Boltz-Jump: Accelerated Sampling of the Conformational Landscape of Biomolecular Structure Prediction Models

Anonymous Authors[1]

Abstract

Conformational ensembles provide valuable insight into the properties and function of biomolecules beyond what can be obtained from single structures. Diffusion-based biomolecular structure models such as AlphaFold 3 (Abramson et al., 2024) and Boltz-2 (Passaro et al., 2025) support generation of an ensemble of structures via repeated diffusion sampling. However, these models have two main limitations for this purpose. First, these models are mostly trained on static 3D structures; even with perfect sampling, the model distribution is not expected to exactly match the distribution from molecular dynamics simulations. Second, sampling ensembles with these models is computationally expensive due to the many function evaluations required to generate the reverse diffusion process. Here, we address the second issue by introducing Boltz-Jump, a method that accelerates the generation of conformational ensembles by up to 10× using the Boltz-2 model *without additional training* using walk-jump sampling (Daigavane et al., 2025; Saremi & Hyvärinen, 2019). Ensembles generated by Boltz-Jump also show significantly improved ability over Boltz-2 ensembles to replicate ensemble properties such as predicting exposed residues, weak and transient contacts in the ATLAS (Vander Meersche et al., 2023) and mdCATH (Mirarchi et al., 2024) datasets. Since Boltz-Jump directly leverages the open-source Boltz-2 model, it supports sampling ensembles for all biomolecular systems (protein, small molecules, and nucleic acids) supported by Boltz-2, as well as steering the sampling process using user-defined potentials (e.g. for guidance towards physically realistic structures). Finally, Boltz-Jump enables the sampling of *folding and unfolding trajectories* for small proteins, unlocking new capabilities beyond the base model.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Submitted to the 2026 Workshop on Generative and Agentic AI for Biology (ICML 2026). Do not distribute.

1. Introduction

Proteins are highly functionally diverse and perform vital roles in human biology making them a main focus of drug discovery. The properties of proteins are in principle entirely contained in their one-dimensional amino acid sequence. However, many of these properties are complicated functions of sequence, and are better understood by considering the distribution of structures these biomolecules adopt in 3D space. Experimentally determined crystal structures frequently closely resemble the most stable configuration of the protein. AlphaFold2 (Jumper et al., 2021) was trained on crystal structures deposited in the Protein Data Bank (Berman et al., 2000) and made a significant breakthrough in the CASP14 competition for protein crystal structure prediction. However, most pharmaceutically important properties, such as function, developability of antibodies, druggability of targets, and binding affinities to various drug-like moieties are not captured in a single structure alone.

In equilibrium, conformations are sampled from the Boltzmann distribution. The native state can be thought of as a high-probability conformation in this ensemble that is often closely related to the crystallized structure. In contrast, the properties we care about are often governed by other diverse structures with lower probability, or are integrals over the Boltzmann distribution of structures. Since proteins with significant sequence homology can frequently exhibit different native structures that correspond to metastable states for their homologues, it is likely that non-local information about metastable states have been learned by folding models (Roney et al., 2025; Roney & Ovchinnikov, 2022). One approach that leverages this hypothesis is MSA (multiple sequence alignment) subsampling (del Alamo et al., 2022; Wayment-Steele et al., 2024) which subsamples the MSA sequences which are evolutionarily related to the input sequences in order to elicit alternative structures. While not strictly principled, MSA subsampling introduces some diversity that can be biologically relevant or dynamically informative (Vani et al., 2023; 2024).

The general approach of exploiting pre-trained models developed originally for single structure prediction to predict ensembles or trajectories has been previously explored in several prior works. AlphaFlow (Jing et al., 2024) converts the non-generative AlphaFold2 model into a genera-

tive model over Cα atoms by adding an input embedding stack (similar to the template input stack); the resulting AlphaFlow samples outperform MSA subsampling. Conformix (Richman et al., 2025) uses Boltz-1 (Wohlwend et al., 2024), combined with guidance and twisted particle sampling to significantly increase diversity in sampled structures. On the other hand, models such as BioEmu (Lewis et al., 2024) are explicitly trained on large datasets of MD data across a variety of simulation conditions.

AlphaFold 3 (Abramson et al., 2024) style models such as Boltz (Wohlwend et al., 2024; Passaro et al., 2025), Chai (Chai Discovery, 2024), and Protenix (Zhang et al., 2026) are all diffusion models which sample structures by reversing a forward noising process using a learned score function. Walk-Jump (Saremi & Hyvärinen, 2019) is another score based approach to generative modeling which generates samples by performing Langevin dynamics in a smoothed space followed by denoising using Tweedie's formula (Miyasawa, 1960; Robbins, 1956) which has recently demonstrated promise for molecular ensemble generation (Daigavane et al., 2025). Importantly walk-jump requires only the same score function that is already learned by diffusion models (technically it requires somewhat less since the score function is only evaluated at one noise level). Therefore walk-jump can be immediately applied as an alternative sampling scheme to pre-trained diffusion models. We aim to explore the distributions obtained from this sampling scheme compared to diffusion, and also from other models that have explicitly been trained on large amounts of MD data for the purposes of ensemble generation.

We choose Boltz-2 because 1) it is open-source and 2) it has been trained on a small amount of molecular dynamics data. However, we wish to emphasize that our method is general enough to be applied to any diffusion-based protein structure model operating over all-atom coordinates (e.g. Chai (Chai Discovery, 2024), AlphaFold 3 (Abramson et al., 2024), Protenix (Zhang et al., 2026)).

2. Methods

Walk-jump sampling (Saremi & Hyvärinen, 2019) has already been applied to discrete sequence generation (Frey et al., 2024), small molecule generation (Pinheiro et al., 2024b;a) and molecular dynamics sampling (Daigavane et al., 2025). Our work builds on JAMUN (Daigavane et al., 2025), which trained a walk-jump sampling model on MD simulations of small peptides, demonstrating an order-of-magnitude faster sampling than standard molecular dynamics. Although JAMUN showed transferability to unseen peptides and across sequence lengths it was trained on converged molecular dynamics trajectories which are expensive to generate at scale for larger biomolecules. Here, we directly utilize the open-source Boltz-2 model *with no additional training*. Boltz-2 supports predictions for monomeric proteins, protein-protein complexes, DNA, RNA, and small molecules, which greatly expands the scope of our applications.

2.1. Walk-Jump Sampling

As before, p_X is the Boltzmann distribution of the clean data X. For a given noise level σ, let p_Y be the distribution of $Y = X + \sigma\varepsilon$ where $\varepsilon \sim \mathcal{N}(0, \mathbb{I}_{N\times 3})$, corresponding to the noisy data Y.

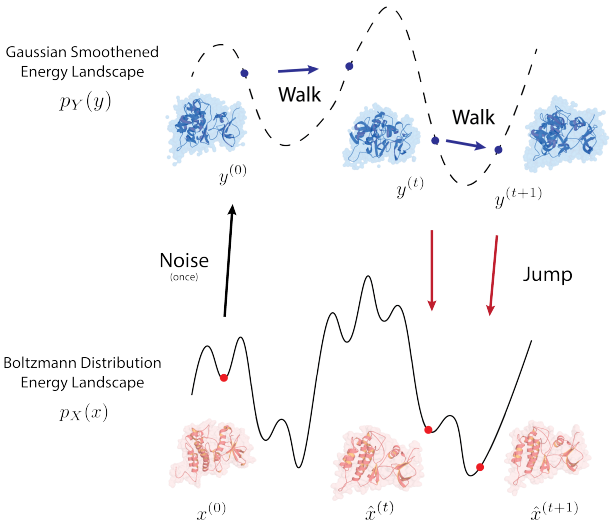


Figure 1. Overview of the walk-jump sampling process. The output is a denoised trajectory $\{\hat{x}^{(t)}\}_{t=0}^{T}$.

A conceptual overview of the walk-jump sampling process is shown in Figure 1. Given the initial sample $x^{(0)} \sim p_X$, walk-jump sampling performs the following steps:

1. **Noise** the initial structure $x^{(0)}$ to create the initial sample $y^{(0)}$ from the noisy data distribution p_Y (Figure 1):
$$y^{(0)} = x^{(0)} + \sigma\varepsilon^{(0)}, \text{ where } \varepsilon^{(0)} \sim \mathcal{N}(0, \mathbb{I}_{N\times 3}). \quad (1)$$

2. **Walk** to obtain samples $y^{(1)}, \ldots, y^{(N)}$ from p_Y using Langevin dynamics which consists of numerically solving the following Stochastic Differential Equation (SDE) (Figure 1):
$$dy = v_y dt, \quad (2)$$
$$dv_y = \nabla_y \log p_Y(y) dt - \gamma v_y dt + M^{-\frac{1}{2}}\sqrt{2} dB_t, \quad (3)$$
where v_y represents the particle velocity, $\nabla_y \log p_Y(y)$ is the gradient of the log of the probability density function (called the score function) of p_Y, γ is friction, M is the mass, and B_t is the standard Wiener process in $N \times 3$-dimensions: $B_t \sim \mathcal{N}(0, t\mathbb{I}_{N\times 3})$. In practice, we employ the BAOAB(Leimkuhler & Matthews,

2013) discretized solver (**??**) to integrate Equation 2 numerically.

3. **Jump** back approximately to p_X to obtain samples $\hat{x}_1, \ldots, \hat{x}_N$ (Figure 1) :

$$\hat{x}_i = \hat{x}(y_i; \sigma) = \mathbb{E}[X \mid Y = y_i], \tag{4}$$

where $\hat{x}(\cdot) \equiv \mathbb{E}[X \mid Y = \cdot]$ is called the denoiser. It corresponds to the minimizer of the ℓ_2-loss between clean samples X and samples denoised back from $Y = X + \sigma\varepsilon$.

$$\hat{x}(\cdot; \sigma) = \arg\min_f \mathbb{E}_{\substack{X \sim p_X, \varepsilon \sim \mathcal{N}(0, \mathbb{I}_{N\times 3}) \\ Y = X + \sigma\varepsilon}} [\|f(Y) - X\|^2], \tag{5}$$

where $f : \mathbb{R}^{N\times 3} \to \mathbb{R}^{N\times 3}$. As shown by Robbins (1956); Miyasawa (1960), the denoiser $\hat{x}$ is closely linked to the score $\nabla_y \log p_Y$:

$$\hat{x}(y; \sigma) = y + \sigma^2 \nabla_y \log p_Y(y). \tag{6}$$

Since the denoiser $\hat{x}$ is already parametrized by Boltz-2 in its `AtomDiffusion` module, we can obtain the score function using Equation 6. Note that the denoiser $\hat{x}$ is conditional on the sequence representation s and pair representation z computed by the `Pairformer` stack in Boltz-2 from the given input sequence and queried MSA information.

3. Datasets

Our experiments focus on the ATLAS (Vander Meersche et al., 2023) and mdCATH (Mirarchi et al., 2024) datasets, as benchmarked in the original Boltz-2 paper. ATLAS consists of molecular dynamics simulations performed with the CHARMM36m (Huang et al., 2017) force field of 1390 monomeric protein chains, varying from 38 to 2128 amino acid residues in size, performed with three independent replicates at a temperature of 300 K and a constant pressure of 1 bar (NPT ensembles). mdCATH consists of molecular dynamics simulations performed with the CHARMM22* force field (Piana et al., 2011) of 5398 monomeric protein chains, varying from 50 to 500 residues in size, performed with five independent replicates each at five temperatures ranging between 320 K to 450 K at a constant volume obtained after equilibration under a constant pressure of 1 atm. We use the mdCATH MD simulations at 320 K for our analysis here, since Boltz-2 was trained on a small subset of simulations at this temperature.

4. Benchmarks

We utilize the comprehensive metrics defined by AlphaFlow (Jing et al., 2024). In particular, we measure:

- **Structural validity**: Precision, recall and diversity as measured by lDDT (local distance difference test) for Cα carbons for each residue, compared to the ground truth starting crystal structure.
- **Flexibility**: Root mean squared deviation (RMSD) and per-residue fluctuation (RMSF) compared to the ground truth MD ensemble.
- **Distributional accuracy**: Root mean Wasserstein distance between the predicted ensemble and the ground truth MD ensemble, projected onto the PCA components obtained from the ground truth MD ensemble.
- **Transient and weak contacts prediction**: Residue pairs which associate (or dissociate) in the ground-truth MD simulation.
- **Exposed residue prediction**: Residues which are originally hidden but eventually exposed to the solvent during the ground truth MD simulation.

We compare Boltz-Jump to AlphaFlow and BioEmu on the ATLAS and mdCATH datasets. The comparison between these models is not ideal due to the lack of consistent splits across these models. In fact, AlphaFlow was omitted from ATLAS comparisons in the original Boltz-2 paper because its training set overlapped significantly with the test set chosen by Boltz-2. Further, AlphaFlow itself was trained on a much larger subset of ATLAS data than Boltz-2 was. The exact training splits for BioEmu and Boltz-2 are not publicly available; in any case, retraining these models would require significant computational resources unfeasible for the current study. For this reason, we emphasize that the crucial comparison is between Boltz-Jump and Boltz-2, since they utilize the same base model but with different sampling strategies.

We utilize the GPU-accelerated `mmseqs` software to compute the MSAs for all input sequences; these are batched, precomputed and shared across all models. This runtime is not accounted for in the total runtimes in Table 3. We describe the MSA generation process and provide our MSA generation script in Appendix A.

Finally, our evaluation here improves on the MD evaluation in the original Boltz-2 paper (Passaro et al., 2025); the original paper only took samples from each of Boltz-2 (with no MSA subsampling), AlphaFlow (only the base model, not distilled), and BioEmu, and these were compared to only 200 samples from the ATLAS and mdCATH trajectories. We choose the same set of 40 evaluation targets from each dataset as in the Boltz-2 MD evaluation. The exact target IDs are listed in Appendix B[1].

[1] Obtained via private correspondence with the Boltz-2 authors; listed here for reproducibility.

5. Results

We set $\sigma = 2.0\,\text{Å}$, with timestep $\delta = 2.0$ and friction $\gamma = 0.1$ in the BAOAB integrator. The timestep and friction are intricately linked to the smoothness of the underlying potential energy surface; usually higher noise levels enable larger timesteps, and friction can be reduced to decorrelate faster. Note that JAMUN was only able to work with $\sigma = 0.5\,\text{Å}$; higher noise levels would affect the quality of the denoised samples with their equivariant GNN architecture. We suspect that the conditioning of the Boltz-2 denoiser on rich sequence and pair embeddings from the `Pairformer` stack enables successful denoising at higher noise levels.

Across both ATLAS and mdCATH, Boltz-Jump is much more predictive of ensemble properties such as identifying weak contacts, transient contacts, and exposed residues, almost matching the bespoke AlphaFlow MD and BioEmu baselines at significantly improved runtime cost. Again, a fair comparison across model families is difficult due to differences in training data; hence we additionally highlight in gray a comparison across the Boltz-2 models. We find that Boltz-Jump maintains most of the recall (as measured by lDDT) but sacrifices some precision; this is expected as we only perform a one-step denoising process, which inherently introduces some error. Another concern is the ensemble distributional metrics (RMWD, PCA $\mathcal{W}_2$ and PC similarity) are diminished under walk-jump sampling. This suggests that walk-jump sampling may be exploring a broader region of conformational space than the ground-truth MD ensemble. However, whether the ground-truth MD ensembles are themselves converged remains unclear. In the future, we plan to benchmark *walk-denoise*, where we follow the reverse diffusion process starting at our chosen noise level σ, which allows the model to correct for these errors.

Due to the use of a single noise level σ throughout the sampling process, Boltz-Jump samples ensembles much faster than any of the baselines, as seen in Table 3. In particular, we observe $5\times$ to $10\times$ faster sampling over the usual full diffusion process which restarts each sample from pure noise.

5.1. Folding Trajectories

Since walk-jump requires a seed structure (just as molecular dynamics trajectories are seeded) to begin sampling, we can sample folding trajectories with Boltz-Jump by initializing the walk-jump trajectory from an unfolded state. Figure 2 shows an example of a folding trajectory for Trp-cage (PDB: `1L2Y`) generated using Boltz-Jump *with no guidance*.

5.2. Biased Sampling with Steering Potentials

In practical situations, we often wish to bias the sampling process to produce samples satisfying certain constraints. One popular method to do this is to adjust the score by adding a force $-\lambda\nabla U_{\text{steer}}$ defined by a steering potential U_{steer} and a strength factor λ; for example, to ensure that certain residues remain bonded. Since we directly build on the Boltz-2 repository, we directly inherit all of the steering potentials supported by Boltz-2. We also added another steering potential to steer the model towards a target state, such as the folded state of a protein, by simply adding the per-atom deviations upto a 1Å threshold. In the full diffusion sampling, the strength λ of the steering potential needs to be adjusted constantly according to a tuned schedule as the noise level σ is reduced; otherwise the sampling collapses onto a few modes. A key advantage of walk-jump sampling is the ability to set a single value for the strength λ throughout the sampling process, removing the need to tune an entire schedule. We demonstrate this by simulating the unfolding trajectory (which would not occur in any reasonable amount of time in an MD simulation) from the folded-to-unfolded state of Trp-cage with Boltz-Jump in Figure 3. The visible kink in the helix is due to a proline residue, which is a known helix breaker due to its unique cyclic structure; Boltz-Jump captures this nuance and does not fully unfold the structure into a simple α-helix.

Similarly, in Figure 4 we initialize from the inactive form of Abl Kinase (PDB: `6XR6`) and guide it towards its (conformationally different) active form (PDB: `6XR7`). The secondary structure remains well-defined throughout this trajectory.

6. Conclusion

Here, we have introduced Boltz-Jump, leveraging the open-source Boltz-2 protein structure prediction model to perform walk-jump sampling. We demonstrated several advantages of this strategy; accelerated sampling of the conformational ensemble which better captures transient and weak contacts, the ability to sample folding trajectories without guidance and unfolding trajectories using guidance. All of these capabilities were performed using the original Boltz-2 model with no additional training. In the future, we plan to benchmark the folding trajectories sampled by our method, and further understand the tradeoffs of walk-jump sampling.

Table 1. Conformational ensemble metrics on ATLAS. **Bold**: best overall. Gray: best among Boltz-2 variants. ↑: higher is better, ↓: lower is better.

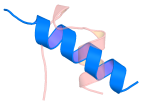
Metric	Boltz-Jump	Boltz-2 Full MSA	Boltz-2 MSA 32	Boltz-2 MSA 64	AlphaFlow MD Base	AlphaFlow MD Distilled	BioEmu
lDDT Precision ↑	0.735	**0.871**	0.860	0.867	0.834	0.811	0.830
lDDT Recall ↑	0.819	0.825	0.821	0.824	**0.831**	0.801	0.819
lDDT Diversity ↑	**0.268**	0.017	0.029	0.022	0.135	0.085	0.182
Pairwise RMSD ρ ↑	0.611	0.622	0.438	0.508	0.610	**0.718**	0.548
Per-residue RMSF ρ ↑	0.559	0.633	0.631	0.624	**0.711**	0.703	0.639
RMWD ↓	4.59	3.98	4.05	3.98	**3.52**	3.74	3.72
PCA $\mathcal{W}_2$ ↓	3.20	2.32	2.35	2.33	**1.72**	2.05	2.34
PC similarity ↑	7.5	30.0	27.5	25.0	**45.0**	32.5	40.0
Weak contacts J ↑	0.531	0.217	0.276	0.234	**0.643**	0.520	0.541
Transient contacts J ↑	0.258	0.263	0.268	0.267	**0.429**	0.299	0.360
Exposed residue J ↑	0.489	0.327	0.327	0.333	**0.719**	0.609	0.591
Exposed MI ρ ↑	0.239	0.085	0.095	0.102	**0.287**	0.134	0.260

Table 2. Conformational ensemble metrics on mdCATH at 320 K. **Bold**: best overall. Gray: best among Boltz-2 variants. ↑: higher is better, ↓: lower is better.

Metric	Boltz-Jump	Boltz-2 Full MSA	Boltz-2 MSA 32	Boltz-2 MSA 64	AlphaFlow MD Base	AlphaFlow MD Distilled	BioEmu
lDDT Precision ↑	0.751	**0.848**	0.841	0.842	0.786	0.776	0.741
lDDT Recall ↑	0.753	0.757	**0.764**	0.760	0.751	0.730	0.740
lDDT Diversity ↑	0.239	0.026	0.033	0.028	0.167	0.078	**0.260**
Pairwise RMSD ρ ↑	0.627	0.550	0.612	**0.629**	0.323	0.335	0.526
Per-residue RMSF ρ ↑	0.622	0.632	0.607	0.622	0.640	0.628	**0.660**
RMWD ↓	4.21	4.62	4.55	4.30	4.27	4.77	**3.97**
PCA $\mathcal{W}_2$ ↓	2.75	2.77	2.70	**2.54**	2.68	3.11	2.59
PC similarity ↑	15.0	25.0	22.5	17.5	22.5	**27.5**	20.0
Weak contacts J ↑	0.590	0.258	0.308	0.262	0.564	0.501	**0.618**
Transient contacts J ↑	0.304	0.219	0.214	0.238	**0.357**	0.238	0.339
Exposed residue J ↑	0.625	0.374	0.408	0.435	**0.650**	0.571	0.631
Exposed MI ρ ↑	0.264	0.041	0.048	0.035	0.231	0.109	**0.376**

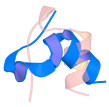
Table 3. Wall-clock runtime (minutes) for generating conformational ensembles. Boltz-Jump achieves significant speedups over all baselines.

Dataset	Method	Min	Mean	Median	Max
ATLAS	**Boltz-Jump**	**0.37**	**0.97**	**0.95**	**1.47**
	Boltz-2 Full MSA	1.35	4.97	4.48	12.72
	Boltz-2 MSA 32	1.32	4.91	4.45	12.57
	Boltz-2 MSA 64	1.37	4.90	4.43	12.57
	AlphaFlow MD Distilled	1.83	8.99	7.39	30.03
	AlphaFlow MD Base	7.15	79.06	63.01	288.75
	BioEmu Sample	0.40	4.34	3.73	14.88
	BioEmu Relax	9.67	38.06	34.53	89.25
mdCATH	**Boltz-Jump**	**0.73**	**0.79**	**0.78**	**0.87**
	Boltz-2 Full MSA	1.43	2.68	2.51	5.15
	Boltz-2 MSA 32	1.45	2.68	2.50	5.13
	Boltz-2 MSA 64	1.43	2.63	2.47	5.10
	AlphaFlow MD Distilled	1.33	3.28	2.80	8.22
	AlphaFlow MD Base	8.42	7.53	22.97	75.48
	BioEmu Sample	1.48	2.82	2.53	5.70
	BioEmu Relax	10.68	19.97	17.73	35.10



(a) Frame 0

(b) Frame 200

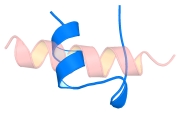


(c) Frame 400

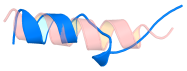
(d) Frame 600

(e) Frame 800

Figure 2. Structural snapshots from the Boltz-Jump trajectory for Trp-cage (PDB: `1L2Y`), starting from the unfolded structure (blue) and reaching the folded structure *with no guidance*.

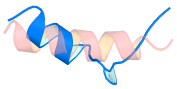


(a) Frame 0

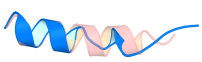


(b) Frame 200

(c) Frame 400

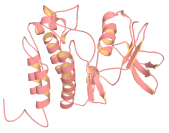


(d) Frame 600



(e) Frame 800

Figure 3. Structural snapshots from the Boltz-Jump trajectory for Trp-cage (PDB: `1L2Y`), starting from the folded structure (blue) and reaching the unfolded structure *with guidance*, due to thermodynamic unfavorability of the unguided process.



(a) Frame 0

(b) Frame 200

(c) Frame 400

(d) Frame 600

(e) Frame 800

Figure 4. Structural snapshots from the Boltz-Jump trajectory for Abl Kinase, starting from the inactive form (PDB: `6XR6`) and guided towards the (conformationally different) active form (PDB: `6XR7`) *with guidance*.

References

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., Bodenstein, S. W., Evans, D. A., Hung, C.-C., O'Neill, M., Reiman, D., Tunyasuvunakool, K., Wu, Z., Žemgulytė, A., Arvaniti, E., Beattie, C., Bertolli, O., Bridgland, A., Cherepanov, A., Congreve, M., Cowen-Rivers, A. I., Cowie, A., Figurnov, M., Fuchs, F. B., Gladman, H., Jain, R., Khan, Y. A., Low, C. M. R., Perlin, K., Potapenko, A., Savy, P., Singh, S., Stecula, A., Thillaisundaram, A., Tong, C., Yakneen, S., Zhong, E. D., Zielinski, M., Žídek, A., Bapst, V., Kohli, P., Jaderberg, M., Hassabis, D., and Jumper, J. M. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, Jun 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07487-w. URL https://doi.org/10.1038/s41586-024-07487-w.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. The Protein Data Bank. *Nucleic Acids Research*, 28(1): 235–242, 01 2000. ISSN 0305-1048. doi: 10.1093/nar/28.1.235. URL https://doi.org/10.1093/nar/28.1.235.

Chai Discovery. Chai-1: Decoding the molecular interactions of life. *bioRxiv*, 2024. doi: 10.1101/2024.10.10.615955. URL https://www.biorxiv.org/content/early/2024/10/11/2024.10.10.615955.

Consortium, T. U. Uniprot: the universal protein knowledgebase in 2025. *Nucleic Acids Research*, 53(D1): D609–D617, 11 2024. ISSN 1362-4962. doi: 10.1093/nar/gkae1010. URL https://doi.org/10.1093/nar/gkae1010.

Daigavane, A., Vani, B. P., Davidson, D., Saremi, S., Rackers, J. A., and Kleinhenz, J. JAMUN: Bridging smoothed molecular dynamics and score-based learning for conformational ensemble generation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=8Z3KnaYtw9.

del Alamo, D., Sala, D., Mchaourab, H. S., and Meiler, J. Sampling alternative conformational states of transporters and receptors with alphafold2. *eLife*, 11:e75751, mar 2022. ISSN 2050-084X. doi: 10.7554/eLife.75751. URL https://doi.org/10.7554/eLife.75751.

Frey, N. C., Berenberg, D., Kleinhenz, J., Hotzel, I., Lafrance-Vanasse, J., Kelly, R. L., Wu, Y., Rajpal, A., Ra, S., Bonneau, R., Cho, K., Loukas, A., Gligorijevic, V., and Saremi, S. Protein discovery with discrete walk-jump sampling. In *International Conference on Learning Representations*, 2024.

Huang, J., Rauscher, S., Nawrocki, G., Ran, T., Feig, M., de Groot, B. L., Grubmüller, H., and MacKerell, A. D. Charmm36m: an improved force field for folded and intrinsically disordered proteins. *Nature Methods*, 14(1): 71–73, Jan 2017. ISSN 1548-7105. doi: 10.1038/nmeth.4067. URL https://doi.org/10.1038/nmeth.4067.

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

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. Highly accurate protein structure prediction with alphafold. *Nature*, 596 (7873):583–589, Aug 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2. URL https://doi.org/10.1038/s41586-021-03819-2.

Leimkuhler, B. and Matthews, C. Robust and efficient configurational molecular sampling via langevin dynamics. *The Journal of Chemical Physics*, 138(17):174102, May 2013. doi: 10.1063/1.4802990.

Lewis, S., Hempel, T., Jiménez-Luna, J., Gastegger, M., Xie, Y., Foong, A. Y. K., Satorras, V. G., Abdin, O., Veeling, B. S., Zaporozhets, I., Chen, Y., Yang, S., Schneuing, A., Nigam, J., Barbero, F., Stimper, V., Campbell, A., Yim, J., Lienen, M., Shi, Y., Zheng, S., Schulz, H., Munir, U., Clementi, C., and Noé, F. Scalable emulation of protein equilibrium ensembles with generative deep learning. *bioRxiv*, 2024. doi: 10.1101/2024.12.05.626885.

Mirarchi, A., Giorgino, T., and De Fabritiis, G. mdcath: A large-scale md dataset for data-driven computational biophysics. *Scientific Data*, 11(1):1299, Nov 2024. ISSN 2052-4463. doi: 10.1038/s41597-024-04140-z. URL https://doi.org/10.1038/s41597-024-04140-z.

Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., and Steinegger, M. Colabfold: making protein folding accessible to all. *Nature Methods*, 19(6): 679–682, Jun 2022. ISSN 1548-7105. doi: 10.1038/s41592-022-01488-1. URL https://doi.org/10.1038/s41592-022-01488-1.

Miyasawa, K. An Empirial Bayes Estimator of the Mean of a Normal Population. *Bulletin de l'Institut international de statistique.*, 38(4):181–188, 1960.

Passaro, S., Corso, G., Wohlwend, J., Reveiz, M., Thaler, S., Somnath, V. R., Getz, N., Portnoi, T., Roy, J., Stark, H., Kwabi-Addo, D., Beaini, D., Jaakkola, T., and Barzilay, R. Boltz-2: Towards accurate and efficient binding affinity prediction. *bioRxiv*, 2025. doi: 10.1101/2025.06.14.659707. URL https://www.biorxiv.org/content/early/2025/06/18/2025.06.14.659707.

Piana, S., Lindorff-Larsen, K., and Shaw, D. E. How robust are protein folding simulations with respect to force field parameterization? *Biophysical Journal*, 100(9):L47–L49, 2011. ISSN 0006-3495. doi: https://doi.org/10.1016/j.bpj.2011.03.051. URL https://www.sciencedirect.com/science/article/pii/S0006349511004097.

Pinheiro, P. O., Jamasb, A., Mahmood, O., Sresht, V., and Saremi, S. Structure-based Drug Design by Denoising Voxel Grids. *arXiv preprint arXiv:2405.03961*, 2024a.

Pinheiro, P. O., Rackers, J., Kleinhenz, J., Maser, M., Mahmood, O., Watkins, A., Ra, S., Sresht, V., and Saremi, S. 3D molecule generation by denoising voxel grids. *Advances in Neural Information Processing Systems*, 36, 2024b.

Richman, D. D., Karaguesian, J., Suomivuori, C.-M., and Dror, R. O. Unlocking hidden biomolecular conformational landscapes in diffusion models at inference time. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=U87XyMPrZp.

Robbins, H. An Empirical Bayes Approach to Statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, volume 3.1, 1956.

Roney, J. P. and Ovchinnikov, S. State-of-the-art estimation of protein model accuracy using alphafold. *Phys. Rev. Lett.*, 129:238101, Nov 2022. doi: 10.1103/PhysRevLett.129.238101. URL https://link.aps.org/doi/10.1103/PhysRevLett.129.238101.

Roney, J. P., Ou, C., and Ovchinnikov, S. Protein diffusion models as statistical potentials. *bioRxiv*, 2025. doi: 10.64898/2025.12.09.693073. URL https://www.biorxiv.org/content/early/2025/12/09/2025.12.09.693073.

Saremi, S. and Hyvärinen, A. Neural Empirical Bayes. *Journal of Machine Learning Research*, 20(181):1–23, 2019.

Steinegger, M. and Söding, J. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11):1026–1028, Nov 2017. ISSN 1546-1696. doi: 10.1038/nbt.3988. URL https://doi.org/10.1038/nbt.3988.

Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R., and Wu, C. H. Uniref: comprehensive and non-redundant uniprot reference clusters. *Bioinformatics*, 23(10):1282–1288, 03 2007. ISSN 1367-4803. doi: 10.1093/bioinformatics/btm098. URL https://doi.org/10.1093/bioinformatics/btm098.

Vander Meersche, Y., Cretin, G., Gheeraert, A., Gelly, J.-C., and Galochkina, T. Atlas: protein flexibility description from atomistic molecular dynamics simulations. *Nucleic Acids Research*, 52(D1):D384–D392, 11 2023. ISSN 0305-1048. doi: 10.1093/nar/gkad1084. URL https://doi.org/10.1093/nar/gkad1084.

Vani, B. P., Aranganathan, A., Wang, D., and Tiwary, P. AlphaFold2-RAVE: From sequence to boltzmann ranking. *J. Chem. Theory Comput.*, 19(14):4351–4354, July 2023.

Vani, B. P., Aranganathan, A., and Tiwary, P. Exploring kinase asp-phe-gly (dfg) loop conformational stability with alphafold2-rave. *Journal of Chemical Information and Modeling*, 64(7):2789–2797, Apr 2024. ISSN 1549-9596. doi: 10.1021/acs.jcim.3c01436. URL https://doi.org/10.1021/acs.jcim.3c01436.

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, Jan 2024. ISSN 1476-4687. doi: 10.1038/s41586-023-06832-9. URL https://doi.org/10.1038/s41586-023-06832-9.

Wohlwend, J., Corso, G., Passaro, S., Reveiz, M., Leidal, K., Swiderski, W., Portnoi, T., Chinn, I., Silterra, J., Jaakkola, T., and Barzilay, R. Boltz-1 democratizing biomolecular interaction modeling. *bioRxiv*, 2024. doi: 10.1101/2024.11.19.624167. URL https://www.biorxiv.org/content/early/2024/11/20/2024.11.19.624167.

Zhang, Y., Gong, C., Zhang, H., Ma, W., Liu, Z., Chen, X., Guan, J., Wang, L., Yang, Y., Xia, Y., and Xiao, W. Protenix-v1: Toward high-accuracy open-source biomolecular structure prediction. *bioRxiv*, 2026. doi: 10.64898/2026.02.05.703733. URL https://www.biorxiv.org/content/early/2026/02/22/2026.02.05.703733.1.

A. Multiple Sequence Alignment (MSA) Generation

We utilize the GPU-accelerated `mmseqs` (Steinegger & Söding, 2017) package to query the ColabFold (Mirdita et al., 2022) and UniRef30 (Consortium, 2024; Suzek et al., 2007) databases:

```
#!/bin/bash

if [[ $# < 1 ]]; then
  echo "Must pass name of FASTA file to $0"
  exit
fi
echo "Reading sequences from $1"
FASTA=$1
OUTDIR="${FASTA%.*}_msa"
mkdir -p $OUTDIR

mmseqs gpuserver /data2/colabfold_db/database/colabfold_envdb_202108_db \
  --max-seqs 10000 --db-load-mode 3 --prefilter-mode 1 & PID1=$! \
&& mmseqs gpuserver /data2/colabfold_db/database/uniref30_2302_db \
  --max-seqs 10000 --db-load-mode 3 --prefilter-mode 1 & PID2=$! \
&& sleep 60

colabfold_search $FASTA /data2/colabfold_db/database/ $OUTDIR \
  --gpu 1 --gpu-server 1 --db-load-mode 3
```

B. Evaluation Targets

We list the exact PDB IDs of the targets we evaluated on, matching the Boltz-2 evaluation.

Table 4. ATLAS test set: 40 protein chains used for evaluation, with sequence lengths ranging from 40 to 567 residues.

PDB ID	Length	PDB ID	Length
4jnu_B	40	2qia_A	262
1ptq_A	50	1ssq_D	267
4pz1_A	98	3ho6_A	267
2p9x_D	99	7p46_A	282
7s86_A	99	1hq0_A	295
2bo1_A	101	1d3y_B	301
6tgk_C	105	1r6w_A	322
1o4k_A	108	3dpg_B	338
1jif_A	122	3lpc_A	340
2w0g_A	129	1l5o_A	356
2pag_A	135	4gv2_A	357
7ead_A	135	7aqx_A	364
2yvq_A	143	7qsu_A	374
3cy4_A	154	7dmn_A	377
2igi_A	180	1dlj_A	402
3frr_A	191	6pnv_A	411
1juv_A	193	1zjc_A	418
3o3x_A	198	7wab_A	484
5w2f_A	206	3djl_A	541
7rm7_A	228	5znj_A	567

Table 5. mdCATH test set: 40 protein chains evaluated at 320 K, with sequence lengths ranging from 52 to 280 residues.

PDB ID	Length	PDB ID	Length
3qiiA00	52	3i32A02	125
1d2dA00	56	2kkmA01	125
2hg7A00	60	4pt1B00	128
2kk2A00	61	3r5dA01	129
3c6fA01	62	2kkuA00	139
1lddA00	74	1zxfA00	155
4i6uB00	77	1f8yA00	156
2k5iA02	78	1uf2C02	157
4i69A00	79	3ufbA01	159
1wosA04	86	2oudA00	177
2gzoA02	89	2kc3A00	183
3fg6A02	94	3jciA00	190
1wwjA00	99	3nv0A00	196
2hgkA01	105	2a7kB01	197
1q6aA00	107	1a87A02	200
3wy8A01	113	1d0bA00	207
4jmfB00	116	4r5qA00	215
2dckA02	119	4fhdA02	227
2k54A00	123	2v0rA00	235
3i3lA02	125	4b9bA02	280