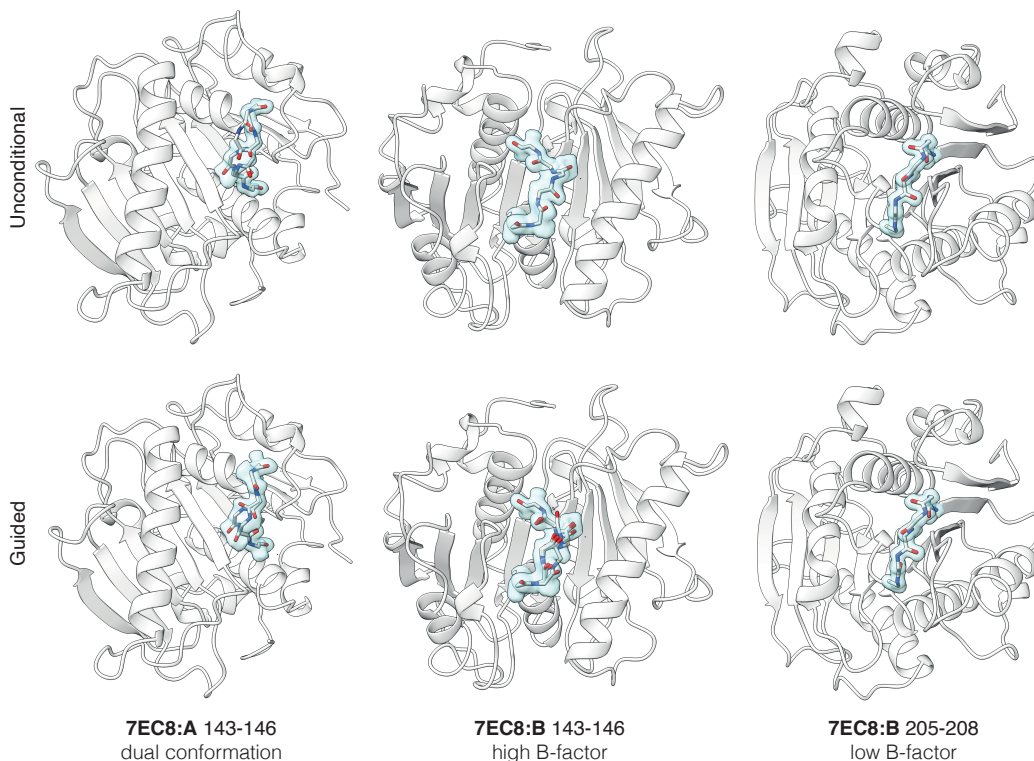high B-factor

7EC8:B 205-208
low B-factor

Figure A1: **Comparison of conditional and unconditional sampling in bimodally- and unimodally-distributed regions of protein** 7EC8:A. The figure illustrates the differences between unconditional sampling (first row) and density-guided (second row) sampling methods in three regions of the protein (left-to-right): residues $143 - 146$ of chain A exhibiting an explicitly modeled dual conformation (two altlocs), the same position in chain B originally modeled as a high B-factors single conformation, and residues $205 - 208$ in chain B, originally modeled as a low B-factors single conformation. Our density-guided sampling consistently describes the flexible region in both chains as a bimodal distribution while producing a tightly distributed ensemble for the third low B-factor region.

A.4 Ablation

We conducted an ablation study on the batched (non-i.i.d) sampling procedure. As discussed in Section 2, our density-guidance approach leverages a batch of non-i.i.d samples, where we compute an expectation over F_c maps and compare it to F_o as depicted in Eq. 2. In contrast, when we generate i.i.d samples by optimizing each sample individually without taking an expectation – the performance deteriorates, as shown in Table A2.
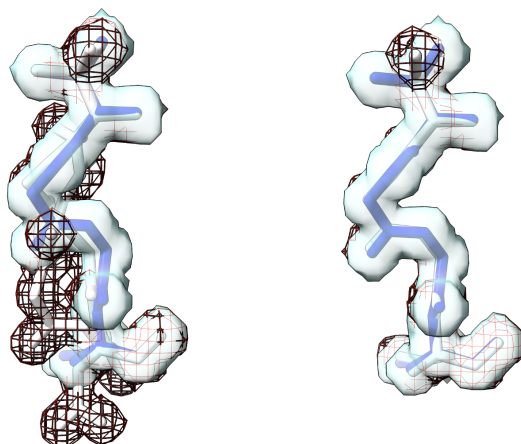
Figure A2: **Comparison of density fitting for residues 205-208 of protein** 7EC8. The unguided approach (left) does not adequately align the calculated density with the observed density, whereas the guided approach (right) demonstrates significantly better alignment. The observed density is represented by the light blue surface, while the calculated density from the samples is shown as a dark red isomesh.
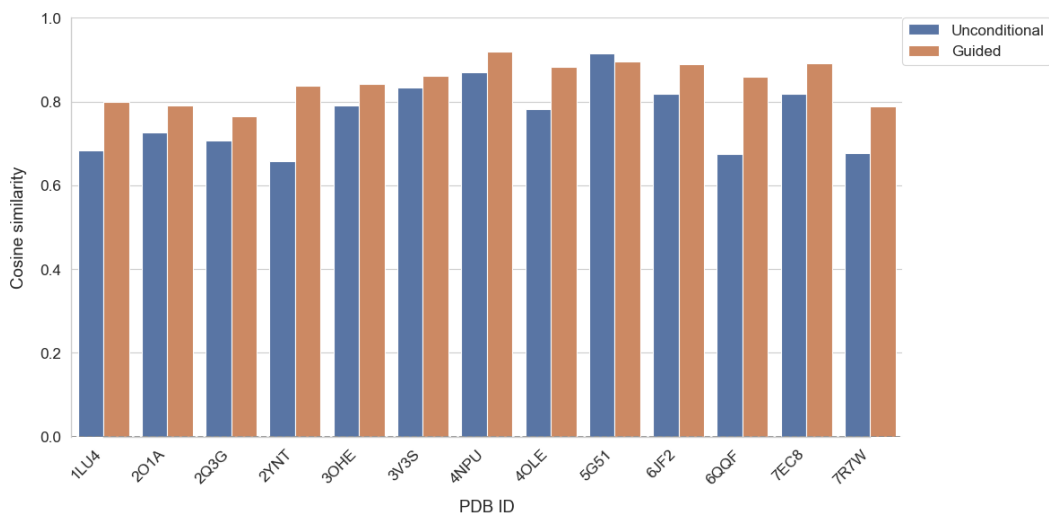

Figure A3: **Comparison of cosine similarities between ensembles obtained using unconditional and density-guided sampling.**

PDB	Guided (batched)	Guided (i.i.d)
2O1A	**0.7903**	0.7459
3OHE	**0.8416**	0.8369
7EC8	**0.8903**	0.8585

Table A2: **Batched sampling of protein ensembles produces superior alignment to F_{o} compared to i.i.d sampling**.

A.5 Algorithms

Algorithm A1 Density-guided backbone sampling

Require: Backbone denoising diffusion model ϵ_θ; χ sampling model χ_θ; χ resample time steps T_χ; number of χ optimization steps S_χ; χ optimization step size λ_χ; target density F_o; number of diffusion steps N; guidance scale λ; preconditioning matrix $\mathbf{P}$; diffusion schedule constants $\{\alpha_n\}_{n=1}^N$, $\{\beta_n\}_{n=1}^N$ as in `chroma`; amino acid sequence of the protein $\mathbf{a}$; F_c density forward model.

1: $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: $\mathbf{x}_N = \mathbf{Pz}$
3: **for** $n = N$ **to** 1 **do**
4: $\quad \mathbf{n} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
5: $\quad \hat{\epsilon} = \epsilon_\theta(\mathbf{x}_n, \mathbf{n})$
6: $\quad \hat{\mu} = \frac{\alpha_n+1}{2(1-\alpha_n)}\mathbf{x}_n - \frac{\sqrt{\alpha_n}}{1-\alpha_n}\hat{\epsilon}$
7: $\quad \hat{\mathbf{x}}_0 = \dfrac{\mathbf{x}_n - \sqrt{1-\alpha_n}\hat{\epsilon}}{\sqrt{\alpha_n}}$
8: $\quad$ **if** $n \in T_\chi$ **then**
9: $\quad\quad \hat{\chi}_n = \chi_\theta(\hat{\mathbf{x}}_0)$
10: $\quad$ **else** $\hfill \triangleright$ Optimize the side-chain angles
11: $\quad\quad$ Initialize $\tilde{\chi}_{n,0} = \hat{\chi}_{n+1}$
12: $\quad\quad$ **for** $i = 0$ **to** $S_\chi - 1$ **do**
13: $\quad\quad\quad \tilde{\chi}_{t,i+1} = \tilde{\chi}_{t,i} - \lambda_\chi \nabla_\chi \|F_c(\texttt{all\_atom}(\hat{\mathbf{x}}_0, \tilde{\chi}_{t,i}, \mathbf{a})) - F_o\|_2^2$
14: $\quad\quad$ **end for**
15: $\quad\quad \hat{\chi}_n = \tilde{\chi}_{n,S_\chi}$
16: $\quad$ **end if**
17: $\quad \mathbf{x}_{n-1} = \hat{\mu} + \lambda\nabla_\mathbf{x}\left[\|F_c(\texttt{all\_atom}(\hat{\mathbf{x}}_0, \hat{\chi}_n, \mathbf{a})) - F_o\|_2^2\right] + \sqrt{\beta_n}\mathbf{Pn}$ $\hfill \triangleright$ Eq. 4
18: **end for**
19: **return** $\mathbf{x}_0$

Algorithm A2 Selecting samples using matching pursuit [8]

Require: $\mathbf{D} = \{d_1, \ldots, d_n\}$: Sample densities ($F_c$) from forward model, d_t: Target density from F_o density map, $\text{corr}(\cdot, \cdot)$: Function to compute correlation metric, $m_{\max}$: Maximum allowed samples to select, $\mathbf{a}$: Amino Acid sequence of the protein

1: $\mathcal{I} = \emptyset$
2: $\mathcal{P} = \{0, 1, 2, \ldots, |D| - 1\}$
3: $s_{\text{current}} = 0$
4: **while** $|\mathcal{I}| < m_{\max}$ **do**
5: $\quad \mathbf{c} = \{\text{corr}(d_{\mathcal{I}\cup\{i\}}, d_t) \mid d_i \in \mathbf{D}, i \notin \mathcal{P}\}$
6: $\quad i_{\text{best}}, s_{\max} = \underset{i}{\arg\max}\,\mathbf{c}, \max\mathbf{c}$ $\hfill \triangleright$ Maximize $\log p(F_o|\mathcal{X}_{\mathcal{I}\cup\{i_{\text{best}}\}}, \mathbf{a})$
7: $\quad \mathcal{I} = \mathcal{I} \cup \{i_{\text{best}}\}$ $\hfill \triangleright$ Add best sample
8: $\quad \mathcal{P} = \mathcal{P} \setminus \{i_{\text{best}}\}$ $\hfill \triangleright$ Update candidates
9: $\quad$ **if** $s_{\max} < s_{\text{current}}$ **then**
10: $\quad\quad$ break
11: $\quad$ **end if**
12: $\quad s_{\text{current}} = s_{\max}$
13: **end while**
14: **return** $\mathcal{I}$

A.6 Dataset

PDB ID	Protein function	Source organism	Expr. system	Altloc site	Altloc sequence	Resolution (Å)
1LU4:A	Oxidoreductase	Mycobacterium tuberculosis	E. coli	95 − 100	YNVPWW	1.12
2O1A:A	Surface active protein	Staphylococcus aureus	E. coli	50 − 53	KQNN	1.60
2Q3G:A	Structural Genomics	Homo sapiens	E. coli	24 − 27	FNVP	1.11
2YNT:A	Hydrolase	Pseudomonas aeruginosa	E. coli	126 − 128	GNG	1.60
3OHE:A	Hydrolase	Marinobacter nauticus VT8	E. coli	97 − 104	YQGDPAWP	1.20
3V3S:B	Hydrolase	Pseudomonas aeruginosa	E. coli	227 − 232	KAQERD	1.90
4NPU:B	Hydrolase	Human immunodeficiency virus 1	E. coli	33 − 36	FEEI	1.50
4OLE:B	Unknown	Homo sapiens	E. coli	59 − 69	ASTEKKDVLVP	2.52
5G51:A	Viral Protein	Deformed wing virus	E. coli	31 − 36	GSASDQ	1.45
6JF2:A	Structural protein	Salmonella enterica subsp. enterica serovar Typhimurium str. LT2	E. coli	50 − 54	VTAGG	2.00
6QQF:A	Hydrolase	Gallus gallus	Gallus gallus	68 − 75	RTPGSRNL	1.95
7EC8:A	Hydrolase	Uncultured bacterium	E. coli	187 − 190	DGGI	1.35
7R7W:B	Immune system	Homo sapiens	E. coli	46 − 50	IEKVE	1.17

Table A3: **Protein structures used for the evaluation in our experiments.** Note that the residue indices of the altloc sites are given with respect to the first residue in the structure (numbered as 1) and may differ from the author-assigned indices in the original structures.