LangBoltz: All-Atom Protein Ensemble Generation via Boltzmann-Aligned Protein Language Models

Yikai Liu [1 2 3] **Samuel Sledzieski** [2 3] **Ming Chen** [4] **Guang Lin** [1] **Sonya M. Hanson** [2 3] **Abhilash Sahoo** [2 3]

Abstract

Efficient sampling of the Boltzmann distribution of proteins is fundamental to understanding their biophysical properties, yet it remains a long-standing challenge in computational biology and chemistry. Traditional physics-based approaches, such as molecular dynamics simulations, are often bottlenecked by prohibitive computational costs and slow convergence in rugged energy landscapes. In this work, we introduce LangBoltz, an all-atom protein language model designed for efficient generative sampling of protein conformation ensembles. LangBoltz combines large-scale all-atom pretraining, which yields a broad generative prior over plausible protein conformations, with an efficient Boltzmann alignment procedure that progressively steers the model toward the Boltzmann distribution of a target protein. Our results demonstrate that the Boltzmann alignment method can correct biases in the pretrained LangBoltz model to achieve consistency with the Boltzmann distribution for mid-size and flexible proteins. By bridging the gap between large-scale structural pretraining and rigorous physics, LangBoltz represents a scalable paradigm for accurate and efficient protein ensemble generation.

1. Introduction

Proteins are the fundamental building blocks of life, carrying out essential functions such as catalysis, signaling, and molecular transport. These functions are intimately linked to conformational flexibility and dynamics, making accurate characterization of protein conformational ensembles crucial for understanding molecular mechanisms of biological processes (Whisstock & Lesk, 2003). Traditional computational approaches, most notably molecular dynamics (MD) simulations, have been developed and refined over decades to study protein folding, unfolding, and functional motions at atomic resolution (Shaw et al., 2010; Robustelli et al., 2018). However, MD simulations remain computationally expensive, and the feasible simulation time is usually orders of magnitude smaller than the timescales of biologically relevant conformational changes (Izaguirre et al., 1999).

Recent advances in deep generative models have introduced powerful alternatives for modeling protein conformational landscapes. A growing class of data-driven generative models aims to learn the equilibrium distribution of protein conformations directly from large structural databases and equilibrium MD trajectories (Lu et al., 2023; Lewis et al., 2025; Jing et al., 2024; Wayment-Steele et al., 2024; Liu et al., 2025). Leveraging large-scale structural data from sources such as the Protein Data Bank (PDB) (Bank, 1971) and AlphaFold Database (AFDB) (Varadi et al., 2022), together with extensive MD datasets (Schweke et al., 2024; Mirarchi et al., 2024; Lewis et al., 2025), these approaches have demonstrated strong empirical performance in recovering equilibrium ensembles and capturing conformational diversity across proteins. However, because training data provides limited and uneven coverage of equilibrium ensembles, these models generally do not guarantee physically correct relative populations of metastable states.

Complementary to data-driven approaches, energy-driven methods seek to explicitly enforce physical consistency by training generative models to approximate Boltzmann distributions defined by a given energy function (Noé et al., 2019; Mahmoud et al., 2022; Klein & Noé, 2024; Zheng et al., 2024). This line of research aims to construct direct samplers for equilibrium distributions, enabling independent and identically distributed sampling without relying on long dynamical trajectories. Despite recent progress, energy-driven approaches have so far lagged behind in scalability and robustness. A central challenge lies in balancing exploration and exploitation: training from scratch is often unstable due to large energy fluctuations (Vanommeslaeghe

[1]Department of Mechanical Engineering, Purdue University, West Lafayette, Indiana 47907, USA [2]Center for Computational Biology, Flatiron Institute, New York, NY, USA [3]Center for Computational Mathematics, Flatiron Institute, New York, NY, USA [4]Department of Chemistry, Purdue University, West Lafayette, Indiana 47907, USA. Correspondence to: Abhilash Sahoo <asahoo@flatironinstitute.org>, Sonya M. Hanson <shanson@flatironinstitute.org>.

et al., 2010), while insufficient exploration prevents discovery of important metastable states, leading to mode collapse. (Midgley et al., 2022; Felardos et al., 2023; Nam et al., 2025; Schopmans & Friederich, 2025)

Based on these gaps, we introduce **LangBoltz**, a generative protein Language model for Boltzmann-aligned ensemble generation. LangBoltz provides a systematic framework for combining large-scale pretrained protein ensemble generators with principled energy-based training. We train an all-atom autoregressive protein language model using large-scale single-structure and MD simulation datasets. Building on this pretrained model, we introduce an annealed importance sampling–based training objective that iteratively refines the model distribution from the pretrained prior toward the target Boltzmann distribution. We demonstrate the effectiveness of LangBoltz on three mid-sized proteins spanning distinct conformational regimes, including intrinsically disordered, fast-folding, and relatively rigid proteins.

In summary, this work makes the following contributions:

1. **Pretrained all-atom protein ensemble model.** We train and release a scalable autoregressive protein language model capable of generating diverse all-atom protein conformational ensembles.

2. **Systematic energy-based fine-tuning framework.** We propose a principled annealed importance sampling–based training strategy that enables stable and scalable energy-driven Boltzmann training on top of large pretrained models, and provide ablation studies analyzing its strengths and failure modes.

2. Background

Sampling from unnormalized probability distributions. A central problem in molecular dynamics and computational biophysics is to generate protein conformational ensembles that follow the Boltzmann distribution defined by a physics-based energy function:

$$P_B(\mathbf{x}) = \frac{\exp(-U(\mathbf{x})/k_B T)}{Z}, \tag{1}$$

where $\mathbf{x}$ denotes all-atom Cartesian coordinates, $U(\mathbf{x})$ is the potential energy, T is the temperature, k_B is the Boltzmann constant, and Z is the partition function. This equilibrium distribution governs the relationship between protein sequence, structure, dynamics, and function.

Traditionally, sampling from P_B is performed using physics-based simulations, most notably molecular dynamics (MD). MD simulates the time evolution of a molecular system by numerically integrating Newton's equations of motion. Recent work has explored the use of deep generative models for Boltzmann sampling. Pioneered by Boltzmann Generators (Noé et al., 2019), this line of research aims to learn direct samplers for equilibrium distributions, enabling independent and identically distributed (i.i.d.) generation once training converges and bypassing the barrier-crossing limitations of dynamical simulations. However, a major challenge in such approaches is mode collapse: if certain metastable states have negligible probability under the learned model, they are unlikely to be discovered or recovered during training. This issue is expected, as generative models receive no training signal for regions of configuration space that are rarely or never sampled. To address these challenges, recent studies have proposed techniques such as temperature annealing (Schopmans & Friederich, 2025), annealed importance sampling (Midgley et al., 2022), and enhanced-sampling-inspired biasing strategies that promote exploration of slow collective variables (Nam et al., 2025) to encourage mode discovery. Another difficulty lies in identifying and accurately modeling important metastable states in high-dimensional conformational spaces, where the probability of discovering all relevant modes decreases rapidly with system size. A growing line of work (Wang et al., 2024; Lu et al., 2025) therefore seeks to initialize Boltzmann samplers from pretrained protein structure or ensemble generators. Nevertheless, existing approaches often still require substantial amounts of equilibrium MD data for fine-tuning, limiting their scalability and practicality.

Protein conformation representation. In this work, we adopt the pretrained ESM3 structure tokenizer (Hayes et al., 2025) to represent protein backbone conformations as discrete token sequences $\mathbf{c}_{\text{bb}} \in \mathbb{Z}^N$, where each residue is assigned one of $4{,}096$ learned structure tokens. These tokens provide a compact representation of the local backbone structural environment around each residue. The discretization is performed using a VQ-VAE encoder (Van Den Oord et al., 2017), and a paired decoder reconstructs generated token sequences back into three-dimensional backbone coordinates. Since the open ESM3 model provides backbone-only representations, we introduce a dedicated side-chain tokenization strategy. Side-chain conformations, being largely local, are encoded independently using a torsion-angle discretization. For each residue, we represent up to four sidechain torsion angles (χ_1–χ_4). Each torsion angle is uniformly discretized into 64 bins over $[0, 2\pi)$, yielding a categorical sidechain token representation $\mathbf{c}_{\text{sc}} \in \mathbb{Z}^{4N}$. This representation captures the dominant rotameric degrees of freedom while remaining compatible with autoregressive sequence modeling.

3. Overview of LangBoltz

In this section, we introduce the high-level framework of LangBoltz, as illustrated in Fig. 1. LangBoltz is an all-atom

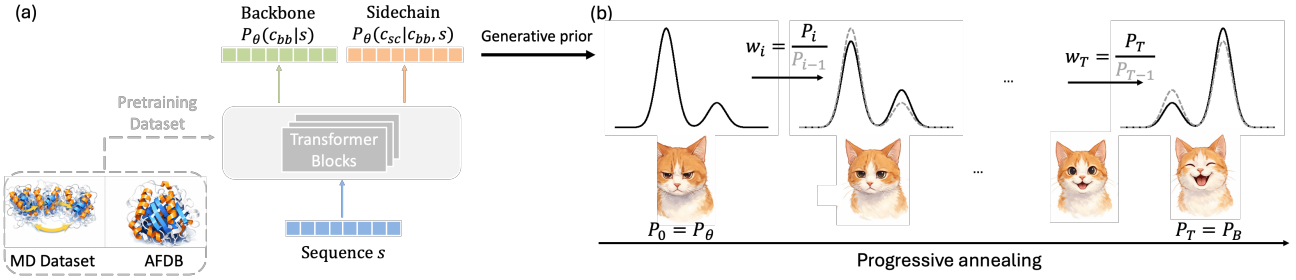

Figure 1. An illustrative framework of LangBoltz. (a) LangBoltz is an all-atom protein language model that operates on discretized representations of protein sequences and structures. It is pretrained on large scale single structure from AFDB and ensemble conformations from MD dataset, and employs a powerful autoregressive transformer to generate equilibrium protein conformational ensembles. (b) The exact likelihood evaluation provided by LangBoltz enables scalable alignment with the target Boltzmann distribution. Using a progressive annealing strategy, LangBoltz gradually transforms the pretrained distribution into the desired Boltzmann distribution.

generative protein language model designed to efficiently generate protein conformational ensembles. The framework consists of two stages: (i) a large-scale pretraining stage (Fig. 1(a)) that learns to approximate the protein conformational energy landscape, and (ii) a Boltzmann alignment protocol (Fig. 1(b)) that refines the pretrained model to accurately sample the Boltzmann distribution of a target protein.

The pretraining stage equips the model with an initial understanding of the protein conformational energy landscape. Specifically, LangBoltz is pretrained on a diverse dataset of protein structures, including single native structures from the AlphaFold database as well as equilibrium conformational ensembles obtained from molecular dynamics simulations. This enables the model to capture not only native folded conformations but also structural thermodynamic fluctuations and alternative metastable states.

However, while the pretrained model can provide reasonable guesses of protein conformational space, it does not enforce consistency with the true Boltzmann distribution defined by a physical energy function. Therefore, when studying a specific protein of interests, an explicit alignment procedure is required for accurate ensemble modeling. To faithfully recover the target Boltzmann distribution while avoiding mode collapse, we introduce a progressive annealing strategy inspired by Flow Annealed Importance Sampling Bootstrap (FAB) (Midgley et al., 2022). During Boltzmann alignment, the pretrained model distribution is progressively transformed toward the target Boltzmann distribution for a specific protein system through a sequence of intermediate distributions. This progressive annealing procedure stabilizes training, improves sample efficiency, and enables accurate learning of equilibrium protein conformational ensembles.

3.1. Pretraining Objective

At the pretraining stage, LangBoltz aims to learn the equilibrium distribution of conformations $P_\theta(\mathbf{c}_{\text{bb}}, \mathbf{c}_{\text{sc}} \mid s)$, where $\mathbf{c}_{\text{bb}}$ and $\mathbf{c}_{\text{sc}}$ denote backbone and sidechain structural tokens, respectively, and s denotes the amino acid sequence. LangBoltz operates in an autoregressive manner, sequentially predicting structural tokens for each residue conditioned on the input sequence and all previously generated tokens. We decompose the learning problem into two tasks:

$$P_\theta(\mathbf{c}_{\text{bb}}, \mathbf{c}_{\text{sc}} \mid s) = P_\theta(\mathbf{c}_{\text{bb}} \mid s)\, P_\theta(\mathbf{c}_{\text{sc}} \mid \mathbf{c}_{\text{bb}}, s). \quad (2)$$

The backbone models learns the equilibrium distribution of protein backbones conditioned on sequence, while the sidechain model learns conditional sidechain conformations given the backbone geometry and sequence.

1. **Backbone.** The backbone model aims to learn the equilibrium distribution of backbone conformations $P_\theta(\mathbf{c}_{\text{bb}} \mid s)$. Specifically, the distribution is modeled autoregressively as:

$$P_\theta(\mathbf{c}_{\text{bb}} \mid s) = \prod_{i=0}^{N-1} P_\theta(\mathbf{c}_i^{\text{bb}} \mid \mathbf{c}_{<i}^{\text{bb}}, s), \quad (3)$$

where $\mathbf{c}_i^{\text{bb}}$ denotes the backbone structure token of residue i, and $\mathbf{c}_{<i}^{\text{bb}}$ represents all preceding backbone tokens. The backbone model is pretrained by minimizing the negative log-likelihood of backbone conformations:

$$\mathcal{L}_{\text{bb}}(\theta) = -\mathbb{E}_{s,\mathbf{c}_{\text{bb}}}\left[\sum_{i=0}^{N-1} \log P_\theta(\mathbf{c}_i^{\text{bb}} \mid \mathbf{c}_{<i}^{\text{bb}}, s)\right]. \quad (4)$$

2. **Sidechain.** The sidechain model learns the conditional distribution of sidechain conformations given the backbone and sequence $P_\theta(\mathbf{c}_{\text{sc}} \mid \mathbf{c}_{\text{bb}}, s)$. Conditioned on a

fixed backbone, sidechain conformations are modeled autoregressively over residues with local dependencies:

$$\begin{aligned} \mathcal{E}_i &= \left(\mathbf{c}^{\mathrm{sc}}_{<i\cap\mathcal{N}(i)}, \mathbf{c}^{\mathrm{bb}}_{\mathcal{N}(i)}, s_{\mathcal{N}(i)}\right), \\ P_\theta(\mathbf{c}_{\mathrm{sc}} \mid \mathbf{c}_{\mathrm{bb}}, s) &= \textstyle\prod_{i=0}^{N-1} P_\theta(\mathbf{c}^{\mathrm{sc}}_i \mid \mathcal{E}_i)\,, \end{aligned} \tag{5}$$

where $\mathbf{c}^{\mathrm{sc}}_i$ denotes the sidechain structure token of residue i, and $\mathcal{N}(i)$ denotes the local environment of residue i. The local environment $\mathcal{N}(i)$ is defined as all residues whose backbone C_α atoms lie within a radius of 12 Å of residue i (Jumper et al., 2018). This locality assumption reflects the short-range nature of sidechain packing, substantially reducing modeling complexity while preserving physical fidelity. The sidechain model is pretrained by minimizing the conditional negative log-likelihood of sidechain conformations:

$$\mathcal{L}_{\mathrm{sc}}(\theta) = -\mathbb{E}\left[\sum_{i=0}^{N-1} \log P_\theta(\mathbf{c}_{\mathrm{sc}} \mid \mathbf{c}_{\mathrm{bb}}, s)\right]. \tag{6}$$

3.2. Pretraining setup

Data Following BioEmu (Lewis et al., 2025), we construct a training dataset that combines single-structure protein data with equilibrium molecular dynamics (MD) ensembles. For the single-structure dataset, we use the Swiss-Prot subset of the AlphaFold database, which contains 542,378 high-confidence sequence–structure pairs. To expose the model to equilibrium conformational variability, we additionally incorporate molecular dynamics (MD) simulation data from three sources: mdCATH (Mirarchi et al., 2024), BioEmu (Lewis et al., 2025), and ATLAS (Schweke et al., 2024). These datasets comprise more than 30,000 distinct proteins and over 75 milliseconds of MD simulation time.

Model details. To obtain rich sequence and structural representations, we adopt the pretrained sequence and backbone structure embedding modules from ESM3 and freeze them throughout training. The backbone model consists of 24 transformer blocks with a total of 1.4 billion parameters, while the sidechain model consists of 8 transformer blocks with 455 million parameters. Both models follow the transformer architecture introduced in ESM3 (Hayes et al., 2025), employing pre-layer normalization (Pre-LN), rotary positional embeddings (RoPE) (Su et al., 2024), and SwiGLU activation functions. The backbone and sidechain models are trained independently during the pretraining stage.

We emphasize that the use of an autoregressive transformer architecture is a deliberate design choice. Autoregressive models combine scalable sequence modeling with exact likelihood evaluation, enabling a principled connection to statistical mechanics. In contrast to diffusion-based generative models (Song et al., 2020), which lack tractable likelihoods, or normalizing flows (Dinh et al., 2016), which require restrictive architectural constraints to ensure invertibility, autoregressive transformers provide a scalable framework for likelihood-based training. This property is essential for the Boltzmann alignment objectives pursued in this work.

3.3. Progressive Annealing Boltzmann Generator Training

The objective of Boltzmann alignment training is to finetune the pretrained model such that it can generate independent and identically distributed (i.i.d.) protein conformational ensembles that follow the Boltzmann distribution defined in Eq. 1. LangBoltz defines a generative distribution over discrete structure tokens $\mathbf{c}$. Let $\mathbf{c} = \mathrm{Enc}(\mathbf{x})$ denote the deterministic structure encoder from an all-atom structure $\mathbf{x}$. The probability distribution over tokens is:

$$P_B(\mathbf{c}) = \int P(\mathbf{c} \mid \mathbf{x})\, P_B(\mathbf{x})\, d\mathbf{x} = \int_{\mathrm{Enc}(\mathbf{x})=\mathbf{c}} P_B(\mathbf{x})\, d\mathbf{x}. \tag{7}$$

For a given structural sequence $\mathbf{c}$, the corresponding preimage $\{\mathbf{x} : \mathrm{Enc}(\mathbf{x}) = \mathbf{c}\}$ may contain many continuous conformations, including both physically plausible and implausible geometries. While the total energy variation within this preimage can be large, the Boltzmann distribution conditioned on a fixed structural sequence is sharply concentrated around a small subset of locally relaxed conformations. High-energy conformations within the same token class carry negligible Boltzmann weight and therefore do not contribute to equilibrium observables. As a result, each sequence effectively represents a single dominant metastable basin, enabling a discrete structural representation that preserves equilibrium thermodynamics. Thus, the Boltzmann weight of a structure sequence can be approximated by:

$$P_B(\mathbf{c}) \;\propto\; \exp(-U(\mathbf{x_c})/k_B T)\,, \tag{8}$$

where $\mathbf{x_c}$ denotes a representative structure associated with token $\mathbf{c}$ and $U(\mathbf{x_c})$ denotes the physics-based energy function of the representative structure. In practice, $\mathbf{x_c}$ is obtained by locally minimizing the energy within the discretized structure tokens. This approximation enables modeling equilibrium distributions in the discrete token space.

Since the autoregressive model provides exact likelihoods for generated samples, it enables alignment with the Boltzmann distribution through Kullback–Leibler divergence:

$$D_{\mathrm{KL}}(P_B \,\|\, P_\theta) = \mathbb{E}_{\mathbf{c}\sim P_B}\left[-\log P_\theta(\mathbf{c})\right] - H(P_B), \tag{9}$$

with $H(P_B) = -\mathbb{E}_{\mathbf{c}\sim P_B}\left[\log P_B(\mathbf{c})\right]$ the entropy independent of the model parameters θ.

While direct sampling from P_B is intractable in many cases, we can alternatively estimate the expectation using impor-

tance sampling with a proposal distribution $q(\mathbf{c})$, yielding:

$$\mathbb{E}_{\mathbf{c}\sim P_B}\left[-\log P_\theta(\mathbf{c})\right] = \mathbb{E}_{\mathbf{c}\sim q}\left[\frac{P_B(\mathbf{c})}{q(\mathbf{c})}\left(-\log P_\theta(\mathbf{c})\right)\right]. \tag{10}$$

In an online learning setting, samples are drawn from the current model distribution, which is therefore used as the proposal distribution, i.e., $q = P_\theta$.

In practice, although Eq. 10 yields unbiased estimators, its effectiveness depends critically on the overlap between the proposal distribution P_θ and the target Boltzmann distribution P_B. When P_θ deviates substantially from P_B, importance weights become highly variable, leading to a low effective sample size and unstable optimization.

Inspired by annealed importance sampling (Neal, 2001) and its recent adaptation in generative Boltzmann generator (Midgley et al., 2022), we introduce a sequence of intermediate distributions $\{p_i\}_{i=0}^{T}$ that smoothly interpolate between the pretrained model and the target distribution. We set $p_0 = P_\theta$ as the pretrained autoregressive model and $p_T = P_B$ as the target Boltzmann distribution. The intermediate distributions are defined recursively as:

$$\log p_i(\mathbf{c}) = \beta_i \log p_{i-1}(\mathbf{c}) + (1-\beta_i)\log P_B(\mathbf{c}), \tag{11}$$

where $1 = \beta_0 > \beta_1 > \cdots > \beta_T = 0$ defines a monotonically decreasing annealing schedule.

At annealing iteration i, we perform the following steps:

1. **Sampling.** Draw a dataset $\mathcal{C}_i = \{\mathbf{c}^{(n)}\}_{n=1}^{N}$ by sampling from the current model distribution p_{i-1}.
2. **Importance weighting.** For each sample $\mathbf{c}^{(n)}$, compute the unnormalized importance weight:

$$w_i(\mathbf{c}^{(n)}) = \frac{p_i(\mathbf{c}^{(n)})}{p_{i-1}(\mathbf{c}^{(n)})} \propto \left(\frac{P_B(\mathbf{c}^{(n)})}{p_{i-1}(\mathbf{c}^{(n)})}\right)^{1-\beta_i}. \tag{12}$$

3. **Weighted training.** Update the model by minimizing the weighted cross-entropy loss:

$$\mathcal{L}_i(\theta) = -\mathbb{E}_{\mathbf{c}\sim p_{i-1}}\left[\mathrm{sg}(w_i(\mathbf{c}))\,\log p_\theta(\mathbf{c})\right], \tag{13}$$

where $\mathrm{sg}(\cdot)$ denotes the stop-gradient operator. This progressive annealing procedure gradually aligns the model distribution from the pretrained prior toward the target distribution, mitigating mode collapse and improving effective sample size compared to direct importance sampling.

4. Experiments

4.1. Pretraining evaluation

To assess the pretrained backbone model, we evaluate its ability to generate metastable conformational states for 20 octapeptides that are unseen during backbone-model training. We quantify agreement with reference molecular dynamics (MD) simulations using the Jensen–Shannon divergence (JSD) computed over several collective variables (CVs), including low-dimensional physical observables and the top two time-lagged independent component analysis (TICA) components capturing the slowest dynamical modes of the system. We compare our results against BioEmu (Lewis et al., 2025), the state-of-the-art protein ensemble generator trained on a scaled MD dataset. For a fairer comparison, we also compare against BioEmu using proteins from CATH1 dataset, where both models are trained upon.

For sidechain modeling, we evaluate whether the model produces thermodynamically consistent conformations by benchmarking against reference MD simulations on the CATH1 dataset. This dataset consists of 50 proteins, each accompanied by 100 μs of MD simulations, ensuring sufficient convergence of sidechain conformational distributions. Performance is assessed using three complementary criteria: (i) similarity of per-residue sidechain torsion-angle (χ) distributions, (ii) steric-clash ratios, and (iii) the number of hydrogen bonds in the reconstructed structures. As a primary baseline, we compare against AttnPacker (McPartlon & Xu, 2023), SOTA deterministic sidechain packing model.

4.2. Boltzmann alignment evaluation

Systems. We evaluate our method on three proteins of different flexibility: PaaA2 (Sterckx et al., 2014), Villin (Friederich et al., 1990), and Ubiquitin (Ubq) (Hershko & Ciechanover, 1998). PaaA2 is an intrinsically disordered protein with helical segments, Villin is a fast-folding protein, and Ubiquitin is a relatively rigid protein exhibiting predominantly local fluctuations. The three proteins contain 35 (Villin), 71 (PaaA2) and 76 (Ubiquitin) residues, substantially larger than systems typically studied in prior purely energy-driven Boltzmann generator work.

Ground-truth simulations. For reference, we perform all-atom molecular dynamics simulations for each protein at 350 K, with a total simulation time of 15 μs per system using the implicit solvent Generalized Born model GBN2 (Nguyen et al., 2013). These trajectories are used to characterize equilibrium ensembles and to evaluate the fidelity of Boltzmann sampling.

Annealing schedules. The annealing schedule is a key design choice in our energy-based training procedure. For a fixed number of annealing iterations, the schedule governs how probability mass is redistributed across intermediate distributions, directly affecting training stability, gradient variance, and sample efficiency. In this work, we primarily adopt a geometric annealing schedule, which is widely used due to its favorable stability properties, particularly the

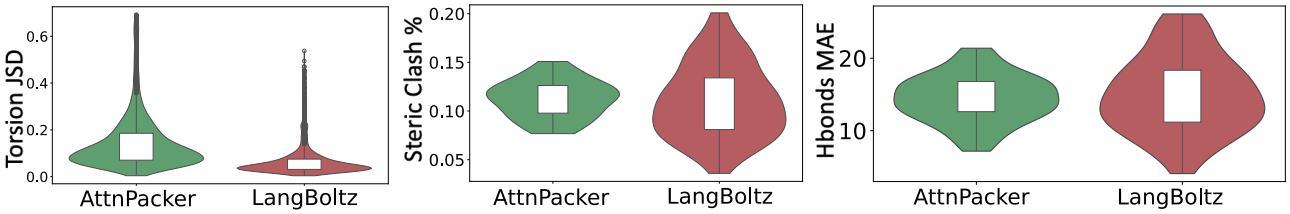

Figure 2. **Comparison of sidechain packing performance against baseline models AttnPacker (deterministic backmapping from backbone) on the CATH1 test dataset (40 protein domains, each with 100 μs MD trajectories).** (a) Jensen–Shannon divergence between sidechain torsion-angle distributions from MD simulations and generated ensembles. (b) Percentage of steric clashes. (c) Mean absolute error of hydrogen-bond counts relative to MD simulations. Lower values indicate better performance for all metrics.

strong distributional overlap at early training stages:

$$\beta_i = (1 - \lambda^{1-i/T})/(1 - \lambda), \qquad i = 0, \ldots, T, \tag{14}$$

where $\lambda \in (0, 1)$ controls the annealing rate. We set $\lambda = 0.2$ throughout to ensure stable optimization.

To assess robustness with respect to annealing design, we additionally perform an ablation study across different numbers of annealing steps $T \in 20, 40, 60$. This evaluates convergence behavior under different computational budgets.

Evaluation metrics. To evaluate the sampling quality of LangBoltz at 350 K, we employ a set of complementary metrics. To assess coverage of metastable states, we perform time-lagged independent component analysis (TICA) on the reference MD trajectories and compute the Jensen–Shannon divergence (JSD) between the projected distributions of ground-truth and generated samples. For physical interpretability, we additionally compare JSD values for the radius of gyration distribution and compute the residue-level mean absolute error of root-mean-square fluctuations (RMSF) between generated ensembles and MD simulations. Finally, we evaluate the reverse effective sample size (ESS) to quantify sampling efficiency and weight degeneracy.

5. Results

5.1. Pretraining

We report the test results of model's generalization capacity in Table 1. We observe strong agreement with reference MD simulations, achieving comparable performance to the BioEmu baseline. Note that these proteins are used as training proteins for the baseline model BioEmu but not for the LangBoltz model. While BioEmu remains better on octapeptides, LangBoltz is competitive despite not training on these peptides. This result highlights the ability of LangBoltz to generalize beyond its training distribution and to recover accurate conformational ensembles on previously unseen systems. Additionally, we observed that the free energy mean absolute error (MAE) over these 17 CATH systems is of 0.5 kcal/mol, a value lower than the state-of-the-art protein ensemble generator BioEmu (0.9 kcal/mol). Results on CATH1 proteins are shown in Appendix Fig. 4 - 6.

Table 1. Distributional similarity evaluation on the Octapeptide test dataset. Metrics are reported as Jensen–Shannon divergence (JSD) over the radius of gyration (Rg), root-mean-square distance (RMSD) w.r.t the native structure, and the top two TICA components capturing slow protein dynamic Note that the these proteins are used as training proteins for BioEmu but not for LangBoltz.

Model	Rg ↓	RMSD ↓	TICA ↓
LangBoltz	0.036	0.025	0.232
BioEmu	**0.031**	**0.020**	**0.134**

We further report sidechain packing results in Fig. 2. Across all evaluated metrics—including distributional agreement, steric clash ratio, and hydrogen-bond statistics—LangBoltz achieves comparable or improved agreement with reference MD simulations relative to the baseline model AttnPacker. In particular, LangBoltz consistently outperforms AttnPacker in capturing sidechain thermodynamics, as reflected by lower Jensen–Shannon divergence (JSD) values for sidechain torsion-angle distributions.

Results from the backbone and sidechain models indicate that the pretrained LangBoltz model captures essential physical diversity and major statistical properties of protein conformations, providing a strong initialization that facilitates effective subsequent energy-based alignment.

5.2. Boltzmann alignment

Progressive recovery of free-energy landscapes. Figure 3 shows the evolution of the free energy surface (FES) projected onto the top two TICA components for annealing with $T = 60$ iterations. At $t = 0$, corresponding to the pretrained model, the sampled distributions deviate substantially from the ground-truth MD reference. These deviations manifest as collapsed or poorly resolved basins, missing metastable states, and insufficient separation between kinet-

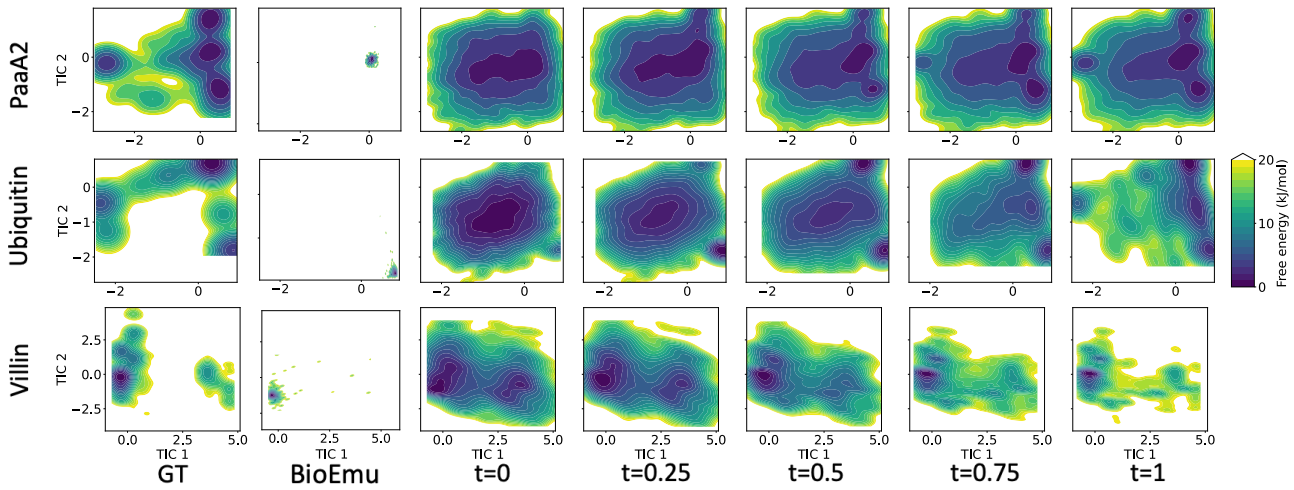

Figure 3. Visualization of the progressive annealing process, showing free energy surfaces along the top two TICA components for samples generated by (1) the pretrained model (t = 0) and (2) intermediate annealing stages at t = 0.25T, 0.5T, 0.75T, and T. The TICA projection is parameterized using backbone torsion angles from ground-truth molecular dynamics (MD) simulations (GT). For reference, results from the baseline model BioEmu is also visualized. Results shown here correspond to T = 60; results for T = 20 and T = 40 are provided in the Appendix.

ically distinct regions. Such artifacts are particularly severe for PaaA2 and Ubiquitin, where the pretrained model fails to capture the multi-basin structure of the energy landscape.

As annealing progresses, the generated distributions are systematically refined towards the target Boltzmann distribution. At intermediate stages ($t = 0.25, 0.5, 0.75$), metastable basins gradually emerge and sharpen, with improved coverage of high-probability regions and clearer kinetic separation. By the final iteration ($t = 1$), the resulting FES for all three proteins aligns better with the MD reference, exhibiting well-resolved metastable states and qualitatively correct basin topology. This improvement is most evident for PaaA2 and Ubiquitin, where Boltzmann alignment successfully recovers all major kinetically separated metastable states absent in the pretrained model. In contrast, for Villin, the pretrained model already captures the dominant two metastable states but produces an overly smooth energy landscape; Boltzmann alignment sharpens basin boundaries and enhances state separation.

Quantitative distributional similarity. As summarized in Table 2, progressive annealing consistently improves distributional agreement with ground-truth MD simulations when evaluated using Jensen–Shannon divergence on both TICA projections and the radius of gyration (R_g), provided a sufficient number of annealing iterations ($T = 60$) is used. In contrast, using a small number of iterations ($T = 20$) leads to a pronounced tendency toward mode collapse in the Boltzmann alignment procedure, as most clearly illustrated in Appendix Figs. 8–11.

Across all three proteins, TICA divergence decreases monotonically with increasing T, indicating progressively improved recovery of metastable states. A similar monotonic improvement is observed for R_g distributions, reflecting enhanced agreement in global conformational statistics. Residue-level flexibility, quantified by RMSF error, also improves with annealing, demonstrating more accurate modeling of local dynamics. The only exception occurs for Villin, where annealing does not yield monotonic improvement; we analyze this failure mode in detail in Appendix D.

Additionally, all annealing schedules achieve significantly higher reverse ESS than the pretrained model. Because reverse ESS does not penalize mode collapse or uncovered regions of the target distribution, it is interpreted alongside additional structural and physical metrics.

Notably, the appendix figures reveal that insufficient annealing iterations systematically concentrate probability mass onto a small number of configurations, leading to collapsed free-energy basins. This behavior highlights the necessity of sufficiently long annealing schedules to maintain overlap between intermediate distributions and avoid mode collapse.

6. Discussion

LangBoltz presents a scalable all-atom protein ensemble generator that matches the generality of diffusion-based protein ensemble models while enabling exact likelihood evaluation in a discrete structural representations. This property allows the model to be explicitly aligned — for a given

Table 2. Distributional similarity evaluation on three test proteins. Metrics include the Jensen–Shannon divergence (JSD) computed over the top two TICA components capturing slow protein dynamics, the radius of gyration (Rg), and the mean absolute error (MAE) of the root-mean-square fluctuation (RMSF). We additionally report the reverse effective sample size (ESS) to assess model likelihood accuracy.

System	Model	TICA JSD ↓	Rg JSD ↓	RMSF MAE (nm) ↓	ESS (%) ↑
PaaA2	Pretrained	0.343	0.035	0.243	0.138
	$T = 20$	0.247	0.116	0.275	0.884
	$T = 40$	0.230	0.021	0.133	**2.135**
	$T = 60$	**0.165**	**0.015**	**0.132**	1.935
Ubiquitin	Pretrained	0.517	0.347	0.761	0.125
	$T = 20$	0.484	0.069	0.078	**1.565**
	$T = 40$	0.427	0.202	0.393	0.817
	$T = 60$	**0.218**	**0.037**	**0.058**	1.444
Villin	Pretrained	0.300	0.537	**0.140**	0.271
	$T = 20$	0.336	**0.132**	0.315	**2.034**
	$T = 40$	0.285	0.354	0.272	1.568
	$T = 60$	**0.265**	0.230	0.193	1.936

target protein, temperature, and chosen energy function — with a target Boltzmann distribution through energy-based training, rather than relying solely on implicit physical consistency learned during pretraining.

The Boltzmann alignment serves two primary purposes. First, it provides a systematic mechanism for correcting deficiencies in the pretrained generative model, as demonstrated in our experiments by recovering missing metastable basins for PaaA2 and Ubiquitin. Such deficiencies are unavoidable in large-scale pretraining due to finite data coverage and model capacity. By progressively annealing toward the target energy function, LangBoltz refines the pretrained distribution faithfully into the target distribution without requiring brute-force exploration of the conformation space.

Second, the framework places no restrictions on the choice of energy function. While this work employs an implicit-solvent energy model for demonstration, the likelihood-based formulation allows users to substitute alternative physical models—such as different solvent descriptions, environment-dependent interactions, problem-specific effective potentials, or even black-box energy functions.

Limitations. The effectiveness of Boltzmann alignment depends on the quality of the pretrained model as an initial proposal distribution. When the pretrained model deviates substantially from the target Boltzmann distribution, longer progressive annealing schedules are required to achieve convergence, as observed in the Villin example. Nevertheless, this behavior reflects the intended design of LangBoltz: rather than performing sampling from scratch, the method leverages pretrained structural priors and focuses computation on correcting residual discrepancies, making all-atom Boltzmann sampling tractable at scale. Finally, the current implementation incurs nontrivial computational overhead due to frequent communication between the machine learning model and the physics-based energy evaluation engine, OpenMM, reflecting an implementation-level limitation.

Outlook. Several directions for future work naturally follow. First, the optimal design of annealing schedules remains an open question; for example, adaptive annealing schemes that dynamically adjust the inverse temperature to prevent importance-weight collapse may improve stability and reduce manual tuning. Second, because alignment requires only energy evaluations, the framework naturally extends to hybrid energy functions that incorporate experimental restraints such as SAXS, NMR, or FRET alongside physics-based energies. Finally, the current approach makes strong assumptions about within-token configurational entropy. Future work could pursue quantitative analyses of within-token entropy and develop the structure tokenizer that are more expressive and better preserve entropic contributions.

7. Conclusion

In this work, we introduce LangBoltz, an all-atom generative model for protein ensemble generation and a systematic and efficient procedure for aligning model outputs to a target Boltzmann distribution. Our results demonstrate that the pretrained model captures and generalizes key features of protein conformational landscapes, including metastable state conformations and side-chain thermodynamics. Building on this foundation, the Boltzmann alignment with progressive annealing provides a scalable and effective mechanism for refining a pretrained generative model into a faithful Boltzmann sampler. Empirical studies on three mid-size proteins spanning different degrees of structural flexibility demonstrate consistent improvements in metastable state coverage and thermodynamic accuracy. Together, these results high-

light the complementary strengths of large-scale data-driven pretraining and physics-based Boltzmann alignment, paving the way for scalable, physically grounded generative modeling of protein thermodynamics across diverse systems.

Impact Statement

This work develops machine learning methods for modeling protein conformational ensembles, with potential applications in basic biological research and computational drug discovery. As a methodological contribution, it does not directly enable deployment in safety-critical or societally sensitive settings. We do not anticipate immediate negative societal impacts; broader implications will depend on downstream applications of the method.

References

Bank, P. D. Protein data bank. *Nature New Biol*, 233(223): 10–1038, 1971.

Charron, N. E., Bonneau, K., Pasos-Trejo, A. S., Guljas, A., Chen, Y., Musil, F., Venturin, J., Gusew, D., Zaporozhets, I., Krämer, A., et al. Navigating protein landscapes with a machine-learned transferable coarse-grained model. *Nature chemistry*, pp. 1–9, 2025.

Dinh, L., Sohl-Dickstein, J., and Bengio, S. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

Eastman, P., Galvelis, R., Peláez, R. P., Abreu, C. R., Farr, S. E., Gallicchio, E., Gorenko, A., Henry, M. M., Hu, F., Huang, J., et al. Openmm 8: molecular dynamics simulation with machine learning potentials. *The Journal of Physical Chemistry B*, 128(1):109–116, 2023.

Felardos, L., Hénin, J., and Charpiat, G. Designing losses for data-free training of normalizing flows on boltzmann distributions. *arXiv preprint arXiv:2301.05475*, 2023.

Friederich, E., Pringault, E., Arpin, M., and Louvard, D. From the structure to the function of villin, an actin-binding protein of the brush border. *Bioessays*, 12(9): 403–408, 1990.

Hayes, T., Rao, R., Akin, H., Sofroniew, N. J., Oktay, D., Lin, Z., Verkuil, R., Tran, V. Q., Deaton, J., Wiggert, M., et al. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025.

Hershko, A. and Ciechanover, A. The ubiquitin system. *Annual review of biochemistry*, 67(1):425–479, 1998.

Izaguirre, J. A., Reich, S., and Skeel, R. D. Longer time steps for molecular dynamics. *The Journal of chemical physics*, 110(20):9853–9864, 1999.

Jing, B., Berger, B., and Jaakkola, T. Alphafold meets flow matching for generating protein ensembles. *arXiv preprint arXiv:2402.04845*, 2024.

Jumper, J. M., Faruk, N. F., Freed, K. F., and Sosnick, T. R. Accurate calculation of side chain packing and free energy with applications to protein molecular dynamics. *PLoS computational biology*, 14(12):e1006342, 2018.

Klein, L. and Noé, F. Transferable boltzmann generators. *Advances in Neural Information Processing Systems*, 37: 45281–45314, 2024.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Lewis, S., Hempel, T., Jiménez-Luna, J., Gastegger, M., Xie, Y., Foong, A. Y., Satorras, V. G., Abdin, O., Veeling, B. S., Zaporozhets, I., et al. Scalable emulation of protein equilibrium ensembles with generative deep learning. *Science*, pp. eadv9817, 2025.

Liu, Y., Zheng, H., Mao, L., Wang, Y., Chen, M., and Lin, G. Protdyn: a foundation protein language model for thermodynamics and dynamics generation. *arXiv preprint arXiv:2510.00013*, 2025.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Lu, J., Zhong, B., Zhang, Z., and Tang, J. Str2str: A score-based framework for zero-shot protein conformation sampling. *arXiv preprint arXiv:2306.03117*, 2023.

Lu, J., Chen, X., Lu, S. Z., Lozano, A., Chenthamarakshan, V., Das, P., and Tang, J. Aligning protein conformation ensemble generation with physical feedback. *arXiv preprint arXiv:2505.24203*, 2025.

Mahmoud, A. H., Masters, M., Lee, S. J., and Lill, M. A. Accurate sampling of macromolecular conformations using adaptive deep learning and coarse-grained representation. *Journal of Chemical Information and Modeling*, 62(7): 1602–1617, 2022.

McPartlon, M. and Xu, J. An end-to-end deep learning method for protein side-chain packing and inverse folding. *Proceedings of the National Academy of Sciences*, 120 (23):e2216438120, 2023.

Midgley, L. I., Stimper, V., Simm, G. N., Schölkopf, B., and Hernández-Lobato, J. M. Flow annealed importance sampling bootstrap. *arXiv preprint arXiv:2208.01893*, 2022.

Mirarchi, A., Giorgino, T., and De Fabritiis, G. mdcath: A large-scale md dataset for data-driven computational biophysics. *Scientific Data*, 11(1):1299, 2024.

Nam, J., Máté, B., Toshev, A. P., Kaniselvan, M., Gómez-Bombarelli, R., Chen, R. T., Wood, B., Liu, G.-H., and Miller, B. K. Enhancing diffusion-based sampling with molecular collective variables. *arXiv preprint arXiv:2510.11923*, 2025.

Neal, R. M. Annealed importance sampling. *Statistics and computing*, 11(2):125–139, 2001.

Nguyen, H., Roe, D. R., and Simmerling, C. Improved generalized born solvent model parameters for protein simulations. *Journal of chemical theory and computation*, 9(4):2020–2034, 2013.

Noé, F., Olsson, S., Köhler, J., and Wu, H. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science*, 365(6457):eaaw1147, 2019.

Robustelli, P., Piana, S., and Shaw, D. E. Developing a molecular dynamics force field for both folded and disordered protein states. *Proceedings of the National Academy of Sciences*, 115(21):E4758–E4766, 2018.

Schopmans, H. and Friederich, P. Temperature-annealed boltzmann generators. *arXiv preprint arXiv:2501.19077*, 2025.

Schweke, H., Pacesa, M., Levin, T., Goverde, C. A., Kumar, P., Duhoo, Y., Dornfeld, L. J., Dubreuil, B., Georgeon, S., Ovchinnikov, S., et al. An atlas of protein homo-oligomerization across domains of life. *Cell*, 187(4): 999–1010, 2024.

Shaw, D. E., Maragakis, P., Lindorff-Larsen, K., Piana, S., Dror, R. O., Eastwood, M. P., Bank, J. A., Jumper, J. M., Salmon, J. K., Shan, Y., et al. Atomic-level characterization of the structural dynamics of proteins. *Science*, 330 (6002):341–346, 2010.

Sillitoe, I., Bordin, N., Dawson, N., Waman, V. P., Ashford, P., Scholes, H. M., Pang, C. S., Woodridge, L., Rauer, C., Sen, N., et al. Cath: increased structural coverage of functional space. *Nucleic acids research*, 49(D1):D266–D273, 2021.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Sterckx, Y. G., Volkov, A. N., Vranken, W. F., Kragelj, J., Jensen, M. R., Buts, L., Garcia-Pino, A., Jové, T., Van Melderen, L., Blackledge, M., et al. Small-angle x-ray scattering-and nuclear magnetic resonance-derived conformational ensemble of the highly flexible antitoxin paaa2. *Structure*, 22(6):854–865, 2014.

Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., and Liu, Y. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Tsuboyama, K., Dauparas, J., Chen, J., Laine, E., Mohseni Behbahani, Y., Weinstein, J. J., Mangan, N. M., Ovchinnikov, S., and Rocklin, G. J. Mega-scale experimental analysis of protein folding stability in biology and design. *Nature*, 620(7973):434–444, 2023.

Van Den Oord, A., Vinyals, O., et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Vanommeslaeghe, K., Hatcher, E., Acharya, C., Kundu, S., Zhong, S., Shim, J., Darian, E., Guvench, O., Lopes, P., Vorobyov, I., et al. Charmm general force field: A force field for drug-like molecules compatible with the charmm all-atom additive biological force fields. *Journal of computational chemistry*, 31(4):671–690, 2010.

Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., et al. Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic acids research*, 50(D1):D439–D444, 2022.

Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A., and Case, D. A. Development and testing of a general amber force field. *Journal of computational chemistry*, 25(9): 1157–1174, 2004.

Wang, Y., Wang, L., Shen, Y., Wang, Y., Yuan, H., Wu, Y., and Gu, Q. Protein conformation generation via force-guided se (3) diffusion models. *arXiv preprint arXiv:2403.14088*, 2024.

Wayment-Steele, H. K., Ojoawo, A., Otten, R., Apitz, J. M., Pitsawong, W., Hömberger, M., Ovchinnikov, S., Colwell, L., and Kern, D. Predicting multiple conformations via sequence clustering and alphafold2. *Nature*, 625(7996): 832–839, 2024.

Whisstock, J. C. and Lesk, A. M. Prediction of protein function from protein sequence and structure. *Quarterly reviews of biophysics*, 36(3):307–340, 2003.

Zheng, S., He, J., Liu, C., Shi, Y., Lu, Z., Feng, W., Ju, F., Wang, J., Zhu, J., Min, Y., et al. Predicting equilibrium distributions for molecular systems with deep learning. *Nature Machine Intelligence*, 6(5):558–567, 2024.