Scalable Normalizing Flows Enable Boltzmann Generators for Macromolecules

Joseph Kim[1,2], David Bloore[1], Karan Kapoor[1], Jun Feng[1], Mengdi Wang[2], Ming-Hong Hao[1].
[1]Ensem Therapeutics, 888 Winter St, Waltham, MA 02451
[2]Princeton University, 41 Olden St. Princeton, New Jersey 08544

Summary

Background: The structural ensemble of a protein determines its functions. The probabilities of the ground and metastable states of a protein at equilibrium for a given temperature determine the interactions of the protein with other proteins, effectors, and drugs, which are keys for pharmaceutical development.
Previous work: Traditional methods for sampling the equilibrium distribution rely on Markov-chain Monte Carlo or molecular dynamics (MD). These methods often get stuck in local minima and are sensitive to initial seeding. Previous ML-based methods employ normalizing flow models [1,2,3]. However, these methods scale poorly and empirical results are often demonstrated for small systems like alanine dipeptide. Some works have utilized coarse graining [4,5] along with normalizing flows, but these methods do not provide the resolution often required for downstream applications such as drug-design.

Method

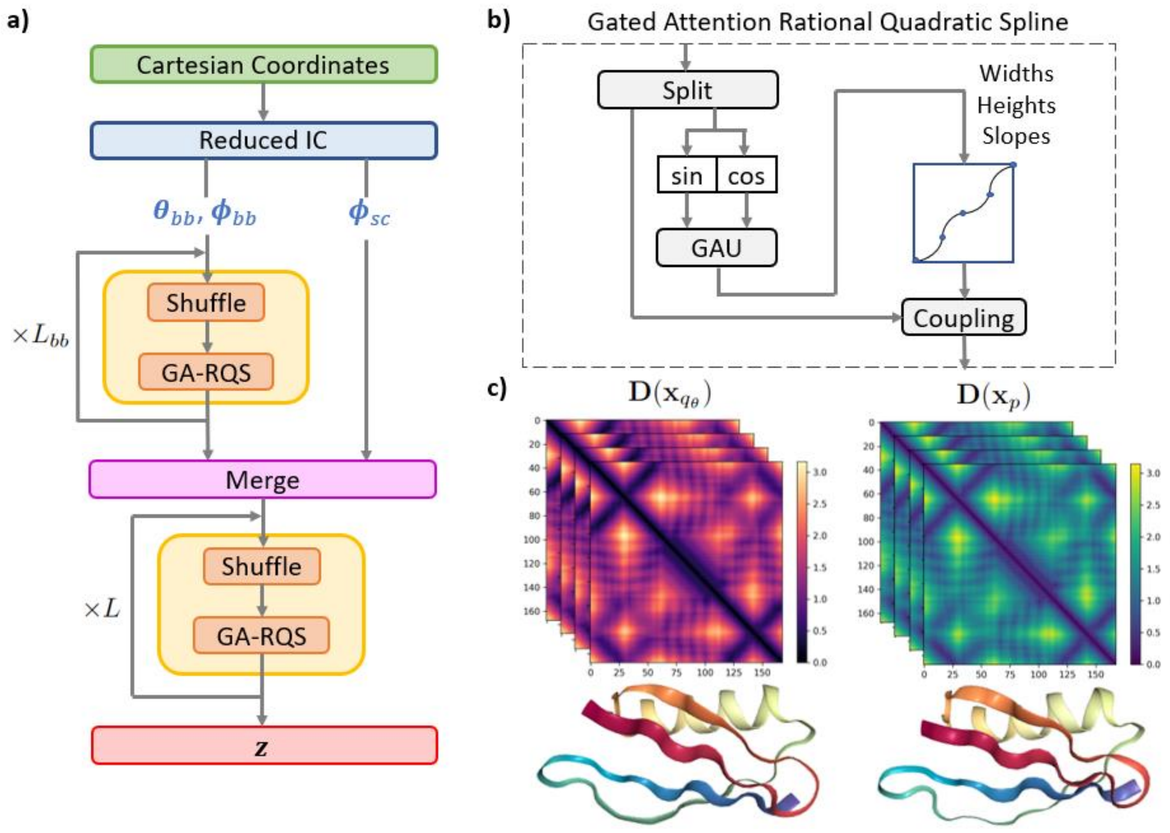

Figure 1: (a) Our split flow architecture. (b) Each transformation block consists of a gated attention rational quadratic spline (RQS) coupling layer. (c) Example structures of protein G from the flow q_θ (left) and from molecular dynamics simulation p (right). We also show sample distance matrices $\mathbf{D}(\mathbf{x}_{q_\theta})$ and $\mathbf{D}(\mathbf{x}_p)$.

Multi-stage training strategy

$$\mathcal{L}_{\mathrm{NLL}}(\theta) := -\mathbb{E}_{\mathbf{x}\sim p}[\log q_\theta(\mathbf{x})]$$
$$\mathcal{L}_{\mathrm{KL}}(\theta) := \mathbb{E}_{\mathbf{x}\sim q_\theta}[\log(q_\theta(\mathbf{x})/p(\mathbf{x}))]$$
$$\mathcal{L}_{\mathrm{W}}(\theta) := \|\boldsymbol{\mu}_{q_\theta} - \boldsymbol{\mu}_p\|_2^2 + \mathrm{trace}(\boldsymbol{\Sigma}_{q_\theta} + \boldsymbol{\Sigma}_p - 2(\boldsymbol{\Sigma}_{q_\theta}\boldsymbol{\Sigma}_p)^{1/2})$$

1. Train with $\mathcal{L} = \mathcal{L}_{\mathrm{NLL}}$.
2. Train with $\mathcal{L} = \mathcal{L}_{\mathrm{NLL}} + \lambda_{\mathrm{W}}\mathcal{L}_{\mathrm{W}}$.
3. Train with $\mathcal{L} = \mathcal{L}_{\mathrm{NLL}} + \lambda'_{\mathrm{W}}\mathcal{L}_{\mathrm{W}} + \lambda_{\mathrm{KL}}\mathcal{L}_{\mathrm{KL}}$, where $\lambda'_{\mathrm{W}} < \lambda_{\mathrm{W}}$.
4. Train with $\mathcal{L} = \mathcal{L}_{\mathrm{NLL}} + \lambda'_{\mathrm{KL}}\mathcal{L}_{\mathrm{KL}}$, where $\lambda'_{\mathrm{KL}} > \lambda_{\mathrm{KL}}$.

Results

Table 1: **Training BGs with different strategies.** We compute ΔD, energy $u(\cdot)$, and mean NLL of 10^6 generated structures after training with different training strategies with ADP, protein G, and Villin HP35. ΔD is computed for batches of 10^3 samples. Means and standard deviations are reported. Statistics for $u(\cdot)$ are reported for structures with energy below the median sample energy. Best results are bold-faced. For reference, the energy for training data structures is -317.5 ± 125.5 kcal/mol for protein G and -1215.5 ± 222.2 kcal/mol for villin HP35. We compare our results against a Neural Spline Flows (NSF) baseline model.

System	Arch.	Training strategy NLL	KL	W2	ΔD (Å)	Energy $u(\mathbf{x})$ (kcal/mol)	$-\mathbb{E}_{p(\mathbf{x})}[\log q_\theta(\mathbf{x})]$
ADP	NSF	✓			0.09 ± 0.01	$(\mathbf{-1.19 \pm 0.61}) \times \mathbf{10^1}$	38.29 ± 0.19
	Ours	✓			0.08 ± 0.01	$(\mathbf{-1.18 \pm 0.65}) \times \mathbf{10^1}$	$\mathbf{36.15 \pm 0.15}$
		✓	✓		0.05 ± 0.01	$(\mathbf{-1.20 \pm 0.59}) \times \mathbf{10^1}$	38.66 ± 0.19
		✓		✓	$\mathbf{0.04 \pm 0.00}$	$(\mathbf{-1.06 \pm 0.74}) \times \mathbf{10^1}$	38.12 ± 0.03
	Ours	✓	✓	✓	$\mathbf{0.03 \pm 0.01}$	$(\mathbf{-1.31 \pm 0.52}) \times \mathbf{10^1}$	37.67 ± 0.09
Protein G	NSF	✓			2.92 ± 0.80	$(2.15 \pm 3.31) \times 10^{10}$	-263.46 ± 0.13
	Ours	✓			1.81 ± 0.14	$(9.47 \pm 15.4) \times 10^8$	$\mathbf{-310.11 \pm 0.08}$
		✓	✓		16.09 ± 1.14	$(2.86 \pm 0.62) \times 10^2$	-308.68 ± 0.08
		✓		✓	$\mathbf{0.18 \pm 0.01}$	$(2.68 \pm 4.31) \times 10^6$	-307.17 ± 0.01
	Ours	✓	✓	✓	$\mathbf{0.19 \pm 0.01}$	$(\mathbf{-3.04 \pm 1.24}) \times \mathbf{10^2}$	$\mathbf{-309.10 \pm 0.91}$
HP35	NSF	✓			0.81 ± 0.06	$(7.78 \pm 17.4) \times 10^7$	687.95 ± 1.92
	Ours	✓			0.65 ± 0.04	$(5.29 \pm 11.7) \times 10^6$	$\mathbf{651.90 \pm 2.88}$
		✓	✓		0.61 ± 0.04	$(6.46 \pm 14.3) \times 10^2$	678.38 ± 0.87
		✓		✓	$\mathbf{0.38 \pm 0.03}$	$(1.15 \pm 1.76) \times 10^7$	678.31 ± 1.55
	Ours	✓	✓	✓	$\mathbf{0.39 \pm 0.03}$	$(\mathbf{-4.66 \pm 3.52}) \times \mathbf{10^2}$	667.45 ± 2.04

Results (cont'd)

Definition 3.1 (Distance Distortion)**.** Let $\mathcal{A}_{bb}$ denote the indices of backbone atoms. Define $\mathbf{D}(\mathbf{x})$ as the pairwise distance matrix for the backbone atoms of $\mathbf{x}$. Define $\mathcal{P} = \{(i,j)|i,j \in \mathcal{A}_{bb} \text{ and } i < j\}$. The distance distortion is defined as

$$\Delta D := \mathop{\mathbb{E}}_{\substack{\mathbf{x}_{q_\theta}\sim q_\theta \\ \mathbf{x}_p \sim p}} \left[\frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} |\mathbf{D}(\mathbf{x}_{q_\theta})_{ij} - \mathbf{D}(\mathbf{x}_p)_{ij}| \right], \tag{3}$$

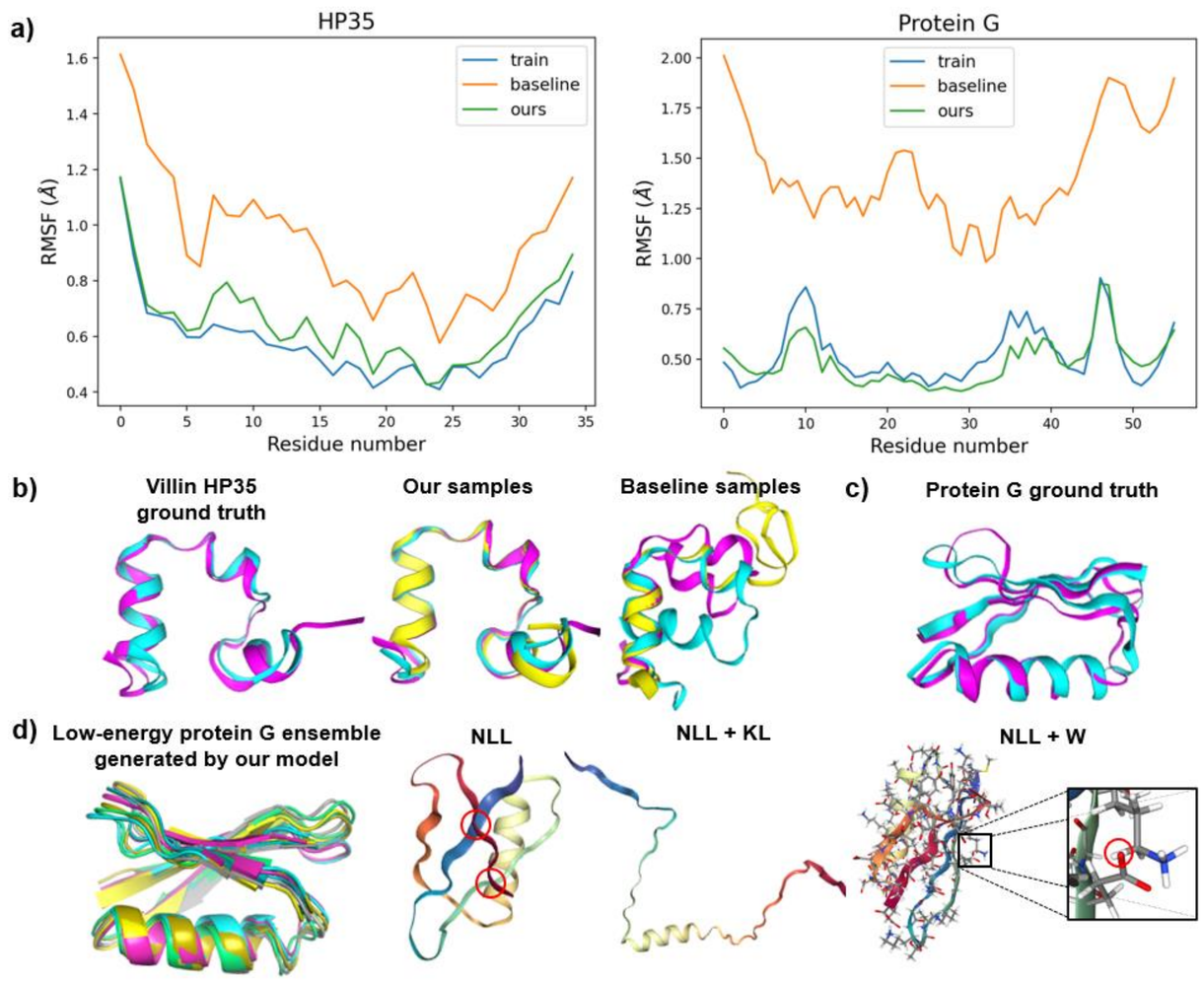

Figure 3: **Sample conformations generated by BG via different training strategies.** (a) Root mean square fluctuation (RMSF) computed for each residue (Cα atoms) in HP35 and protein G. Matching the training dataset's plot is desirable. (b) Examples of HP35 from ground truth training data, generated samples from our model, and generated samples from the baseline model. (c) Example of two metastable states from protein G training data. (d) Low-energy conformations of protein G generated by our model superimposed on each other. We also show some examples of pathological structures generated after training with different training paradigms: NLL (maximum likelihood), both NLL and KL divergence, and NLL and the 2-Wasserstein loss. Atom clashes are highlighted with red circles.

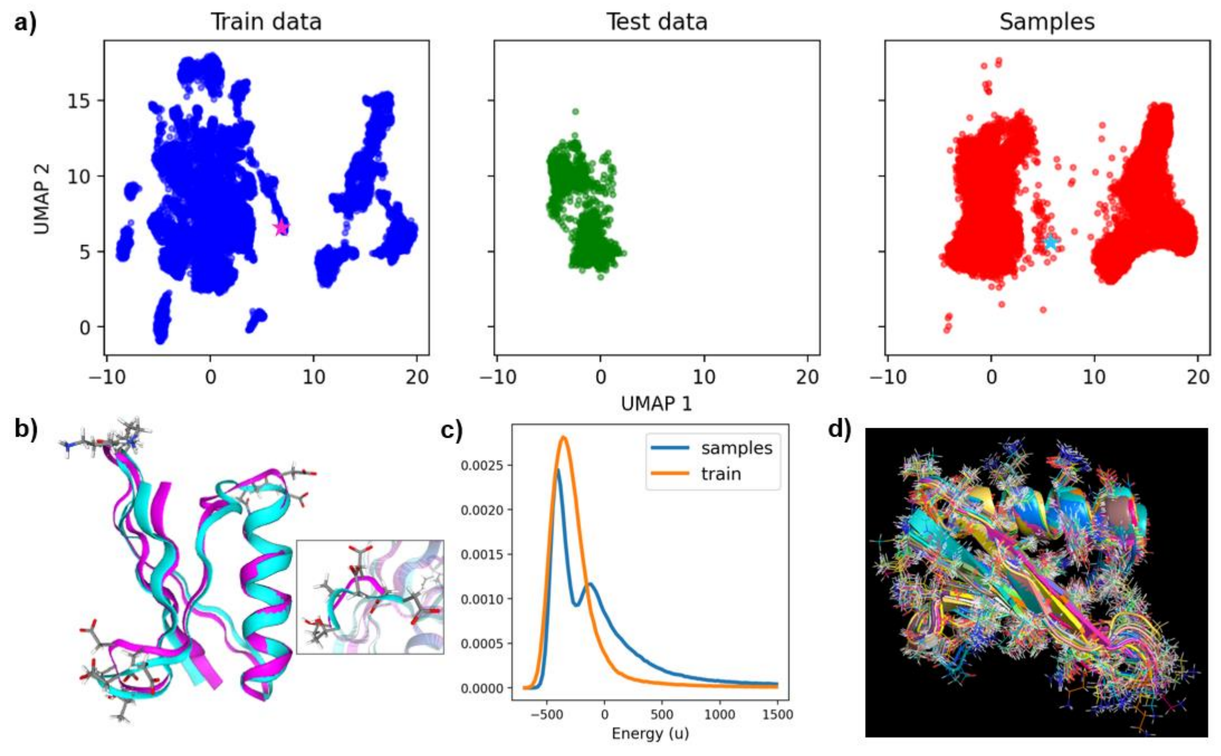

Figure 4: **BGs can generate novel sample conformations.** (a) Protein G 2D UMAP embeddings for the training data, test data, and 2×10^5 generated samples. (b) A representative example of generated structures by the BG model which was not found in training data (cyan) and the closest structure in the training dataset (magenta) by RMSD. Both structures are depicted as stars with their respective structural colors in (a). (c) Protein G energy distribution of training dataset (orange) and samples (blue) generated by our model. The second energy peak of the sampled conformations covers the novel structure shown in (b). (d) An overlay of high-resolution, lowest-energy all-atom structures of protein G generated by the BG model. This demonstrates that our model is capable of sampling low-energy conformations at atomic resolution.

References

[1] Noe et al. Boltzmann generators: sampling equilibrium states of many-body systems with deep learning. *Science*, 365(6457). 2019.
[2] Kohler et al. Smooth normalizing flows. *NeurIPS*. 2021.
[3] Midgley et al. Flow annealed importance sampling bootstrap. *ICLR*. 2023.
[4] Mahmoud et al. Accurate sampling of macromolecular conformations using adaptive deep learning and coarse-grained representation. *J. Chem. Inf. Model*, 62(7), 2022.
[5] Kohler et al. Flow-matching – efficient coarse-graining molecular dynamics without forces. *J. Chem. Theory Comput.*, 19(3), 2023.