Score-based generative models for binding peptide backbones

Matthew Greenig*[1], John Boom*[2], Pietro Sormanni[1], and Pietro Liò[3]

[1]Yusuf Hamied Department of Chemistry, University of Cambridge, Cambridge
[2]Department of Engineering, University of Cambridge, Cambridge
[3]Department of Computer Science, University of Cambridge, Cambridge

*Equal contribution

Introduction

- Score-based generative models (SGMs) are capable of generating diverse, novel protein backbone structures[1,2].
- A key application of SGMs in protein design is the **generation of protein backbones that bind a pre-specified target protein**.
  - ➢ In order to obtain a functional binder, amino acid sequences for these generated backbones are typically designed post-hoc using an inverse folding model[4].
- Thus far, little is known about the key model design choices that affect performance for this task. We attribute this to two factors:
  - ➢ The various existing SGMs for backbone design[1,2,3] differ so substantially in their architecture that it is **impossible to identify individual features that improve performance.**
  - ➢ **Appropriate metrics** for evaluating designed backbones for protein binding *in silico* **do not exist.**
- We present a generative model and evaluation pipeline – named **LoopGen** – that enables controlled comparison of model design choices in the context of binding protein backbone generation.

Methods

- Score-based generative models generate samples by learning to reverse a forward process that gradually adds noise to data.
- Protein structures are commonly represented as a sequence of frames – one rotation and one translation per residue – since each residue's internal structure is **rigid**:

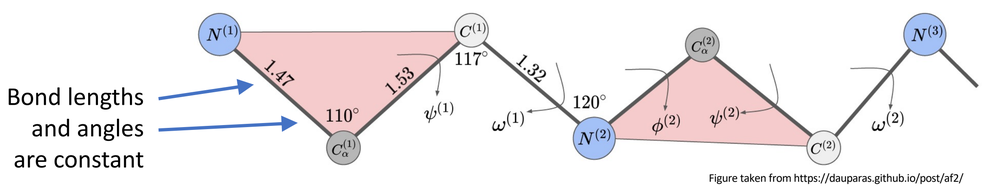

Figure taken from https://dauparas.github.io/post/af2/

- Yim et al.[3] established that the noising/denoising process in an SGM can be conducted separately over the rotational and translational component of each residue in this representation:

$$\mathbf{T}^{(t)} = (\mathbf{R}^{(t)}, \mathbf{X}^{(t)})$$

Each residue is associated with a 3-D rotation **R** and a 3-D translation **t**, indexed by time in the noising process

$$\mathrm{d}\mathbf{T}^{(t)} = [0, -\frac{1}{2}\mathrm{P}\mathbf{X}^{(t)}]\mathrm{d}t + \left[\mathrm{d}\mathbf{B}^{(t)}_{\mathrm{SO}(3)}, \mathrm{dP}\mathbf{B}^{(t)}_{\mathbb{R}^3}\right]$$

Forward (noising) process for both rotations and translations

$$\mathrm{d}\mathbf{R}^{(t)} = \nabla_{\mathbf{R}} \log p_t(\mathbf{T}^{(t)})\mathrm{d}t + \mathrm{d}\mathbf{B}^{(t)}_{\mathrm{SO}(3)}$$

Reverse (denoising) process for rotations

$$\mathrm{d}\mathbf{X}^{(t)} = \mathrm{P}\left(\frac{1}{2}\mathbf{X}^{(t)} + \nabla_{\mathbf{x}} \log p_t(\mathbf{T}^{(t)})\right)\mathrm{d}t + \mathrm{P}\mathrm{d}\mathbf{B}^{(t)}_{\mathbb{R}^3}$$

Reverse (denoising) process for translations

$$s_\theta(\mathbf{T}^{(t)}, t) \approx \left[\nabla \log p_t(\mathbf{R}^{(t)}), \nabla \log p_t(\mathbf{x}^{(t)})\right]$$

The network is trained to approximate the score functions at each time *t*, given the protein's current (noised) structure

- We implemented LoopGen using the above SGM with a GVP-GNN[5] as the score estimator and trained the model to generate binding peptide structures conditional on a target protein.
  - ➢ For training, we curated a dataset of antibody complementary determining region (CDR) loop structures with maximum 90% sequence similarity, each in complex with its target protein.
- Using LoopGen, we evaluated:
  - ➢ The effect of modelling entire residue frames (rotations + translations) compared to Cα atoms (translations only)
  - ➢ The effect of different variance schedules, specifically the interaction between the schedules for rotations and translations
  - ➢ The dependence of the model on the target epitope, using three novel tests: permutation, sequence scrambling, and translation of the epitope

Results

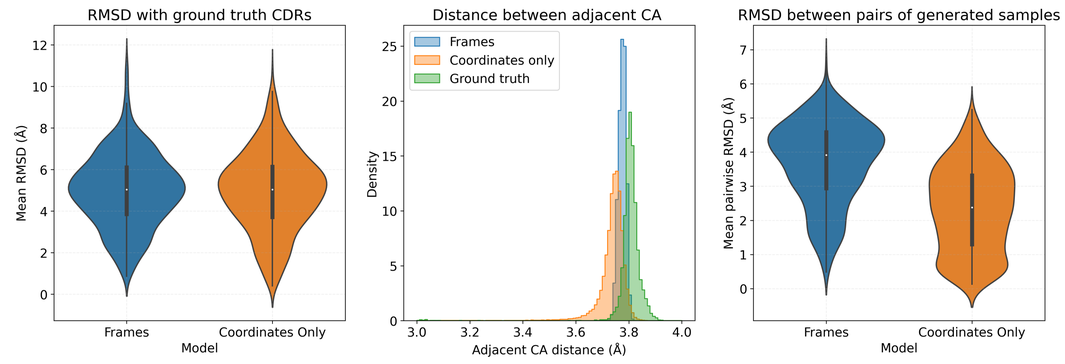

Figure 1. Comparison between a frame-based generative model (blue) and a Cα atom generative model (orange) for antibody CDR loop structures trained using LoopGen. Although the models are comparable in terms of the RMSD of their generated structures to each ground truth CDR (left), the frame generative model more effectively captures the true distribution of distances between adjacent Cα atoms (middle) and generates samples with much higher diversity, measured by Cα atom RMSD between samples (right).

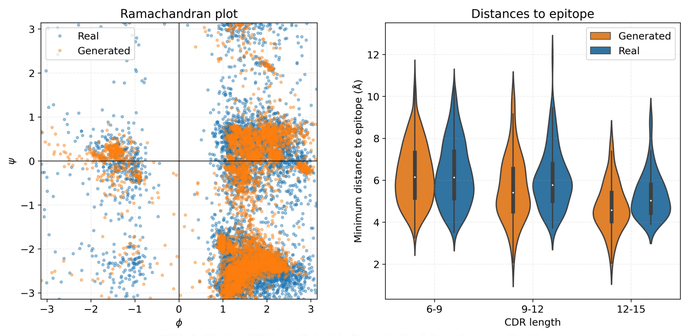

Table 1. Choice of Variance Schedule Dramatically Affects Structure Quality

Trans. Sched.	Rot. Sched.	RMSD (Å)	Internal Clashes (%)	Bond Length (%)	Bond Angle (%)	Epi.-CDR Clash (%)	Any Struct. Viol. (%)
Lin.	Log.	4.98 ± 2.14	0.3	20.6	3.9	3.3	22.4
Quad.	Log.	4.93 ± 2.15	0.0	6.3	0.6	2.7	8.7
Log.	Log.	4.85 ± 2.08	88.9	96.2	43.9	5.7	97.1
Sig.	Log.	4.93 ± 2.18	0.6	6.1	1.8	3.4	9.2
Lin.	Lin.	5.04 ± 2.08	1.0	29.4	4.0	3.4	30.9
Lin.	Quad.	5.13 ± 2.17	0.8	18.6	1.7	3.5	20.5

Figure 2. Evaluation of LoopGen-generated loops in terms of physicochemical plausibility. Ramachandran plot (top left) showing the backbone dihedral angle distribution for generated loops (orange) aligns with the distribution for real CDR loops from the test set (blue). Likewise, the minimum distance between Cα atoms in the CDR loop and the epitope is similar between real and generated backbones (top right). Finally, we conduct an analysis of variance schedule combinations, showing that, while the differences between variance schedules are not obvious when using ground-truth RMSD as a metric, they become very clear when examining the physicochemical violation rates in generated loops (bottom right). Notably, the best-performing variance schedule corresponds to a sequential denoising, where the Cα atom positions (translations) are denoised before the residue orientations (rotations).

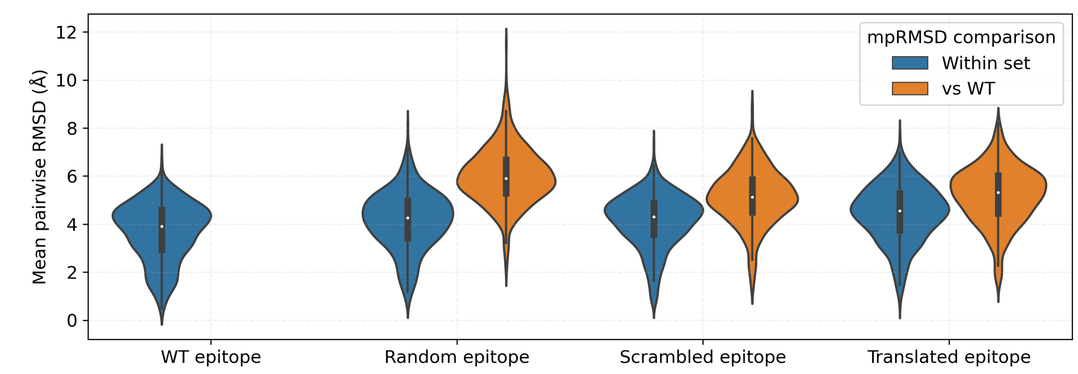

Figure 3. Mean pairwise RMSD between samples generated under various transformations of each test set epitope and the WT epitope. For each epitope in the test set, we perform the following perturbations: permutation/alignment with another random epitope in the test set ("Random epitope"), permutation of sequence identities within the WT epitope structure ("Scrambled epitope"), and translation by 20Å in the direction opposite to the CDR centroid ("Translated epitope"). For each of these transformations of the epitope, as well as the WT epitope, we use LoopGen to sample 10 CDR backbones. For these 10 CDR backbones in each epitope condition, we then plot the mean pairwise RMSD (mpRMSD) *within* the set generated for that epitope condition (blue) and the mpRMSD *between* the generated backbones for that condition and the CDR backbones generated for the WT epitope (orange).

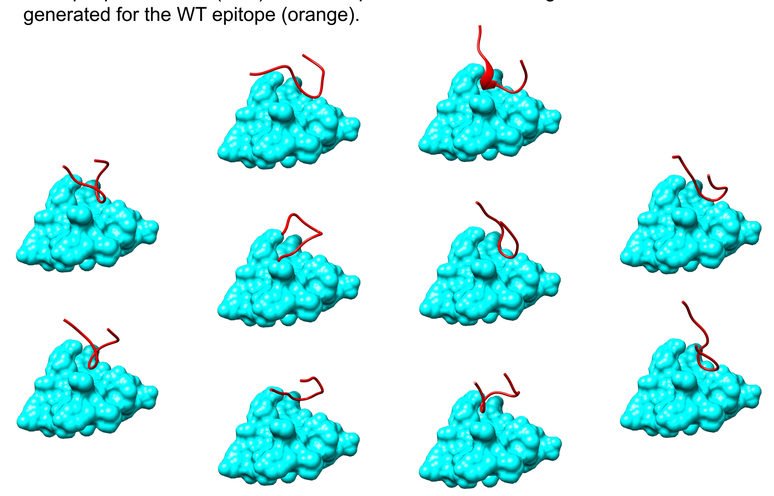

Figure 4. Example generated loop structures for a test set epitope (PDB ID: 3ULU). Notably, all loops are oriented correctly with respect to the target, with room to accommodate an antibody scaffold.

Future directions

- CDR-specific inverse folding model for sequence design.
- Exact likelihood computation for ranking designs – either using probability flow ODE or flow matching framework.
- Grafting – how to identify the optimal antibody framework for a designed loop?

1. Ingraham, J.B., Baranov, M., Costello, Z. *et al.* Illuminating protein space with a programmable generative model. *Nature* **623**, 1070–1078 (2023). https://doi.org/10.1038/s41586-023-06728-8
2. Watson, J.L., Juergens, D., Bennett, N.R. *et al.* De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089–1100 (2023). https://doi.org/10.1038/s41586-023-06415-8
3. Yim, J., Trippe, B. L., De Bortoli, V., Mathieu, E., Doucet, A., Barzilay, R., & Jaakkola, T. SE(3) diffusion model with application to protein backbone generation. arXiv (2023). https://doi.org/10.48550/arxiv.2302.02277
4. J. Dauparas et al., Robust deep learning–based protein sequence design using ProteinMPNN. Science 378, 49-56 (2022). https://doi.org/10.1126/science.add2187
5. Jing, B., Eismann, S., Soni, P. N., & Dror, R. O. Equivariant Graph Neural Networks for 3D Macromolecular Structure (Version 2). arXiv (2021). https://doi.org/10.48550/ARXIV.2106.03843

Read more!