Ambient Proteins: Training Diffusion Models on Low Quality Structures

Giannis Daras*, Jeffrey Ouyang-Zhang*, Krithika Ravishankar, William Daspit, Costis Daskalakis, Qiang Liu, Adam Klivans, Daniel J. Diaz

Massachusetts Institute of Technology

The University of Texas at Austin

Problem

- State-of-the-art generation train on synthetic data from AlphaFold
- Low confidence structures are typically discarded
- Goal: Train on all available data

Approach

Key Insight: distributions contract with noise

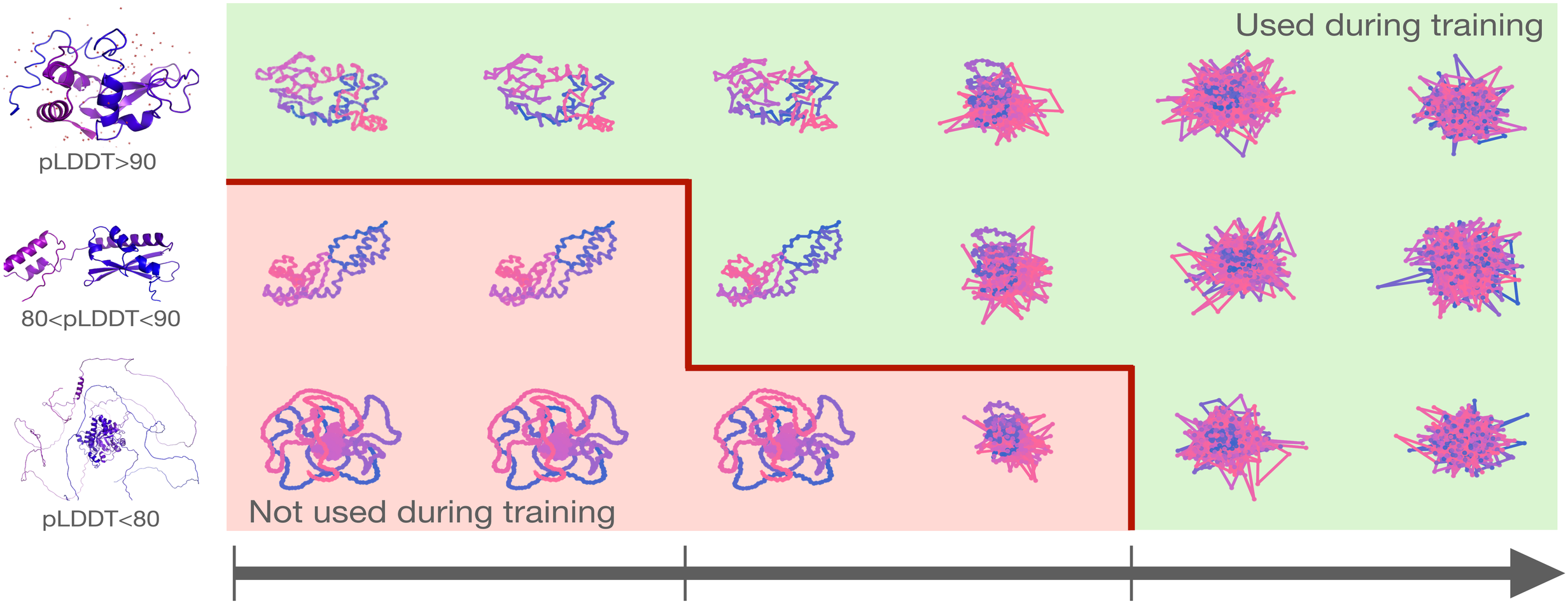
$$\mathrm{D_{KL}}(p_t||\tilde{p}_t) \leq \mathrm{D_{KL}}(p_{t'}\,||\,\tilde{p}_{t'}), \quad \forall t \geq t'$$

p_t : data distribution at noise t

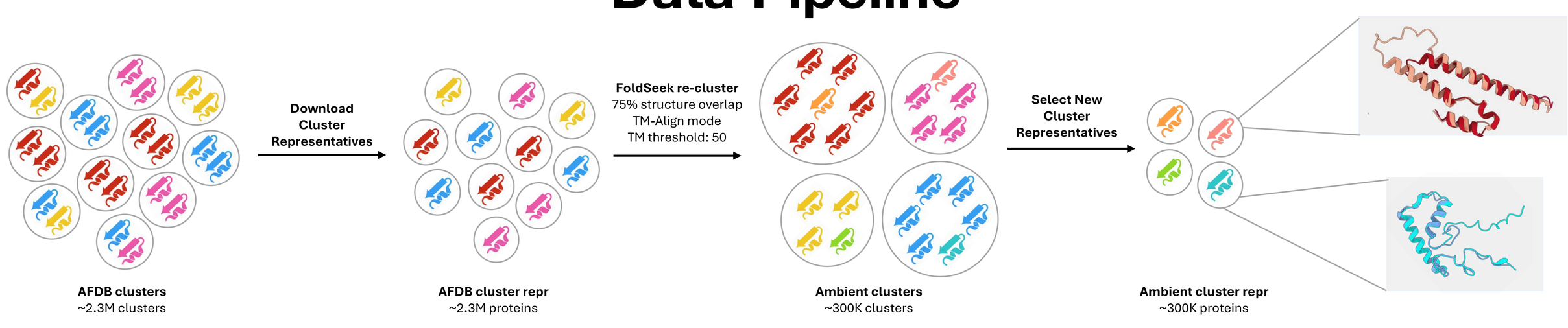
$\tilde{p}_t$: synthetic distribution at noise t

Method

Ambient Diffusion: Train with low-quality proteins at sufficiently high noise levels

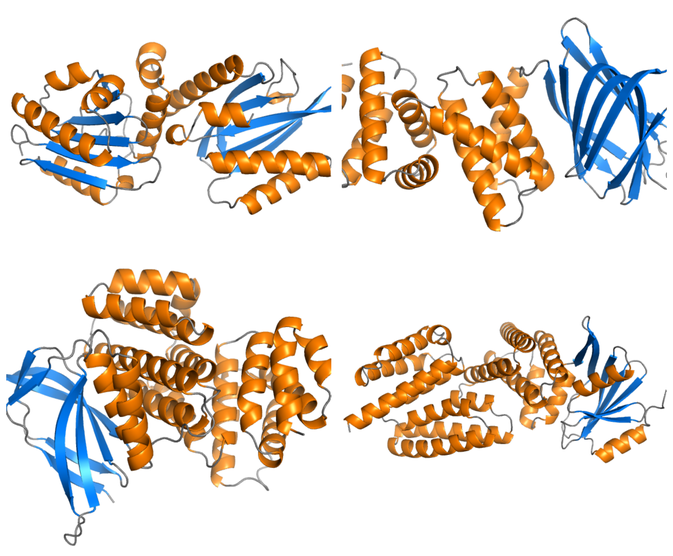


Data Pipeline

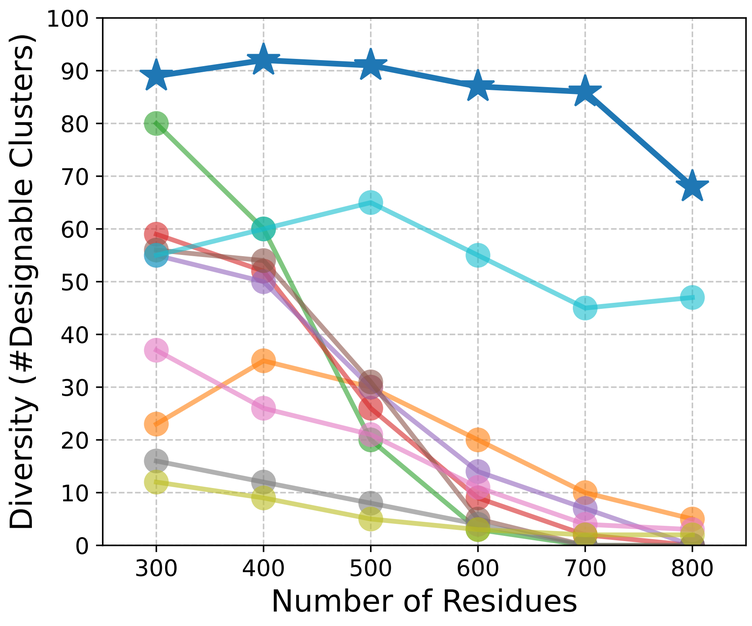


Results

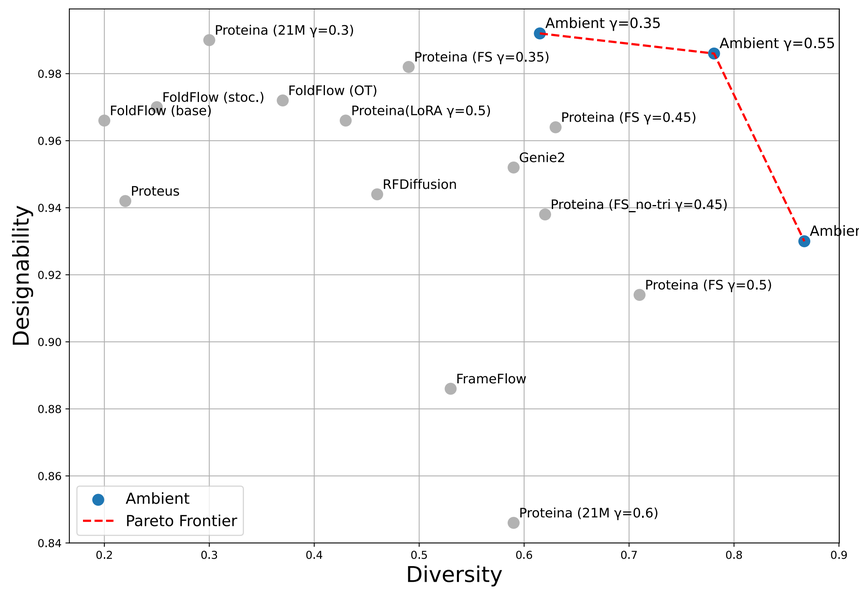
Qualitative Visualizations



State-of-the-art Generation



Pareto Frontier



[1] Highly accurate protein structure prediction with AlphaFold. Jumper et al. 2021.
[2] Out of many, one: Designing and scaffolding proteins at the scale of the structural universe with genie2. Lin et al. 2024.