Generating All-Atom Protein Structure from Sequence-Only Training Data

Amy X. Lu[1,2], Wilson Yan[1], Sarah A. Robinson[2], Kevin K. Yang[3], Vladimir Gligorijevic[2], Kyunghyun Cho[2,4], Richard Bonneau[2], Pieter Abbeel[1], Nathan Frey[2]

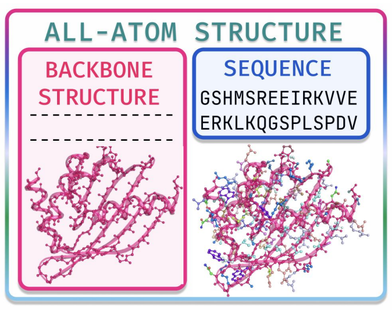
[1]UC Berkeley [2]Prescient Design, Genentech [3]Microsoft Research [4]New York University

Prescient Design
A Genentech Accelerator

tl;dr: by training a diffusion model in the latent space of ESMFold, we generate diverse & high quality all-atom proteins!

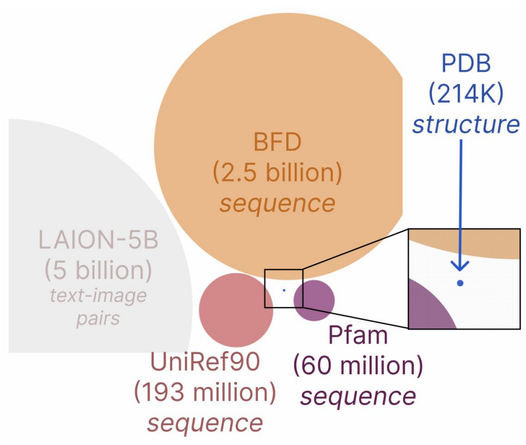
Problem: All-Atom Protein Generation

- Sidechains are crucial in mediating function, but often ignored in popular structure generation methods [1,2]
- Generating the **all-atom structure** is a **multimodal generation problem** requiring simultaneous generation of sequence and structure.

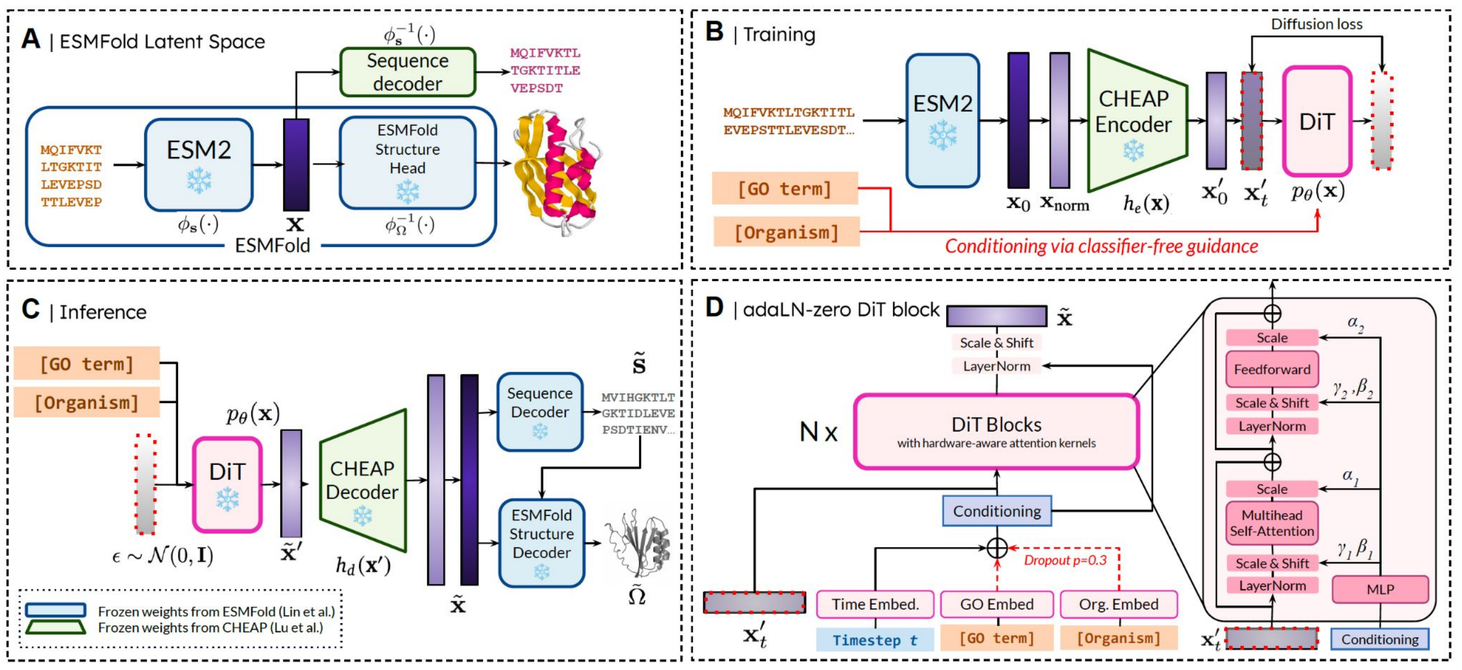


Motivations

1. Given the **scarcity and biases of structural data**, how can we use sequence data to get better coverage of protein evolution?
2. Can we go beyond structure and use **functional/taxonomic information for conditioning**?
3. How can we **leverage the information stored in weights of pretrained protein folding models** for generation?
4. Can we **avoid separate structure-to-sequence and sequence-to-structure steps** for all-atom generation?



PLAID (Protein Latent Induced Diffusion)



(A) Overview of ESMFold [3], which predicts structure from sequence. We use the latent representation just before the structure module for generation.
(B) During training, since we only need sequence to obtain the representation, we can train on sequence databases. We train a denoising diffusion model [4] in the compressed CHEAP latent space [5], following works for high-resolution image generation [6].
(C) During inference, we generate the latent embedding, and use frozen decoders to obtain sequence and all-atom structure.
(D) Function and taxonomic conditioning is added via classifier-free guidance [7]. We use the DiT [8] architecture for incorporating conditioning information; since the architecture only uses attention and linear layers, we can make use of hardware-aware fused attention kernels for faster training/inference. We train two models with 2B and 100M parameters each.

Results: Unconditional Generation

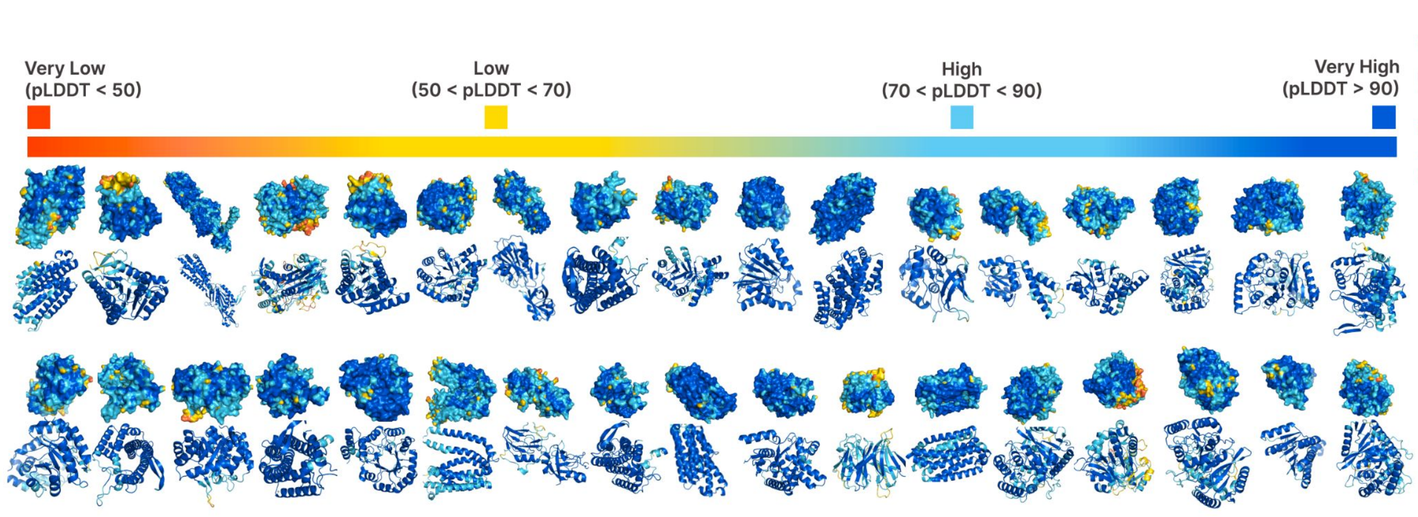


Fig 1 | Despite not requiring structures to train the diffusion model, PLAID can generate high-quality and diverse all-atom structures.

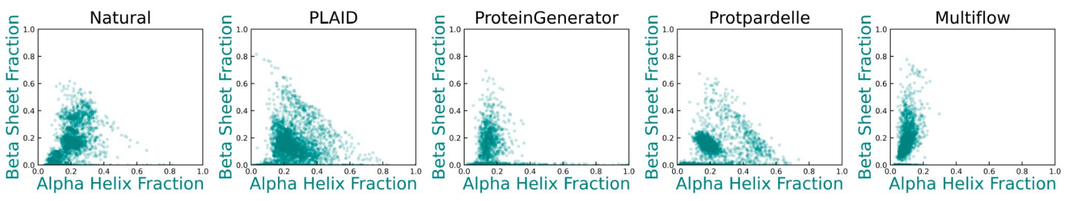


Fig 2 | PLAID samples better balance β-sheet and α-helix content compared to previous methods.

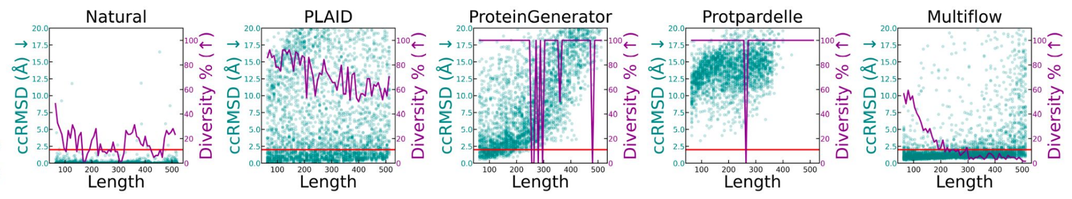


Fig 3 | Scatterplot of sample quality, overlaid with lineplot of sample diversity, examined by length. PLAID better balances diversity and quality, especially for longer sequences.

Results: Conditional Generation

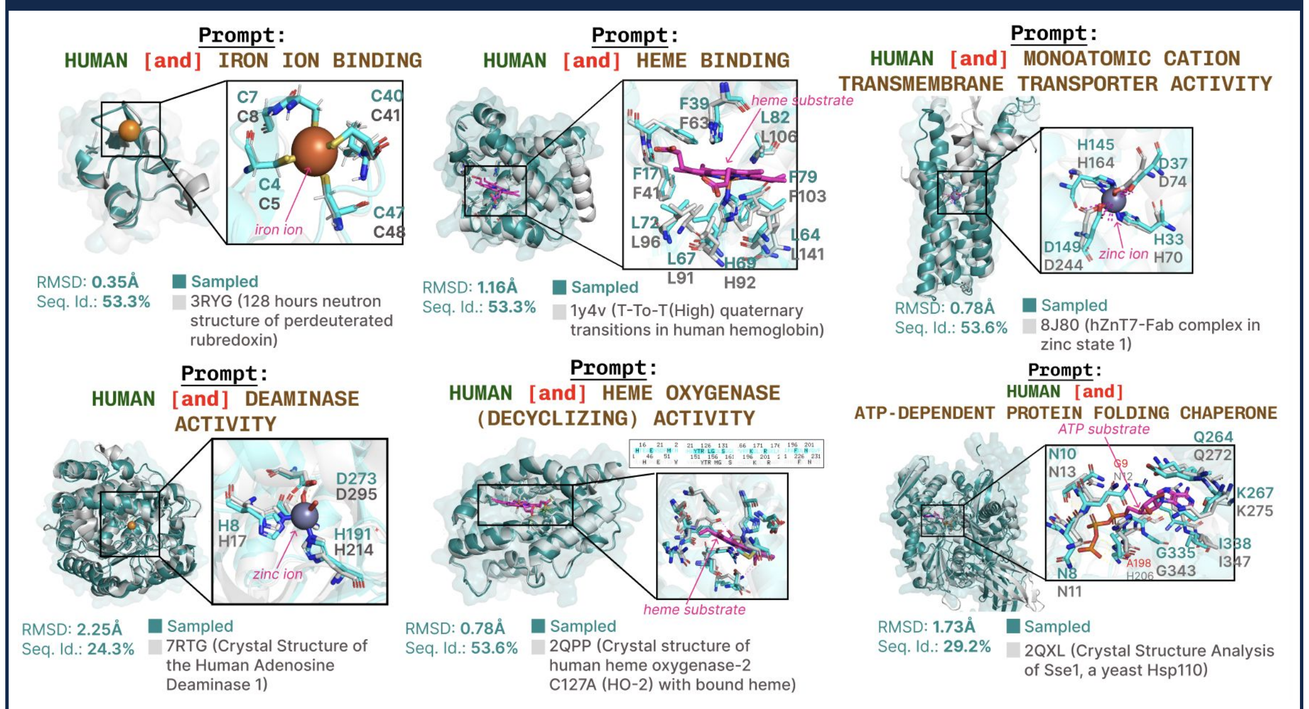


Fig 4 | Function-conditioned samples capture both active-site sequence motifs and correct sidechain placements. Global sequence similarity to closest known sequence is low, suggesting learning rather than memorizing training data.

References

[1] Watson et al. De novo design of protein structure and function with RFdiffusion. *Nature*, 2023.
[2] Ingraham et al. Illuminating protein space with a programmable generative model. *Nature*, 2023.
[3] Lin et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 2023.
[4] Ho et al. Denoising Diffusion Probabilistic Models. NeurIPS, 2020.
[5] Lu et al. Tokenized and Continuous Embedding Compressions of Protein Sequence and Structure. *bioRxiv,* 2024.
[6] Rombach et al. High-Resolution Image Synthesis with Latent Diffusion Models. CVPR, 2022.
[7] Ho et al. Classifier-Free Diffusion Guidance. arXiv, 2022.
[8] Peebles et al. Scalable Diffusion Models with Transformers. ICCV, 2023.

Code **PLAID preprint** **CHEAP preprint** **Weights**

amyxlu@berkeley.edu
@amyxlu
amyxlu.github.io

If you are interested in wet-lab verification and/or using PLAID for functional generation, please reach out!