# CRYOSAMPLER: AMORTIZED ATOMIC MODEL BUILDING ACROSS CRYO-EM CONFORMATIONAL ENSEMBLES

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Cryo-electron microscopy (cryo-EM) increasingly resolves proteins in motion, recovering an ensemble of maps spanning a molecule's conformational landscape. Converting these maps into atomic coordinates is an inverse problem known as model building, a necessary step for every structure deposited in the Protein Data Bank, which in turn supplies the training data for the next generation of AlphaFold-like models. However, existing model building methods treat each map as an independent inverse problem, ignoring that every map in an ensemble is an observation of the same molecule. We ask whether solving this inverse problem jointly across the ensemble with a single amortized model yields better atomic models than per-map inference. In this work, we propose CryoSampler, an amortized framework for ensemble model building that trains a 3D variational autoencoder to map each cryo-EM map to a neural deformation field over a shared reference structure from a frozen Boltz-2 model, supervised end-to-end through a differentiable cryo-EM forward model and stereochemical restraints. We find that amortization acts as an implicit structural prior that produces atomic models of higher geometric quality at lower compute cost, achieving state-of-the-art performance across several experimental datasets. These results establish amortized inference across conformational ensembles as a simple and effective principle for converting heterogeneous cryo-EM maps into high-quality atomic structures.

## 1 INTRODUCTION

Proteins carry out their functions by changing shape. An ion channel opens and closes to let ions cross a cell membrane, and a transporter alternates between inward- and outward-facing states to move molecules across it. Each of these shapes is called a *conformation*, and the set of conformations a protein adopts, together with their relative populations, is its *conformational ensemble*. Static structure predictors such as AlphaFold (Jumper et al., 2021; Abramson et al., 2024) output a single conformation, and predicting ensembles from sequence is widely regarded as the next frontier in computational structural biology (Lane, 2023; Sala et al., 2023; Ourmazd et al., 2022). Emerging ensemble predictors such as AlphaFlow (Jing et al., 2024) and BioEmu (Lewis et al., 2025) draw much of their conformational training data from molecular dynamics simulations, which become costly for large complexes. Just as AlphaFold relied on decades of structures accumulated in the Protein Data Bank (PDB) (Berman et al., 2000), ensemble predictors will depend on large collections of experimentally grounded atomic ensembles, which remain scarce today.

Cryo-electron microscopy (cryo-EM) is a natural source of such data. In a cryo-EM experiment, millions of copies of a molecule are flash-frozen in a thin layer of ice and imaged with an electron microscope; each resulting particle image is a noisy 2D projection of a single molecule in whatever conformation it occupied when frozen. Turning these images into atomic structures involves two inverse problems. The first, *reconstruction*, recovers 3D density maps (grids of the molecule's electron scattering potential) from the particle images, and heterogeneous reconstruction methods recover not one averaged map but an ensemble of maps along the molecule's dominant motions (Punjani & Fleet, 2021; Zhong et al., 2021; Punjani & Fleet, 2023; Levy et al., 2025). The second, *model building*, places every atom of the known sequence into each map so that the resulting atomic model both fits the density and is chemically valid. These problems are central to structure determination



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

and have drawn sustained attention from the machine learning community, both for reconstruction (Zhong et al., 2021; Levy et al., 2022b; Shekarforoush et al., 2024; Levy et al., 2025) and model building (Jamali et al., 2023; 2024; Raghu et al., 2025). Only model building, however, yields the atomic coordinates that are deposited in the PDB and used to train structure predictors.

This division creates a bottleneck for dynamics. Heterogeneous reconstruction now routinely produces map ensembles, but model building methods were designed for one map at a time, and building accurate models into lower-resolution heterogeneous maps remains difficult. As a result, practitioners often deposit a single ensemble-averaged structure, and the conformational variability captured by the experiment rarely reaches the PDB. Accurate *ensemble model building*, which converts an ensemble of maps into an ensemble of atomic models, is therefore a prerequisite for turning cryo-EM data into training data for dynamic structure predictors, and it is the problem we address in this work.

Model building is an underdetermined inverse problem: a density map, often at modest local resolution, must be explained by thousands of atomic coordinates that also satisfy strict stereochemical constraints. Learned de novo methods such as ModelAngelo (Jamali et al., 2023; 2024) trace chains directly from density but produce incomplete models in low-resolution regions. More recent methods use structure predictors as priors, optimizing AlphaFold's input embeddings (Fadini et al., 2026), guiding Boltz's diffusion sampler toward the map (Raghu et al., 2025), or refining coordinates under differentiable geometry restraints (Chen, 2025). When applied to the maps produced by a heterogeneous reconstruction, these map-based methods share one design choice: each map is solved independently. They treat $N$ conformational states as $N$ unrelated problems and must recover physically valid geometry in each state from that state's density alone.

We argue that this per-map formulation discards the most informative structure in heterogeneous cryo-EM data. The maps in an ensemble are not independent observations as they typically arise from the same particle images and depict the same molecule, with identical sequence, chemistry, and topology, differing only by a continuous deformation. Noise and reconstruction artifacts in one state are unlikely to recur in the same form in the others, while valid stereochemistry must hold in every state. Per-map inference cannot exploit either property. In particular, directly optimizing atomic coordinates against each map, even under explicit stereochemical restraints, is free to distort local geometry to explain noise, trading chemical validity for density fit.

To address this gap, we introduce **CryoSampler**, a framework that casts ensemble model building as a single *amortized* inverse problem. Rather than optimizing coordinates per map, CryoSampler trains a 3D variational autoencoder (VAE) whose encoder maps each cryo-EM map to a latent code and whose shared decoder produces a neural deformation field over a single reference structure predicted by a frozen Boltz-2 model (Passaro et al., 2025). The deformed coordinates are rendered through a differentiable cryo-EM forward model and supervised end-to-end against the input maps, together with differentiable stereochemical and backbone restraints. Because every conformational state is explained by the same encoder, decoder, and reference, the network must account for the entire ensemble with one low-dimensional family of smooth deformations. This shared parameterization acts as an implicit structural prior that complements the explicit restraints and yields a set of atomic models with consistent atom correspondence across states (Figure 1). Because CryoSampler operates on reconstructed maps rather than particle images, it is compatible with any heterogeneous reconstruction method that outputs a set of maps. Our main contributions are summarized below.

- We identify per-map independence as a key limitation of existing map-based model building methods when applied to heterogeneous cryo-EM, and formulate *ensemble model building* as a single amortized inverse problem over all maps from a reconstruction.
- We introduce CryoSampler, a 3D VAE that parameterizes each conformational state as a neural deformation field over a shared Boltz-2 reference, trained end-to-end through a differentiable cryo-EM forward model with geometric restraints.
- Across four experimental datasets, CryoSampler achieves the highest or tied-highest map–model fit among current model building methods while maintaining high geometric quality, and degrades gracefully when system-specific map pre-processing is removed.
- We show that amortization acts as an implicit structural prior: compared to direct coordinate optimization, CryoSampler matches map–model fit with $8.5\times$ fewer steric clashes and $20\times$ fewer rotamer outliers.

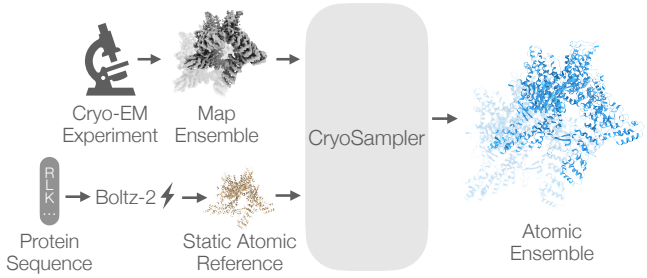


Figure 1: **Overview of CryoSampler.** Heterogeneous cryo-EM reconstruction yields an ensemble of density maps from a single experiment. Rather than building an atomic model into each map independently, CryoSampler solves all maps jointly with a single amortized network that deforms a shared static reference structure from Boltz-2 (Passaro et al., 2025) into an atomic model for every conformational state. The network is trained end-to-end through a differentiable cryo-EM forward model, producing an atomic ensemble with high map–model fit, valid stereochemistry, and consistent atom correspondence across states.

## 2 Related Work

**Atomic model building.** Model building into cryo-EM maps has traditionally been a manual process using interactive tools such as Coot (Emsley et al., 2010) and ISOLDE (Croll, 2018). ModelAngelo (Jamali et al., 2023; 2024) automates de novo building with a network trained on thousands of maps, but often yields incomplete models in ambiguous or low-resolution regions. Recent methods incorporate pre-trained structure predictors as priors: ROCKET (Fadini et al., 2026) optimizes the MSA embeddings of AlphaFold 2, CryoBoltz (Raghu et al., 2025) guides the diffusion sampler of Boltz-1 (Wohlwend et al., 2025) toward a target map, and Chen (2025) refine models under differentiable approximations of MolProbity metrics. These methods take a single map as input and are applied independently to each map of an ensemble. A separate line of work incorporates atomic models into heterogeneous reconstruction from particle images: CryoSTAR (Li et al., 2024) trains a VAE that deforms a reference atomic model to explain conformational heterogeneity, and Chen (2025) generate series of models within a Gaussian-mixture heterogeneity framework. These approaches are tied to a specific particle-level reconstruction pipeline and produce coarse-grained or pipeline-specific models. CryoSampler instead operates on the reconstructed maps themselves, making it agnostic to the upstream reconstruction method, and builds full-atom models with a pre-trained structure predictor as the reference.

**Amortized inference and implicit priors.** Amortized inference trains a shared network to map observations to latent variables rather than optimizing each latent independently (Kingma & Welling, 2014). In cryo-EM, amortization has been used to accelerate pose and heterogeneity inference from particle images (Zhong et al., 2021; Levy et al., 2022a;b; 2025), and deformation-based methods model heterogeneity as a learned warp of a consensus density (Punjani & Fleet, 2023; Schwab et al., 2024). While amortization is usually motivated by speed and is often associated with an accuracy cost relative to per-instance optimization (Cremer et al., 2018), network parameterizations can also act as implicit priors that regularize ill-posed inverse problems (Ulyanov et al., 2018). CryoSampler applies this perspective to atomic model building, where we find that amortizing across an ensemble improves stereochemistry at matched map–model fit.

**Protein ensemble prediction.** Sequence-to-ensemble methods either perturb pre-trained static predictors at inference time (Wayment-Steele et al., 2024; Monteiro da Silva et al., 2024; Suzuki & Amagasa, 2026; Richman et al., 2025) or train on molecular dynamics data (Jing et al., 2024; Lewis et al., 2025); see Sledzieski & Hanson (2026) for a review. Both are limited by the scarcity of experimentally grounded atomic ensembles, which ensemble model building from cryo-EM can help supply.

## 3 Problem Formulation

**Background: from particle images to atomic ensembles.** In single-particle cryo-EM, each particle image is modeled as $I_n = C_n * \Pi_{R_n,t_n}\big(V(\xi_n)\big) + \eta_n$: a tomographic projection $\Pi$ of the molecule's 3D density $V$ at an unknown orientation $R_n \in SO(3)$ and in-plane shift $t_n$, blurred by the microscope's contrast transfer function $C_n$ and corrupted by noise $\eta_n$ (Punjani et al., 2017). Homogeneous reconstruction assumes a single density shared by all images; heterogeneous reconstruction additionally infers a per-image conformational variable $\xi_n$ and returns a set of maps sampled along the learned conformational landscape (Punjani & Fleet, 2021; Zhong et al., 2021). These maps are the input to model building.

**Model building.** Let $\{V_i\}_{i=1}^N$, $V_i \in \mathbb{R}^{R^3}$, denote an ensemble of density maps from a heterogeneous reconstruction of a molecule with known sequence, and let $\mathcal{F}_{\text{EM}} : \mathbb{R}^{A\times 3} \to \mathbb{R}^{R^3}$ be a differentiable forward model that renders $A$ atomic coordinates into a density map. Model building seeks estimated atomic models $\hat{X}_i \in \mathbb{R}^{A\times 3}$ such that the rendered maps $\hat{V}_i = \mathcal{F}_{\text{EM}}(\hat{X}_i)$ match $V_i$ and each $\hat{X}_i$ is stereochemically valid. Because maps have limited local resolution, this inverse problem is underdetermined.

**Per-map inference.** Existing map-based methods solve $N$ independent problems,

$$\hat{X}_i = \arg\min_X \; \mathcal{L}_{\text{recon}}\big(V_i, \mathcal{F}_{\text{EM}}(X)\big) + \mathcal{L}_{\text{structural}}(X), \qquad i = 1, \dots, N, \tag{1}$$

where $\mathcal{L}_{\text{recon}}$ measures map–model agreement and $\mathcal{L}_{\text{structural}}$ encodes stereochemical restraints (both defined in Section 4). The simplest instance, which we call *direct optimization*, initializes $X$ at a reference structure $X_{\text{ref}}$ and optimizes the coordinates by gradient descent.

**Amortized ensemble inference.** We instead parameterize every state through a shared encoder $\mathcal{F}_\phi$, decoder $\mathcal{F}_\theta$, and atomic $\text{MLP}_\psi$ that together map each $V_i$ to atomic offsets $\Delta\hat{X}_i$ applied to a single reference $X_{\text{ref}}$, and solve one joint problem,

$$\min_{\phi,\theta,\psi} \sum_{i=1}^{N} \Big[\mathcal{L}_{\text{recon}}\big(V_i, \mathcal{F}_{\text{EM}}(\hat{X}_i)\big) + \mathcal{L}_{\text{structural}}(\hat{X}_i)\Big], \qquad \hat{X}_i = X_{\text{ref}} + \Delta\hat{X}_i\big(\mathcal{F}_\phi(V_i); \theta, \psi\big). \tag{2}$$

Equation 2 differs from Equation 1 in two ways. First, parameters are shared across states, so evidence from every map constrains the model of every other map. Second, each $\hat{X}_i$ is restricted to the range of the decoder, a smooth deformation field sampled at the atom positions, which acts as an implicit prior on top of $\mathcal{L}_{\text{structural}}$. Per-map amortized fitting ($N = 1$) and direct optimization are therefore natural ablations that remove the first and both of these properties, respectively (Section 5.3).

## 4 Methods

CryoSampler instantiates Equation 2 with a 3D VAE (Figure 2). The weights of Boltz-2 and its reference prediction $X_{\text{ref}}$ are frozen throughout; only the VAE is trained.

**Encoder.** The encoder $\mathcal{F}_\phi$ compresses each input map $V$ into a latent grid distribution $\mathcal{N}(\mu_\phi(V), \sigma^2_\phi(V))$ at $1/8$ of the input spatial resolution using ConvNeXt-style blocks. We sample a latent code with the reparameterization trick, $z = \mu_\phi(V) + \sigma_\phi(V) \odot \epsilon$, $\epsilon \sim \mathcal{N}(0, \mathbf{I})$.

**Deformation decoder.** The decoder $\mathcal{F}_\theta$ expands $z$ into a 64-channel feature grid $Y = \mathcal{F}_\theta(z)$ at the original map resolution. We sample $Y$ at the reference coordinates by trilinear interpolation and pass the per-atom features, together with the reference coordinates, through a lightweight atomic $\text{MLP}_\psi$ that predicts per-atom offsets $\Delta\hat{X}$. The decoder is thus a hybrid neural field whose output is rendered into an estimated map:

$$\hat{X} = X_{\text{ref}} + \Delta\hat{X}, \qquad \Delta\hat{X} = \text{MLP}_\psi\left(X_{\text{ref}}, \text{GridSample}(Y, X_{\text{ref}})\right), \qquad \hat{V} = \mathcal{F}_{\text{EM}}(\hat{X}). \tag{3}$$

**Forward model.** $\mathcal{F}_{\text{EM}}$ renders each atom as an isotropic Gaussian whose width is set by the resolution of the target map (Tang et al., 2007), estimated with `phenix.mtriage` for maps we reconstruct and taken from the deposition otherwise. We do not model the contrast transfer function, which is

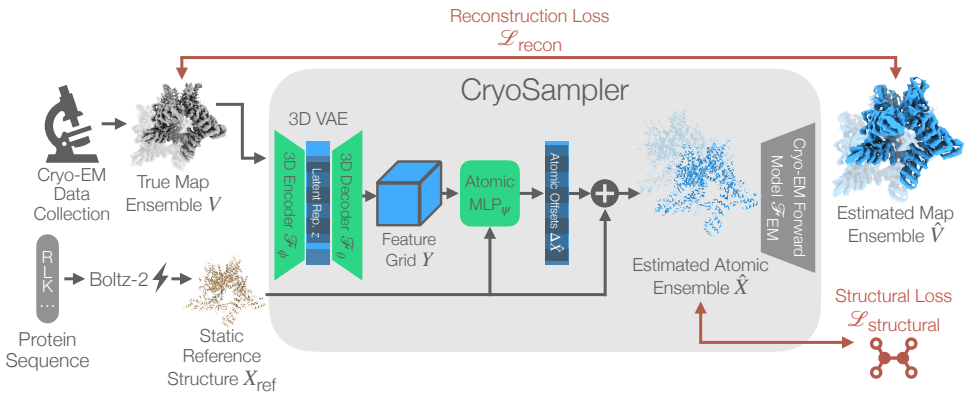


Figure 2: **CryoSampler architecture.** A 3D VAE encodes each input map $V_i$ into a latent grid; a shared decoder produces a feature grid that is sampled at the coordinates of a static Boltz-2 reference $X_{\text{ref}}$ and decoded by an atomic MLP into per-atom offsets. The deformed model $\hat{X}_i$ is rendered by a differentiable cryo-EM forward model into $\hat{V}_i$ and supervised against $V_i$, together with structural losses on $\hat{X}_i$.

corrected during reconstruction, or per-atom B-factors, which are refined for all methods at evaluation time (Section 5.1).

**Objective.** The reconstruction loss is one minus the normalized cross-correlation (NCC) between input and rendered maps,

$$\mathcal{L}_{\text{recon}} = 1 - \text{NCC}(\hat{V}, V), \qquad \text{NCC}(\hat{V}, V) = \frac{\sum_j (\hat{V}_j - \bar{\hat{V}})(V_j - \bar{V})}{\sqrt{\sum_j (\hat{V}_j - \bar{\hat{V}})^2 + \epsilon}\,\sqrt{\sum_j (V_j - \bar{V})^2 + \epsilon} + \epsilon}, \tag{4}$$

where $j$ indexes voxels and $\epsilon = 10^{-8}$. The structural loss $\mathcal{L}_{\text{structural}} = \lambda_M \mathcal{L}_{\text{MolProb}} + \lambda_B \mathcal{L}_{\text{backbone}}$ combines differentiable approximations of MolProbity metrics introduced by Chen (2025), penalizing clashes and Ramachandran and rotamer outliers, with a soft rigid-body backbone loss that preserves local secondary structure of the reference:

$$\mathcal{L}_{\text{backbone}} = \frac{1}{M \cdot |P_B|} \sum_{m=1}^{M} \sum_{(k,l) \in P_B} \left( \|\hat{X}_k^{(m)} - \hat{X}_l^{(m)}\|_2 - \|X_{\text{ref},k}^{(m)} - X_{\text{ref},l}^{(m)}\|_2 \right)^2, \tag{5}$$

where $M = N_{\text{res}} - W + 1$ is the number of sliding windows of $W$ residues and $P_B$ is the set of pairs among the 20 backbone atoms in each window. The full objective, summed over all maps in the ensemble, is

$$\mathcal{L}_{\text{VAE}} = \sum_{i=1}^{N} \Big[ \mathcal{L}_{\text{recon}}(V_i, \hat{V}_i) + \lambda_{\text{KL}} D_{\text{KL}}\big(q(z|V_i) \,\|\, \mathcal{N}(0, \mathbf{I})\big) + \mathcal{L}_{\text{structural}}(\hat{X}_i) \Big], \tag{6}$$

with $\lambda_{\text{KL}} = 10^{-4}$, $\lambda_M = 10^{-6}$, and $\lambda_B = 10^2$ for all systems. After training, the atomic model for each map is obtained by decoding its posterior mean $\mu_\phi(V_i)$.

## 5 Experiments

We first compare CryoSampler against existing model building methods on four experimental cryo-EM systems (Section 5.2). We then isolate the source of its gains through two controlled ablations under identical losses: replacing the amortized VAE with direct coordinate optimization, and replacing joint training with per-map training (Section 5.3). Finally, we evaluate robustness by repeating model building without system-specific map pre-processing (Section 5.4).

### 5.1 Experimental Setup

**Systems.** We evaluate on four experimental systems: the TRPV3 ion channel undergoing a "breathing" motion (Shimada et al., 2020), human P-glycoprotein (P-gp) during drug transport (Culbertson & Liao, 2025), integrin $\alpha V\beta 8$ with a flexible tail (Campbell et al., 2020), and the neurokinin-1 (NK-1) GPCR (Harris et al., 2022). Each ensemble contains four maps. For TRPV3, integrin, and NK-1, we reconstruct the ensemble from raw particles with 3D Variability Analysis (Punjani & Fleet, 2021) and build all four maps jointly. The four P-gp maps were deposited separately in the EMDB rather than obtained from a single heterogeneous reconstruction; we include P-gp in the comparison to existing methods (Table 1) to test whether CryoSampler also outperforms them on separately deposited maps, and we therefore fit each P-gp map individually ($N = 1$) and omit it from the joint versus per-map comparison. The P-gp and NK-1 maps were masked within 5 Å of the deposited atomic models to remove background (Appendix A.5), which uses information from those models and applies to every method. Pre-processing details are given in Appendix A.5.

**Baselines.** We compare against ModelAngelo (Jamali et al., 2023; 2024), CryoBoltz (Raghu et al., 2025), and E2GMM (Chen, 2025), each run per map with default settings. CryoBoltz and E2GMM are initialized from the same Boltz-2 reference as CryoSampler. We also report the unrefined Boltz-2 prediction as a reference point that does not use the map.

**Metrics.** Ground-truth structures are unavailable for experimental maps, so we follow the validation conventions used for PDB depositions, which assess a model jointly by its fit to the map and by its stereochemical quality as summarized by MolProbity (Chen et al., 2010; Liebschner et al., 2019; Lawson et al., 2021). For the same reason, we do not treat deposited models, such as those available for P-gp, as ground truth: each is one expert-built solution that was itself validated by these same criteria, so agreement with it would measure similarity to another model rather than accuracy. Our claim of state-of-the-art performance therefore refers to this combination of map–model fit and geometry across four experimental systems. After rigid alignment and B-factor refinement with `phenix.real_space_refine` (applied identically to every method), we report map–model cross-correlation ($CC_{volume}$, $CC_{mask}$), the MolProbity score, and two metrics that are not optimized by any term in our objective: EMRinger, which scores side-chain placement relative to density, and CaBLAM, which validates backbone geometry from the C$\alpha$ trace. Because cross-correlation is close to our training objective and can be inflated at the expense of geometry (Section 5.3), we always interpret it together with geometry and the independent metrics. Unless noted otherwise, results are mean $\pm$ standard deviation over three training replicates; full metrics and runtimes are in Appendix A.1.

### 5.2 Comparison to Existing Model Building Methods

Table 1 and Figure 3 summarize model building performance. CryoSampler combines the best or tied-best map–model fit ($CC_{volume}$, $CC_{mask}$) with the best or tied-best MolProbity score on three of four systems (TRPV3, NK-1, and integrin); on the fourth, P-gp, it retains the highest map–model fit and places second in geometry. No other method is in the top tier for both fit and geometry on more than one system, so CryoSampler achieves the best overall balance between the two. ModelAngelo is the next-best method on TRPV3 but produces incomplete models (Figure 3) and has the highest MolProbity scores on three of four systems. E2GMM attains the lowest MolProbity scores on three of four systems, driven by low clashscores, but has the lowest map–model fit on three of four systems and roughly 5–7$\times$ higher CaBLAM disfavored rates than CryoSampler. On the independent metrics, CryoSampler's CaBLAM statistics remain close to those of the Boltz-2 reference on all four systems, and its EMRinger scores are in the top tier on TRPV3 and NK-1 (all methods are tied within one s.d. on P-gp). ModelAngelo achieves the highest EMRinger means, with large variance. On the separately deposited P-gp maps, which we fit per map, CryoSampler still achieves the highest $CC_{volume}$ and $CC_{mask}$.

### 5.3 Amortization as an Implicit Structural Prior

**Amortized vs. direct optimization.** CryoSampler's geometric quality comes from its amortized parameterization rather than from its loss functions. To test this, we replace the VAE with direct gradient-based optimization of per-atom coordinates (Equation 1) under the identical reconstruction and structural losses and the same Boltz-2 initialization. Table 2 shows that direct optimization

Table 1: **Model building on four experimental systems.** Map–model fit ($CC$), MolProbity score, and independent validation metrics ($^\dagger$: not optimized by any term of our objective). Mean $\pm$ s.d. over 3 replicates. Among model building methods, bold marks the best mean and all methods within one s.d. of it (a heuristic, not a statistical test); underline marks the next tier under the same rule. Boltz-2 (gray) is the unrefined reference prediction and is not ranked. —: out of memory. $^*$ModelAngelo built models for only 3 of 4 maps and produces incomplete models, which can yield favorable geometry statistics on the residues it builds. Full metrics are in Table 5.

| | Metric | Boltz-2 (ref.) | CryoBoltz | E2GMM | ModelAngelo | Ours |
|---|---|---|---|---|---|---|
| TRPV3 | $CC_{volume}$ (↑) | 0.31±0.03 | — | 0.25±0.02 | 0.79±0.08 | **0.84±0.02** |
| | $CC_{mask}$ (↑) | 0.30±0.03 | — | 0.23±0.02 | 0.79±0.07 | **0.84±0.02** |
| | MolProbity (↓) | 1.83±0.04 | — | **1.57±0.07** | 3.00±0.16 | **1.64±0.04** |
| | EMRinger$^\dagger$ (↑) | 0.08±0.22 | — | 0.01±0.17 | **1.15±0.67** | **1.05±0.32** |
| | CaBLAM disf. (%)$^\dagger$ (↓) | 3.23±0.17 | — | 22.30±1.03 | **3.83±0.72** | **3.87±0.38** |
| NK-1 GPCR | $CC_{volume}$ (↑) | 0.55±0.06 | **0.73±0.05** | **0.75±0.04** | 0.69±0.06 | **0.73±0.04** |
| | $CC_{mask}$ (↑) | 0.55±0.06 | **0.72±0.05** | **0.75±0.04** | 0.70±0.06 | **0.72±0.04** |
| | MolProbity (↓) | 1.87±0.02 | 3.01±0.05 | **1.75±0.20** | 3.61±0.20 | **1.93±0.03** |
| | EMRinger$^\dagger$ (↑) | 0.29±0.19 | 0.47±0.24 | **0.66±0.33** | **0.92±0.31** | **0.69±0.36** |
| | CaBLAM disf. (%)$^\dagger$ (↓) | 3.73±0.45 | 6.52±0.70 | 25.70±3.79 | 5.23±3.60 | **3.80±0.53** |
| Integrin | $CC_{volume}$ (↑) | 0.62±0.07 | 0.38±0.13 | 0.35±0.05 | 0.58±0.03$^*$ | **0.68±0.05** |
| | $CC_{mask}$ (↑) | 0.61±0.06 | 0.39±0.12 | 0.29±0.06 | 0.57±0.02$^*$ | **0.68±0.05** |
| | MolProbity (↓) | 1.80±0.07 | 3.56±0.08 | 1.88±0.09 | 4.03±0.13$^*$ | **1.80±0.07** |
| | EMRinger$^\dagger$ (↑) | 0.33±0.14 | 0.35±0.24 | 0.13±0.14 | **1.97±0.61**$^*$ | 0.35±0.18 |
| | CaBLAM disf. (%)$^\dagger$ (↓) | 5.22±0.19 | 11.00±0.99 | 34.10±1.24 | 13.27±4.94$^*$ | **5.22±0.20** |
| P-gp | $CC_{volume}$ (↑) | 0.39±0.04 | 0.73±0.03 | 0.26±0.02 | 0.59±0.15$^*$ | **0.77±0.02** |
| | $CC_{mask}$ (↑) | 0.40±0.03 | 0.74±0.03 | 0.26±0.03 | 0.59±0.14$^*$ | **0.77±0.02** |
| | MolProbity (↓) | 1.82±0.06 | 2.98±0.41 | **1.60±0.11** | 2.89±1.07$^*$ | 2.67±0.26 |
| | EMRinger$^\dagger$ (↑) | 0.31±0.27 | **0.15±0.28** | **0.31±0.27** | **0.89±2.76**$^*$ | **0.09±0.37** |
| | CaBLAM disf. (%)$^\dagger$ (↓) | 2.85±0.23 | 4.86±0.73 | 22.31±1.99 | **0.00±0.00**$^*$ | 3.23±0.31 |

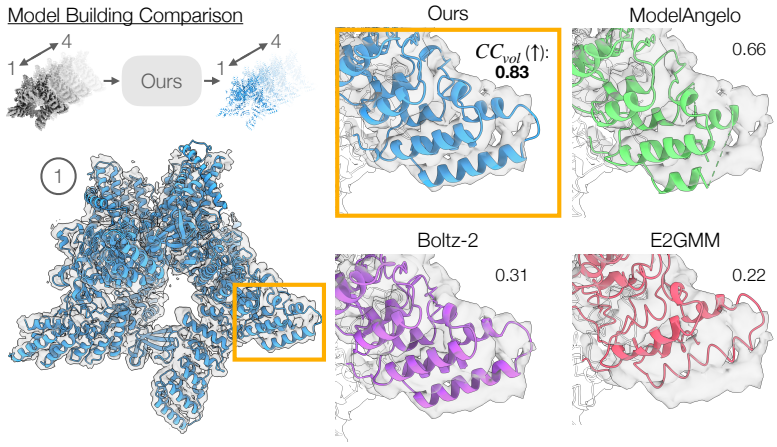


Figure 3: **Qualitative model building on TRPV3.** The experimental map (one of four states) is shown in transparent gray with $CC_{volume}$ for each method. CryoSampler, run on all four maps jointly, achieves the best map–model fit. ModelAngelo achieves the second-highest fit but leaves regions of the map unmodeled (right of inset); Boltz-2 and E2GMM output complete but poorly fitting structures.

matches CryoSampler's map–model fit but produces an $8.5\times$ higher clashscore (168.7 vs. 19.9) and $20\times$ more rotamer outliers (3.97% vs. 0.19%), nearly doubling the MolProbity score (3.67 vs. 1.90). Because the losses are shared, this gap is attributable to the parameterization: when every coordinate is free, the optimizer distorts local geometry to explain density, whereas CryoSampler's shared deformation decoder restricts solutions to smooth, low-dimensional deformations of the reference.

Table 2: **Amortized vs. direct optimization.** Identical losses and Boltz-2 initialization. Mean ± s.d. pooled over TRPV3, integrin, and NK-1 (12 states); separate replicate set from Table 1. Bold: best mean and all entries within one s.d.; underline: next tier; Boltz-2 (gray) is not ranked.

| | $CC_{mask}$ (↑) | Rama. out. (%, ↓) | Rota. out. (%, ↓) | Clashscore (↓) | MolProbity (↓) |
|---|---|---|---|---|---|
| Boltz-2 | 0.50±0.14 | 0.07±0.07 | 0.05±0.04 | 20.3±0.8 | 1.81±0.01 |
| Direct Opt. | **0.76±0.03** | **0.09±0.08** | 3.97±0.49 | 168.7±29.1 | 3.67±0.10 |
| Ours | **0.76±0.06** | **0.08±0.04** | **0.19±0.16** | **19.9±3.7** | **1.90±0.07** |

Table 3: **Joint ($N$=4) vs. per-map ($N$=1) training** with identical losses. Mean ± s.d. across the 4 conformational states of each system (s.d. therefore includes state-to-state variation in map quality); runtime is total GPU-hours. Integrin and NK-1 joint models are replicate 0 of Table 1, re-aligned and re-refined together with the per-map models in a single evaluation run; TRPV3 is a separate replicate set. Because alignment and refinement are repeated, absolute values differ from Table 1 by more than its replicate s.d. (e.g., NK-1 joint MolProbity 1.81 vs. 1.93). Bold/underline as above, per system. MolProbity components (Ramachandran, rotamer, clashscore) are in Table 6.

| | TRPV3 | | NK-1 GPCR | | Integrin | |
|---|---|---|---|---|---|---|
| | Per-map | Joint | Per-map | Joint | Per-map | Joint |
| $CC_{volume}$ (↑) | **0.850±0.014** | **0.842±0.017** | **0.749±0.040** | **0.735±0.043** | **0.722±0.047** | **0.692±0.056** |
| $CC_{mask}$ (↑) | **0.849±0.014** | **0.840±0.017** | **0.744±0.043** | **0.729±0.046** | **0.721±0.039** | **0.692±0.049** |
| MolProbity (↓) | 1.91±0.08 | **1.82±0.01** | 2.05±0.06 | **1.81±0.03** | **1.83±0.01** | **1.83±0.01** |
| Runtime (GPU-h, ↓) | 2.83 | **2.29** | 2.10 | **1.63** | 1.66 | **1.27** |

Amortization thus acts as an implicit structural prior, matching the fit of direct optimization at roughly half its MolProbity score.

**Joint vs. per-map training.** We next train the same VAE either jointly on all four maps of a system or separately on each map ($N = 1$) with the same losses, on TRPV3, NK-1, and integrin (Table 3). Joint training makes CryoSampler faster on every system and improves its geometry where per-map fitting degrades it, at a small cost in map–model fit. Per-map training achieves slightly higher map–model fit on all three systems ($CC_{volume}$ higher by 0.008–0.030 and $CC_{mask}$ by 0.009–0.029), within the across-state s.d. but consistently so: per-map $CC_{volume}$ and $CC_{mask}$ are higher in every one of the four states on all three systems. In exchange, joint training regularizes geometry whereas per-map fitting distorts it. On TRPV3, joint training lowers the MolProbity score from 1.91 to 1.82 in all four states and yields more uniform geometry across states (clashscore s.d. 1.81 vs. 0.10; Table 6). On NK-1, joint training lowers the MolProbity score from 2.05 to 1.81 in all four states, mainly through fewer clashes. On integrin, where per-map fitting already preserves the reference geometry, MolProbity scores are identical (1.83) and joint training provides no geometric benefit. On all three systems, joint training is also 1.24–1.31× faster, because all states are batched through a single network. Joint training thus trades a small amount of density fit for faster training and, where needed, additional regularization, while the dominant regularizing effect across our experiments comes from amortization itself (Table 2).

**Loss ablations.** Both of CryoSampler's structural losses are necessary for valid geometry. Removing the structural loss raises $CC_{volume}$ (e.g., 0.84 to 0.93 on TRPV3) but produces invalid structures, with Ramachandran outliers above 50% on TRPV3 and P-gp (Table 5, Appendix). Removing the backbone loss slightly increases map–model fit while more than doubling clashscore, and removing the KL term has no measurable effect on model building metrics (Appendix A.3). The explicit restraints are therefore essential complements to the implicit prior of amortization.

### 5.4 Robustness to Map Pre-processing

Our main experiments apply system-specific pre-processing: masking density of chains absent from the Boltz-2 reference for integrin, truncating 25 N-terminal residues of one chain in the NK-1 reference, and masking the P-gp and NK-1 maps around the deposited models. We re-ran the model building methods with the integrin masking and the NK-1 truncation removed (Table 4); the NK-1 map mask is retained, so this experiment does not remove all use of deposited models. CryoSampler

Table 4: **Model building without system-specific pre-processing.** For NK-1, we remove the 25-residue truncation of the Boltz-2 reference but retain the 5 Å map mask around the deposited model; for integrin, we remove both the exclusion of chains E and F from the reference and the masking and cropping of the map. ModelAngelo does not use a Boltz-2 reference, so the NK-1 change does not apply to it, and we report it only for the unmasked integrin map. Mean ± s.d. across the 4 states. Bold/underline as in Table 1. Boltz-2 (gray) is the unrefined reference and is not ranked.

| *NK-1 GPCR (untruncated Boltz-2 reference)* | | | | |
|---|---|---|---|---|
| Metric | Boltz-2 | CryoBoltz | E2GMM | Ours |
| $CC_{volume}$ (↑) | 0.57±0.06 | 0.69±0.04 | **0.76±0.04** | **0.72±0.04** |
| $CC_{mask}$ (↑) | 0.57±0.06 | 0.69±0.04 | **0.75±0.04** | **0.72±0.04** |
| Rama. outliers (%, ↓) | 0.00±0.00 | 0.91±0.06 | 2.69±1.04 | **0.00±0.00** |
| Rota. outliers (%, ↓) | 0.00±0.00 | 0.97±0.10 | 0.55±0.25 | **0.25±0.08** |
| Clashscore (↓) | 21.1±0.1 | 172.5±13.2 | **4.4±1.3** | 22.3±1.1 |
| MolProbity (↓) | 1.82±0.00 | 3.00±0.03 | 2.06±0.15 | **1.84±0.02** |
| *Integrin $\alpha V\beta 8$ (untruncated reference, unmasked map)* | | | | |
| Metric | Boltz-2 | E2GMM | ModelAngelo | Ours |
| $CC_{volume}$ (↑) | 0.29±0.16 | **0.40±0.14** | **0.54±0.21** | **0.54±0.02** |
| $CC_{mask}$ (↑) | 0.25±0.17 | **0.34±0.16** | **0.53±0.20** | **0.52±0.02** |
| Rama. outliers (%, ↓) | 0.05±0.00 | 5.19±0.32 | 8.15±2.73 | **0.38±0.02** |
| Rota. outliers (%, ↓) | 0.05±0.00 | 2.22±0.26 | 10.10±1.57 | **2.08±0.02** |
| Clashscore (↓) | 26.4±0.2 | **21.1±3.9** | 95.2±20.2 | 81.9±0.2 |
| MolProbity (↓) | 1.91±0.00 | 3.05±0.11 | 3.89±0.10 | **2.89±0.01** |

retains the best combination of fit and geometry among model building methods without this pre-processing. On NK-1, every method's fit and geometry remain within or near their original tiers; CryoSampler remains tied within one s.d. for the highest $CC_{volume}$ and $CC_{mask}$ and has the lowest MolProbity score among model building methods (1.84 vs. 2.06 for E2GMM). On the unmasked integrin map, map–model fit drops for most methods; CryoSampler retains the highest or tied-highest $CC_{volume}$ and $CC_{mask}$ and the lowest Ramachandran and rotamer outlier rates and MolProbity score among model building methods (2.89 vs. 3.05 for E2GMM and 3.89 for ModelAngelo). Its clashscore, however, rises from 19 to 82 as the model is pulled toward unmodeled density. Pre-processing thus benefits most methods, but CryoSampler degrades gracefully and remains the best-balanced method without it.

## 6 Conclusion

We introduce CryoSampler, which reframes atomic model building for heterogeneous cryo-EM as a single amortized inverse problem over an ensemble of maps. CryoSampler matches the map–model fit of direct coordinate optimization while avoiding its stereochemical failures, and achieves the highest or tied-highest map–model fit among current model building methods on four experimental systems. These results suggest that, for underdetermined inverse problems over related observations, amortization not only accelerates inference but also serves as a structural prior. Because CryoSampler can process maps from any heterogeneous reconstruction method, it offers a path toward depositing experimentally grounded atomic ensembles at scale.

Future work could use the same differentiable forward model to train structure predictors end-to-end on raw cryo-EM maps, learning conformational variability directly from experimental measurements, and to fuse complementary modalities such as crystallographic, NMR, and cross-linking data into a single framework for ensemble structure determination. Such models would move structural biology from static snapshots toward the moving machines that carry out biological function.

**Limitations.** CryoSampler deforms a single reference structure, so it cannot correct topological errors in the Boltz-2 prediction or model chains absent from it, and it currently relies on system-specific masking when maps contain unmodeled density. Our evaluation covers four systems with four states each, and joint training offers a system-dependent trade-off between map–model fit and geometry rather than a uniform improvement. Broader community-wide benchmarks of heterogeneous cryo-EM ensembles would strengthen these conclusions.

## AI USE STATEMENT

In this work, we used generative AI tools to assist with polishing portions of the manuscript text, facilitating the running of experiments, and understanding technical concepts. All authors reviewed the AI-assisted text, and all reported numerical results were produced by our own experimental pipeline and verified by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

## REPRODUCIBILITY STATEMENT

The CryoSampler architecture and training objective are described in Section 4, with loss weights and hyperparameters stated in the main text. Data sources (EMPIAR and EMDB accession codes), reconstruction settings, and all map pre-processing steps are detailed in Appendix A.5, and baseline configurations and evaluation protocols (Phenix commands and settings) are given in Appendix A.6. Anonymized code and processed input maps are provided in the supplementary material.

## 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.

Miro A Astore, David Silva-Sánchez, Robert Blackwell, Pilar Cossio, and Sonya M Hanson. Comparing cryo-EM methods and molecular dynamics simulation to investigate heterogeneity in ligand-bound TRPV1. *Nature Communications*, 2025.

Ahmet Bakan, Lidio M Meireles, and Ivet Bahar. Prody: protein dynamics inferred from theory and experiments. *Bioinformatics*, 27(11):1575–1577, 2011.

Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1): 235–242, 2000.

Melody G Campbell, Anthony Cormier, Saburo Ito, Robert I Seed, Andrew J Bondesson, Jianlong Lou, James D Marks, Jody L Baron, Yifan Cheng, and Stephen L Nishimura. Cryo-em reveals integrin-mediated tgf-$\beta$ activation without release from latent tgf-$\beta$. *Cell*, 180(3):490–501, 2020.

Muyuan Chen. Building molecular model series from heterogeneous cryoem structures using gaussian mixture models and deep neural networks. *Communications Biology*, 8(1):798, 2025.

Vincent B Chen, W Bryan Arendall, Jeffrey J Headd, Daniel A Keedy, Robert M Immormino, Gary J Kapral, Laura W Murray, Jane S Richardson, and David C Richardson. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D*, 66(1): 12–21, 2010.

Chris Cremer, Xuechen Li, and David Duvenaud. Inference suboptimality in variational autoencoders. In *International Conference on Machine Learning*, 2018.

Tristan Ian Croll. *ISOLDE*: a physically realistic environment for model building into low-resolution electron-density maps. *Acta Crystallographica Section D*, 74(6):519–530, Jun 2018. doi: 10.1107/S2059798318002425. URL https://doi.org/10.1107/S2059798318002425.

Alan T Culbertson and Maofu Liao. Cryo-em of human p-glycoprotein reveals an intermediate occluded conformation during active drug transport. *Nature Communications*, 16(1):3619, 2025.

Shangyu Dang, Mark K van Goor, Daniel Asarnow, YongQiang Wang, David Julius, Yifan Cheng, and Jenny van der Wijst. Structural insight into trpv5 channel function and modulation. *Proceedings of the National Academy of Sciences*, 116(18):8869–8878, 2019.

P. Emsley, B. Lohkamp, W. G. Scott, and K. Cowtan. Features and development of *Coot*. *Acta Crystallographica Section D*, 66(4):486–501, Apr 2010. doi: 10.1107/S0907444910007493. URL https://doi.org/10.1107/S0907444910007493.

Alisia Fadini, Minhuan Li, Airlie J. McCoy, Suresh Banjara, Hiroki Okumura, Eve Napier, Pietro Fontana, Amir R. Khan, Luca Jovine, Thomas C. Terwilliger, Randy J. Read, Doeke R. Hekstra, and Mohammed AlQuraishi. Alphafold as a prior: experimental structure determination conditioned on a pretrained neural network. *Nature Methods*, April 2026. ISSN 1548-7105. doi: 10.1038/s41592-026-03047-4. URL http://dx.doi.org/10.1038/s41592-026-03047-4.

Julian A Harris, Bryan Faust, Arisbel B Gondin, Marc André Dämgen, Carl-Mikael Suomivuori, Nicholas A Veldhuis, Yifan Cheng, Ron O Dror, David M Thal, and Aashish Manglik. Selective g protein signaling driven by substance p–neurokinin receptor dynamics. *Nature chemical biology*, 18(1):109–115, 2022.

Kiarash Jamali, Dari Kimanius, and Sjors H W Scheres. A graph neural network approach to automated model building in cryo-EM maps. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=65XDF_nwI61.

Kiarash Jamali, Lukas Käll, Rui Zhang, Alan Brown, Dari Kimanius, and Sjors HW Scheres. Automated model building and protein identification in cryo-em maps. *Nature*, 628(8007):450–457, 2024.

Bowen Jing, Bonnie Berger, and Tommi Jaakkola. Alphafold meets flow matching for generating protein ensembles. In *Forty-first International Conference on Machine Learning*, 2024.

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.

Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.

Thomas J Lane. Protein structure prediction has reached the single-structure frontier. *Nature Methods*, 20(2):170–173, 2023.

Catherine L Lawson, Andriy Kryshtafovych, Paul D Adams, Pavel V Afonine, Matthew L Baker, Benjamin A Barad, Paul Bond, Tom Burnley, Renzhi Cao, Jianlin Cheng, et al. Cryo-EM model validation recommendations based on outcomes of the 2019 EMDataResource challenge. *Nature Methods*, 18(2):156–164, 2021.

Axel Levy, Frédéric Poitevin, Julien Martel, Youssef Nashed, Ariana Peck, Nina Miolane, Daniel Ratner, Mike Dunne, and Gordon Wetzstein. CryoAI: Amortized inference of poses for ab initio reconstruction of 3D molecular volumes from real cryo-EM images. In *European Conference on Computer Vision*, 2022a.

Axel Levy, Gordon Wetzstein, Julien Martel, Frédéric Poitevin, and Ellen D Zhong. Amortized inference for heterogeneous reconstruction in cryo-EM. In *Advances in Neural Information Processing Systems*, 2022b.

Axel Levy, Rishwanth Raghu, Ryan Feathers, Michal Grzadkowski, Frederic Poitevin, Jake D Johnston, Francesca Vallese, Oliver B Clarke, Gordon Wetzstein, and Ellen D Zhong. Cryodrgn-ai: neural ab initio reconstruction of challenging cryo-em and cryo-et datasets. *Nature Methods*, jun 2025. doi: 10.1038/s41592-025-02720-4. URL nature.com/articles/s41592-025-02720-4.

Sarah Lewis, Tim Hempel, José Jiménez-Luna, Michael Gastegger, Yu Xie, Andrew YK Foong, Victor García Satorras, Osama Abdin, Bastiaan S Veeling, Iryna Zaporozhets, et al. Scalable emulation of protein equilibrium ensembles with generative deep learning. *Science*, pp. eadv9817, 2025.

Yilai Li, Yi Zhou, Jing Yuan, Fei Ye, and Quanquan Gu. CryoSTAR: leveraging structural priors and constraints for cryo-EM heterogeneous reconstruction. *Nature Methods*, 21:2318–2326, 2024.

Dorothee Liebschner, Pavel V Afonine, Matthew L Baker, Gábor Bunkóczi, Vincent B Chen, Tristan I Croll, Bradley Hintze, L-W Hung, Swati Jain, Airlie J McCoy, et al. Macromolecular structure determination using x-rays, neutrons and electrons: recent developments in phenix. *Biological Crystallography*, 75(10):861–877, 2019.

Elaine C Meng, Thomas D Goddard, Eric F Pettersen, Greg S Couch, Zach J Pearson, John H Morris, and Thomas E Ferrin. Ucsf chimerax: Tools for structure building and analysis. *Protein Science*, 32(11):e4792, 2023.

Gabriel Monteiro da Silva, Jennifer Y Cui, David C Dalgarno, George P Lisi, and Brenda M Rubenstein. High-throughput prediction of protein conformational distributions with subsampled alphafold2. *nature communications*, 15(1):2464, 2024.

Abbas Ourmazd, Keith Moffat, and Eaton Edward Lattman. Structural biology is solved—now what? *Nature methods*, 19(1):24–26, 2022.

Saro Passaro, Gabriele Corso, Jeremy Wohlwend, Mateo Reveiz, Stephan Thaler, Vignesh Ram Somnath, Noah Getz, Tally Portnoi, Julien Roy, Hannes Stark, et al. Boltz-2: Towards accurate and efficient binding affinity prediction. *BioRxiv*, 2025.

Ali Punjani and David J Fleet. 3d variability analysis: Resolving continuous flexibility and discrete heterogeneity from single particle cryo-em. *Journal of structural biology*, 213(2):107702, 2021.

Ali Punjani and David J Fleet. 3DFlex: determining structure and motion of flexible proteins from cryo-EM. *Nature Methods*, 20:860–870, 2023.

Ali Punjani, John L Rubinstein, David J Fleet, and Marcus A Brubaker. cryosparc: algorithms for rapid unsupervised cryo-em structure determination. *Nature methods*, 14(3):290–296, 2017.

Rishwanth Raghu, Axel Levy, Gordon Wetzstein, and Ellen D Zhong. Multiscale guidance of protein structure prediction with heterogeneous cryo-EM data. In *Advances in Neural Information Processing Systems*, 2025.

Daniel D Richman, Jessica Karaguesian, Carl-Mikael Suomivuori, and Ron O Dror. Unlocking hidden biomolecular conformational landscapes in diffusion models at inference time. *39th Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

Khondamir R Rustamov and Artyom Y Baev. MSA clustering enhances AF-Multimer's ability to predict conformational landscapes of protein–protein interactions. *Bioinformatics Advances*, 5(1): vbae197, 2024.

Davide Sala, Felix Engelberger, Hassane S Mchaourab, and Jens Meiler. Modeling conformational states of proteins with alphafold. *Current Opinion in Structural Biology*, 81:102645, 2023.

Johannes Schwab, Dari Kimanius, Alister Burt, Tom Dendooven, and Sjors H W Scheres. DynaMight: estimating molecular motions with improved reconstruction from cryo-EM images. *Nature Methods*, 21:1855–1862, 2024.

Shayan Shekarforoush, David B Lindell, Marcus A Brubaker, and David J Fleet. CryoSPIN: Improving ab-initio cryo-EM reconstruction with semi-amortized pose inference. In *Advances in Neural Information Processing Systems*, 2024.

Hiroto Shimada, Tsukasa Kusakizako, TH Dung Nguyen, Tomohiro Nishizawa, Tomoya Hino, Makoto Tominaga, and Osamu Nureki. The structure of lipid nanodisc-reconstituted trpv3 reveals the gating mechanism. *Nature Structural & Molecular Biology*, 27(7):645–652, 2020.

Samuel Sledzieski and Sonya M Hanson. The landscape of machine learning approaches for modeling protein conformational ensembles. *Current Opinion in Structural Biology*, 98:103253, 2026.

Shosuke Suzuki and Toshiyuki Amagasa. Steering conformational sampling in boltz-2 via pair representation scaling. *bioRxiv*, pp. 2026–01, 2026.

Guang Tang, Liwei Peng, Philip R Baldwin, Deepinder S Mann, Wen Jiang, Ian Rees, and Steven J Ludtke. Eman2: an extensible image processing suite for electron microscopy. *Journal of structural biology*, 157(1):38–46, 2007.

Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

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.

Jeremy Wohlwend, Gabriele Corso, Saro Passaro, Noah Getz, Mateo Reveiz, Ken Leidal, Wojtek Swiderski, Liam Atkinson, Tally Portnoi, Itamar Chinn, et al. Boltz-1 democratizing biomolecular interaction modeling. *BioRxiv*, pp. 2024–11, 2025.

Ellen D Zhong, Tristan Bepler, Bonnie Berger, and Joseph H Davis. Cryodrgn: reconstruction of heterogeneous cryo-em structures using neural networks. *Nature methods*, 18(2):176–185, 2021.

# A Appendix

## A.1 Full Model Building Metrics, Runtime, and Structural Loss Ablation

We provide a complete version of the model building method comparison in Table 5, including additional information about all model geometry metrics (Ramachandran outliers, rotamer outliers, and clash scores), method runtimes, and an ablation of our structural loss. On top of the results shown in the main paper, we find that our method preserves valid stereochemistry Ramachandran and rotamer outliers compared to the unguided output of Boltz-2, and outputs structures with higher clashscores than Boltz-2 and E2GMM on some systems, most notably P-gp, in exchange for higher map–model fit. Runtimes are reported in terms of the average time taken to complete model building per map in the map series: since our method is trained jointly against all four maps per protein system (apart from the P-glycoprotein), we report the single map-averaged runtime in the table, finding that our method is slower than CryoBoltz and ModelAngelo on most systems but substantially faster than E2GMM.

The ablation results in the rightmost column highlight the importance of the structural loss $\mathcal{L}_{\text{structural}}$: although removing this loss significantly improves map–model fit, it does so at the cost of poor stereochemical metrics, producing geometrically invalid structures. This experiment shows the importance of our structural loss at regularizing model building to output chemically valid conformations.

## A.2 Joint vs. Per-map Training: Full Geometry Breakdown

Table 6 expands the MolProbity score in Table 3 into its components. On NK-1, the lower MolProbity score of joint training comes from a lower clashscore (20.2 vs. 25.2, lower in all four states), partly offset by more rotamer outliers (0.75% vs. 0.40%). On integrin, joint training has slightly more rotamer outliers (0.15% vs. 0.07%) with unchanged clashscore and Ramachandran statistics.

## A.3 Backbone and KL Loss Ablations

Table 7 ablates the backbone loss and KL term on TRPV3. Removing the backbone loss slightly increases map–model fit but more than doubles clashscore and raises rotamer and CaBLAM outlier rates. Removing the KL term has no measurable effect on model building metrics; we retain it because the exploratory transfer experiments (Appendix A.4) train a flow matching model on the latent space, which the KL term keeps close to a standard normal.

## A.4 Exploratory: In-Family Conformational Transfer

Because CryoSampler's decoder produces a deformation field that is queried at reference atom positions, the same decoder can in principle be applied to the Boltz-2 reference of a different protein. We explore whether conformational variability learned from one protein's map ensemble transfers to related proteins without any cryo-EM data for the target. To sample new conformations, we add a latent flow matching model trained on the VAE's latent codes.

### A.4.1 Latent Flow Matching Model

We train a flow matching model on the latent embeddings produced by the VAE of Section 4. The weights of the VAE are frozen, and the mean latent codes $\mu$ for all conformational states in the training set of cryo-EM maps are extracted. This stage of training employs a time-conditioned $\text{MLP}_\beta$ that operates on the compressed $V/8$ latent grid, optimized with a continuous-time flow matching objective. The VAE latent grid of shape $(C \times (V/8)^3)$ is flattened into a single vector, which is concatenated with a sinusoidal embedding of the continuous timestep $t$ before being passed through four fully-connected layers of width $d$ with SiLU activations and a linear readout. The model learns a vector field that smoothly interpolates between a standard normal prior distribution $z_0 \sim \mathcal{N}(0, \mathbf{I})$ and the target distribution of latent codes $z_1 = \mu$. The linear interpolation between these two distributions can be written as:

$$z_t = tz_1 + (1-t)z_0, \quad t \in [0,1] \tag{7}$$

Table 5: Model building performance across systems, with full model geometry metrics, method runtimes, and an ablation study on our method's structural loss ($\mathcal{L}_{\text{structural}}$), compared against Boltz-2 (Passaro et al., 2025), CryoBoltz (Raghu et al., 2025), E2GMM (Chen, 2025), and ModelAngelo (Jamali et al., 2023; 2024). Our method achieves competitive Ramachandran outlier and rotamer outlier scores, with slightly higher clash scores than other methods in some cases. While our method takes slightly longer than CryoBoltz and ModelAngelo, it achieves higher map–model fit as a result. The ablation study, reported in the rightmost column, demonstrates the importance of the structural loss at producing regularized atomic structures. *Method only produced structures for three out of the four maps.

| | Metric | Boltz-2 | CryoBoltz | E2GMM | ModelAngelo | Ours | Ours w/o Structural Loss |
|---|---|---|---|---|---|---|---|
| TRPV3 | *Map–Model Fit* | | | | | | |
| | $CC_{volume}$ (↑) | 0.31±0.03 | — | 0.25±0.02 | 0.79±0.08 | 0.84±0.02 | 0.93±0.01 |
| | $CC_{mask}$ (↑) | 0.30±0.03 | — | 0.23±0.02 | 0.79±0.07 | 0.84±0.02 | 0.93±0.01 |
| | $CC_{peaks}$ (↑) | 0.29±0.03 | — | 0.21±0.02 | 0.71±0.08 | 0.79±0.02 | 0.95±0.01 |
| | $CC_{box}$ (↑) | 0.62±0.02 | — | 0.55±0.01 | 0.85±0.04 | 0.89±0.01 | 0.95±0.01 |
| | *Model Geometry* | | | | | | |
| | Rama. Outlier (%, ↓) | 0.00±0.00 | — | 1.37±0.23 | 1.92±0.65 | 0.00±0.00 | 55.94±1.94 |
| | Rota. Outlier (%, ↓) | 0.06±0.08 | — | 1.17±0.19 | 8.30±1.15 | 0.16±0.08 | 64.64±0.95 |
| | Clash Score (↓) | 21±2 | — | 3±0 | 43±6 | 12±1 | 1661±17 |
| | Runtime (h:mm) | — | — | 5:39±0:12 | 0:07±0:01 | 0:33±0:00 | — |
| P-glycoprotein | *Map–Model Fit* | | | | | | |
| | $CC_{volume}$ (↑) | 0.39±0.04 | 0.73±0.03 | 0.26±0.02 | 0.59±0.15* | 0.77±0.02 | 0.96±0.01* |
| | $CC_{mask}$ (↑) | 0.40±0.03 | 0.74±0.03 | 0.26±0.03 | 0.59±0.14 | 0.77±0.02 | 0.96±0.01 |
| | $CC_{peaks}$ (↑) | 0.53±0.12 | 0.70±0.03 | 0.44±0.13 | -0.11±0.16 | 0.75±0.05 | 0.97±0.01 |
| | $CC_{box}$ (↑) | 0.62±0.07 | 0.85±0.02 | 0.54±0.08 | 0.33±0.09 | 0.85±0.02 | 0.96±0.01 |
| | *Model Geometry* | | | | | | |
| | Rama. Outlier (%, ↓) | 0.05±0.04 | 3.72±6.02 | 1.29±0.34 | 2.13±3.80 | 0.11±0.07 | 62.50±14.87 |
| | Rota. Outlier (%, ↓) | 0.01±0.02 | 0.19±0.17 | 0.62±0.17 | 14.27±8.21 | 1.16±0.72 | 70.09±1.11 |
| | Clash Score (↓) | 21±4 | 152±37 | 4±1 | 70±43 | 109±44 | 2224±175 |
| | Runtime (h:mm) | — | 0:15±0:01 | 3:39±0:05 | 0:06±0:00 | 0:46±0:00 | — |
| Integrin $\alpha V\beta 8$ | *Map–Model Fit* | | | | | | |
| | $CC_{volume}$ (↑) | 0.62±0.07 | 0.38±0.13 | 0.35±0.05 | 0.58±0.03* | 0.68±0.05 | 0.74±0.04 |
| | $CC_{mask}$ (↑) | 0.61±0.06 | 0.39±0.12 | 0.29±0.06 | 0.57±0.02 | 0.68±0.05 | 0.75±0.04 |
| | $CC_{peaks}$ (↑) | 0.75±0.04 | 0.72±0.05 | 0.59±0.04 | -0.22±0.03 | 0.81±0.03 | 0.84±0.03 |
| | $CC_{box}$ (↑) | 0.75±0.04 | 0.72±0.05 | 0.59±0.04 | 0.28±0.02 | 0.81±0.03 | 0.84±0.03 |
| | *Model Geometry* | | | | | | |
| | Rama. Outlier (%, ↓) | 0.02±0.03 | 1.64±0.46 | 2.26±0.43 | 16.59±2.73 | 0.02±0.03 | 1.01±0.76 |
| | Rota. Outlier (%, ↓) | 0.12±0.04 | 1.08±0.29 | 1.17±0.28 | 13.07±0.52 | 0.12±0.04 | 9.34±0.85 |
| | Clash Score (↓) | 19±3 | 419±25 | 6±1 | 97±19 | 19±3 | 220±46 |
| | Runtime (h:mm) | — | 0:21±0:00 | 4:16±0:05 | 0:07±0:00 | 0:19±0:00 | — |
| Neurokinin-1 GPCR | *Map–Model Fit* | | | | | | |
| | $CC_{volume}$ (↑) | 0.55±0.06 | 0.73±0.05 | 0.75±0.04 | 0.69±0.06 | 0.73±0.04 | 0.80±0.03 |
| | $CC_{mask}$ (↑) | 0.55±0.06 | 0.72±0.05 | 0.75±0.04 | 0.70±0.06 | 0.72±0.04 | 0.79±0.02 |
| | $CC_{peaks}$ (↑) | 0.55±0.10 | 0.70±0.05 | 0.74±0.07 | 0.29±0.30 | 0.71±0.08 | 0.87±0.02 |
| | $CC_{box}$ (↑) | 0.75±0.04 | 0.84±0.03 | 0.85±0.02 | 0.59±0.22 | 0.83±0.03 | 0.86±0.02 |
| | *Model Geometry* | | | | | | |
| | Rama. Outlier (%, ↓) | 0.00±0.00 | 0.84±0.22 | 2.31±0.74 | 5.15±2.03 | 0.00±0.00 | 24.97±2.14 |
| | Rota. Outlier (%, ↓) | 0.13±0.18 | 0.92±0.24 | 0.30±0.27 | 14.51±1.33 | 0.39±0.28 | 43.78±0.94 |
| | Clash Score (↓) | 22±2 | 157±13 | 6±2 | 80±17 | 23±3 | 872±29 |
| | Runtime (h:mm) | — | 0:10±0:01 | 3:40±0:07 | 0:06±0:01 | 0:24±0:00 | — |

The target vector field is the straight line vector from the pure noise sample to the latent sample:

$$u_t = z_1 - z_0 \tag{8}$$

During optimization, the flow matching objective aims to optimize the weights of the velocity network $\text{MLP}_\beta$ to match the target vector field, over all possible samples $z_0, z_1$ and interpolation times $t$:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{t,z_0,z_1}\left[\|\text{MLP}_\beta(z_t, t) - u_t\|^2\right] \tag{9}$$

Table 6: Components of the MolProbity score for joint ($N$=4) vs. per-map ($N$=1) training (mean ± s.d. across 4 states). Bold: best mean and all entries within one s.d.; underline: next tier, per system.

| | TRPV3 | | NK-1 GPCR | | Integrin | |
|---|---|---|---|---|---|---|
| | Per-map | Joint | Per-map | Joint | Per-map | Joint |
| Rama. out. (%, ↓) | 0.14±0.11 | **0.10±0.02** | **0.00±0.00** | **0.00±0.00** | **0.06±0.00** | **0.06±0.00** |
| Rota. out. (%, ↓) | **0.36±0.11** | **0.44±0.10** | **0.40±0.18** | 0.75±0.14 | **0.07±0.00** | 0.15±0.00 |
| Clashscore (↓) | 14.85±1.81 | **13.95±0.10** | 25.18±4.01 | **20.18±1.30** | **21.40±0.75** | **21.51±0.62** |

Table 7: Loss ablations on TRPV3 (mean ± s.d.). Bold: best mean and all entries within one s.d.; underline: next tier. These runs form a separate replicate set from Table 1.

| Metric | Ours (no KL) | Ours (no backbone loss) | Ours |
|---|---|---|---|
| $CC_{volume}$ (↑) | **0.84±0.02** | **0.86±0.02** | **0.84±0.02** |
| Rama. outliers (%, ↓) | **0.00±0.00** | 0.06±0.03 | **0.00±0.00** |
| Rota. outliers (%, ↓) | 0.30±0.10 | 0.25±0.06 | **0.09±0.06** |
| Clashscore (↓) | 13.3±0.3 | 29.0±0.2 | **13.0±0.1** |
| EMRinger (↑) | **1.229±0.201** | **1.264±0.178** | **1.279±0.328** |
| CaBLAM outliers (%, ↓) | **0.67±0.04** | 1.30±0.00 | **0.70±0.00** |
| CaBLAM disfavored (%, ↓) | **3.92±0.08** | 6.07±0.04 | 4.05±0.09 |

### A.4.2 Sampling Conformations for a New Protein

After training the latent diffusion model on cryo-EM map ensembles, it is able to generate atomic ensembles for unseen proteins at inference time, bypassing the need for experimental density maps. This capability is validated on new proteins that lie in the same family as the training set and exhibit similar motions, displaying in-domain generalization. During inference, the sampling process starts with pure Gaussian noise samples in the latent space:

$$z_0 \sim \mathcal{N}(0, \mathbf{I}) \tag{10}$$

The latent model $\text{MLP}_\beta$ is evaluated over 50 discrete integration steps to yield a generated ensemble of latent grids $z_1$. The sampled latent grids are then passed through the VAE decoder to produce the 3D feature grid. The deformation process then proceeds in the same manner as in training, except using the coordinates of the target protein $\tilde{X}_{\text{ref}}$ (generated by Boltz-2) for sampling the feature grid. The trained atomic MLP outputs the offsets used to deform the reference structure of the target protein. The analytical form of the generated coordinates is then:

$$X_{\text{final}} = \text{MLP}_\psi \left( \tilde{X}_{\text{ref}}, \text{GridSample}(\mathcal{F}_\theta(z_1), \tilde{X}_{\text{ref}}) \right) \tag{11}$$

### A.4.3 Results

We train CryoSampler on TRPV3 maps and sample 16 conformations each for two held-out TRP channels, TRPV5 (Dang et al., 2019) and TRPV1 (Astore et al., 2025), evaluating against held-out experimental maps of each target. Because generated structures and maps are unpaired, we compute $CC_{volume}$ between every generated structure and every held-out map, convert it to a distance $\sqrt{1 - CC_{volume}}$, and report ensemble precision, recall, and Wasserstein ($\mathcal{W}_2$) distance. AlphaFlow (Jing et al., 2024) and BioEmu (Lewis et al., 2025) do not support multi-chain proteins, and ConforNets and Boltz-sample (Suzuki & Amagasa, 2026) exceeded 80 GB of GPU memory on these tetramers (over 19,000 atoms). As shown in Table 8, CryoSampler and AF-Multimer-Cluster perform comparably and substantially better than Boltz-2, normal mode analysis, and MSA subsampling: CryoSampler attains the highest precision on both targets, while AF-Multimer-Cluster attains higher recall and lower $\mathcal{W}_2$ distance. CryoSampler requires a single structure prediction for the target, after which conformations are sampled by the lightweight latent model, whereas AF-Multimer-Cluster requires a full structure prediction per sample. We view these results as preliminary: they involve two targets within one protein family, and demonstrating broader generalization would require substantially more heterogeneous cryo-EM training data than is currently available.

Table 8: In-family transfer from TRPV3 to held-out TRPV1 and TRPV5 maps, with no cryo-EM data for the target. Mean ± s.d. over replicates. Bold: best mean and all methods within one s.d.; underline: next tier.

| | Metric | Boltz-2 | NMA | Boltz-2 (MSA sub.) | AF-Multimer-Cluster | Ours |
|---|---|---|---|---|---|---|
| TRPV1 | Precision (↑) | 0.39±0.00 | 0.38±0.01 | 0.39±0.01 | 0.45±0.02 | **0.46±0.00** |
| | Recall (↑) | 0.37±0.00 | 0.37±0.01 | 0.37±0.01 | **0.44±0.02** | 0.40±0.00 |
| | $\mathcal{W}_2$ (↓) | 0.64±0.00 | 0.65±0.01 | 0.64±0.01 | **0.59±0.02** | **0.60±0.00** |
| TRPV5 | Precision (↑) | 0.32±0.01 | 0.31±0.01 | 0.32±0.01 | 0.48±0.01 | **0.50±0.00** |
| | Recall (↑) | 0.29±0.01 | 0.29±0.01 | 0.30±0.02 | **0.45±0.01** | 0.43±0.00 |
| | $\mathcal{W}_2$ (↓) | 0.72±0.01 | 0.71±0.01 | 0.72±0.01 | **0.55±0.01** | 0.57±0.00 |

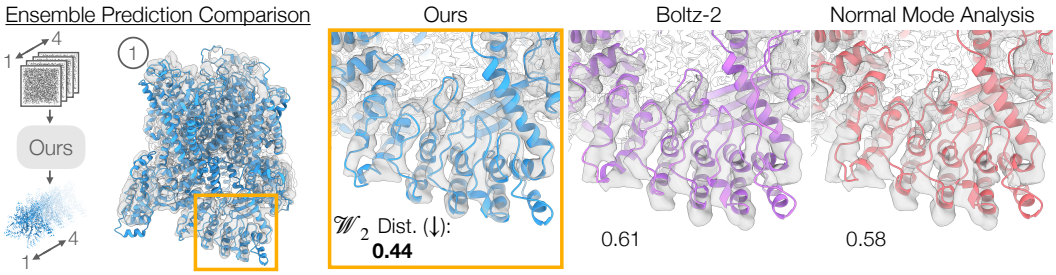


Figure 4: Qualitative in-family transfer example (TRPV3 → TRPV5) from the original single-principal-component protocol with four maps per system. One held-out TRPV5 map is shown in transparent gray with the closest generated structure for each method. Quantitative results under the updated protocol are in Table 8.

### A.5 Cryo-EM Data Pre-processing

In the following sub-sections, we describe the steps taken to process each cryo-EM dataset analyzed in the main text.

#### A.5.1 TRPV3

Picked particles from EMPIAR-10400 (Shimada et al., 2020) were downloaded and reconstructed via backprojection in CryoSPARC (Punjani et al., 2017) with C4 symmetry. These were fed to a heterogeneous reconstruction job with 6 junk classes. The highest resolution class with 79,765 particles was subjected to homogeneous reconstruction with C4 symmetry, yielding a 3.2 angstrom reconstruction.

Heterogeneous reconstruction was subsequently performed with 3D Variability Analysis (Punjani & Fleet, 2021), yielding five principal components with 20 maps each. We selected four maps evenly spaced along the fourth principal component, representing a “breathing" motion of the protein. These maps were directly fed as input to each model building method without further pre-processing, using the `phenix.mtriage` command to additionally input estimated resolutions for each map to the individual algorithms. This command was used for all subsequent datasets except for the P-glycoprotein, for which per-map resolutions were available on the Electron Microscopy Data Bank.

#### A.5.2 TRPV5

Picked particles from EMPIAR-10255 (Dang et al., 2019) were cleaned for corrupt particles and subjected to ab initio reconstruction with a maximum resolution of 12 angstrom. Subsequent homogeneous refinement with C4 Symmetry resulted in a reconstruction of 3.0 angstrom. 3DVA (Punjani & Fleet, 2021) was used to perform heterogeneity analysis, producing five principal components with 20 maps each. For the transfer experiments in Appendix A.4, we use maps spanning all five principal components (10 maps) as a held-out test set; these maps are never used for training or model building.

#### A.5.3 TRPV1

For the second transfer target, we use six previously published heterogeneous reconstructions of ligand-bound TRPV1 (Astore et al., 2025) as a held-out test set.

#### A.5.4 P-GLYCOPROTEIN

Four maps of the P-glycoprotein undergoing drug transport were downloaded from the Electron Microscopy Data Bank (`EMD-40226, 40259, 40258, 40227`) along with their corresponding deposited atomic models (`PDB: 8GMG, 8SA1, 8SA0, 8GMJ`) (Culbertson & Liao, 2025). These maps were then masked with a 5 angstrom extent around the PDB models using the `volume zone` command in ChimeraX (Meng et al., 2023) in order to remove background.

#### A.5.5 INTEGRIN $\alpha V\beta 8$

The data were acquired from EMPIAR-10345 (Campbell et al., 2020) and processed with CryoSPARC and 3DVA to obtain heterogeneous volumes. Four evenly-spaced maps were selected along the first principal component, and the volumes were resampled at 2.5 angstrom per voxel resolution and thresholded at minimum 0.28 intensity in ChimeraX. The density corresponding to the E and F chains of the protein were masked out using the `Segment Map` feature as Boltz-2 cannot accurately predict the binding of these dimers. Finally, the maps were masked to a tight box of size $(50, 100, 90)$.

#### A.5.6 NEUROKININ-1 GPCR

The data were acquired from EMPIAR-10786 (Harris et al., 2022) and processed with CryoSPARC and 3DVA to obtain heterogeneous volumes. Four evenly-spaced maps were selected along the first principal component, and the volumes were resampled at 2.5 angstrom per voxel resolution. The maps were masked with a 5 angstrom extent around the PDB model using the `volume zone` command, and floating density was removed using the command `surface dust size 5`.

### A.6 SETTINGS FOR BASELINE METHODS AND EVALUATION

All experiments were carried out on a single NVIDIA A100 GPU with 40 GB or 80 GB of memory (adjusted to the memory requirements of the evaluated method).

#### A.6.1 MODEL BUILDING

Each baseline is run with default parameters on each of the maps separately over 3 replicates. Our method, E2GMM, and CryoBoltz require an initial structure, which we generate with Boltz-2 and align to the input map before performing model building. Sequences for each protein were acquired from the PDB (Berman et al., 2000). In order to better match the structure output to the experimental density maps, the E and F chains for integrin were removed from the sequence input, and the A and B chains were truncated to only include the first 958 and 636 residues, respectively. Similarly, for the neurokinin-1 GPCR, the first 25 residues of the B chain were removed.

The atomic models output by each method were aligned to the corresponding map using the `fitmap` command in ChimeraX in order to remove any residual rigid body transformational error. B-factors were estimated for each model using the `phenix.real_space_refine` command in Phenix (Liebschner et al., 2019) with the `run=adp` argument. The map–model fit and MolProbity geometry metrics were then computed using `phenix.map_correlations` and `phenix.molprobity`.

#### A.6.2 ENSEMBLE PREDICTION

We evaluated our method, Boltz-2, normal mode analysis, Boltz-2 with MSA subsampling (Monteiro da Silva et al., 2024), and AF-Multimer-Cluster (Rustamov & Baev, 2024) at generating ensembles of TRPV5 and TRPV1. Our method was first trained on TRPV3 maps spanning all five principal components. To train the flow matching model, we augmented the latent codes of the training maps with latents interpolated evenly between adjacent pairs of true latents. After training, our method was given a static Boltz-2 structure of each target (TRPV5 or TRPV1) and generated 16 conformations. Similarly, normal mode analysis in ProDy (Bakan et al., 2011) was used to generate

16 conformations along the first principal component, with modes computed using the $C\alpha$ atoms with an RMSD extent of 1 angstrom. Boltz-2 was run 16 times independently to output 16 different conformations.

Unlike for model building, where we have access to corresponding pairs of atomic models and maps, the generated conformations in this stage are unpaired. As such, we computed pairwise analysis of every generated conformation from each method against each map in the held-out set for TRPV5. Each atomic conformation was aligned to each map using `fitmap` in ChimeraX, after which B-factors were computed using `phenix.real_space_refine`. Pairwise map–model fit was calculated using `phenix.map_correlations`, and the resulting $CC_{volume}$ metric was used to compute pairwise structure–map distances using the formula $\sqrt{1 - CC_{volume}}$, from which precision, recall, and Wasserstein distances were calculated.