
Ambient Proteins

Training Diffusion Models on Noisy Structures

Supplement

Anonymous Author(s)

Affiliation

Address

email

1 A Evaluation Metrics

2 Evaluation of a protein generative model is challenging and there have been a few metrics that have
3 been proposed. In what follows, we explain standard metrics in the protein-generative modeling
4 literature that we will use in our Experimental Results section. Our experiments report using Proteína's
5 definitions of the metrics when possible.

6 **Designability** (also referred to as refoldability) assesses the structural plausibility of generated
7 proteins. Given a generated backbone, ProteinMPNN [1] generates eight plausible amino acid
8 sequences for that backbone. ESMFold then folds each sequence and the resulting eight structures
9 are compared to the original backbone. The self-consistency RMSD (scRMSD) is defined as the
10 smallest root mean squared deviation between the generated backbone and each of the eight refolded
11 structures. A backbone is considered *designable* if $\text{scRMSD} < 2 \text{ \AA}$ and designability is defined as
12 the percentage of generated backbones that meet this criterion.

Diversity quantifies the structural variability among the generated proteins. Designable backbones
are clustered using Foldseek with a TM-score threshold of 0.5. Diversity is then defined as:

$$\text{Diversity} = \frac{\text{Number of Designable Clusters}}{\text{Number of Designable Samples}}.$$

13 This metric reflects the proportion of structurally distinct (i.e., non-redundant) designable backbones
14 among all designable samples.

15 B Secondary Structure conditioning

16 Previous work [2] has explored conditioning protein structure generation on CATH labels, a form of
17 hierarchical classification derived from the orientation and spatial organization of protein secondary
18 structures [5]. In this setting, every residue in a protein sequence is typically assigned the same CATH
19 label. In contrast, we propose a more fine-grained approach. Rather than relying on the manually
20 curated and coarse-grained CATH classification, we condition our model directly on secondary
21 structure annotations at the residue level. Each residue is assigned a label corresponding to its local
22 secondary structure (e.g., helix, strand, coil), allowing the model to leverage localized structural
23 context during generation.

24 We train a variant of our model with partial conditioning, in which the model is conditioned on
25 the secondary structure sequence, without introducing any additional modifications to the input or
26 architecture. We show designable samples conditioned on the secondary structure extracted from real

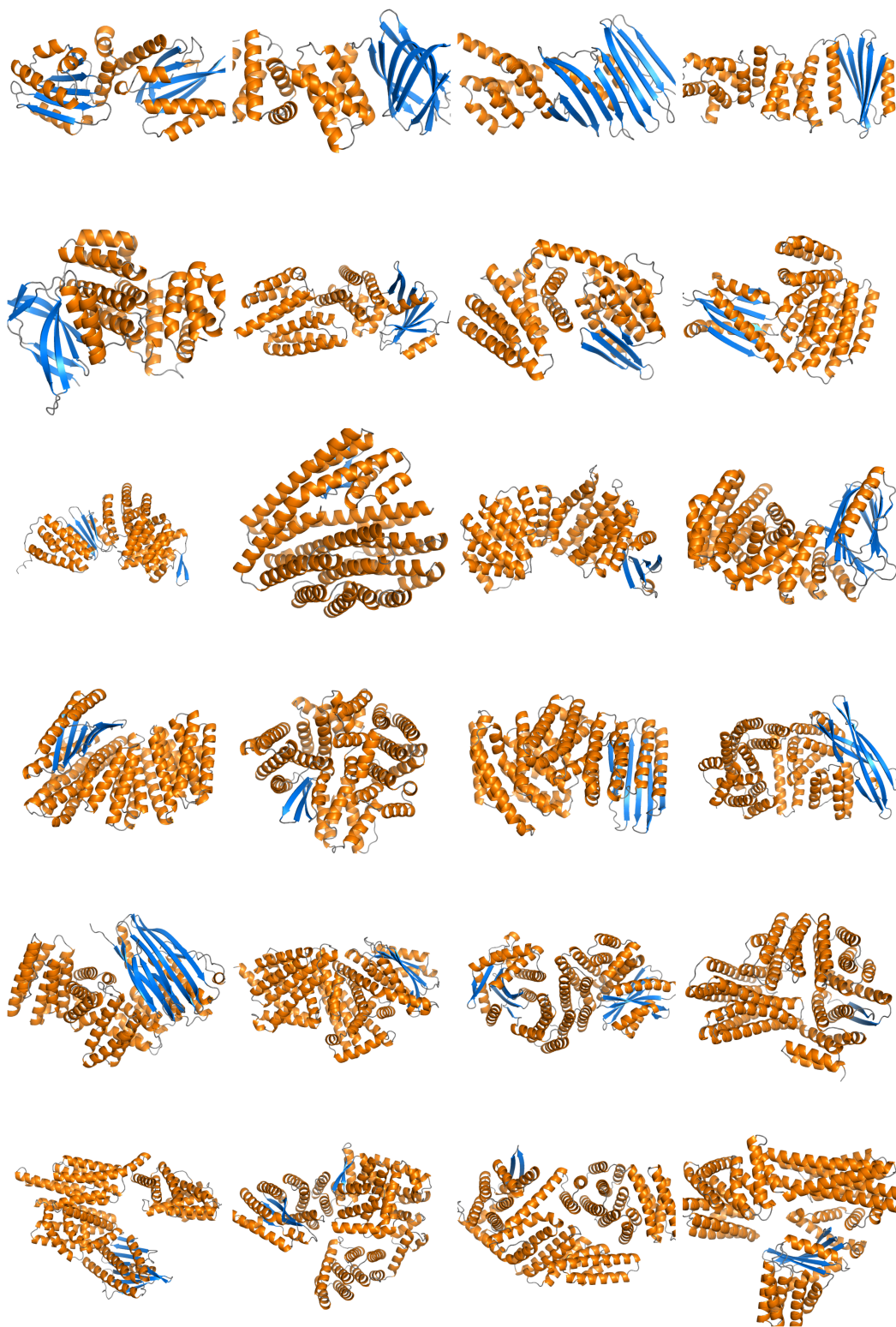


Figure 1: **Qualitative visualizations of unconditional generations.** Our model is capable of producing diverse, multi-domain long proteins.

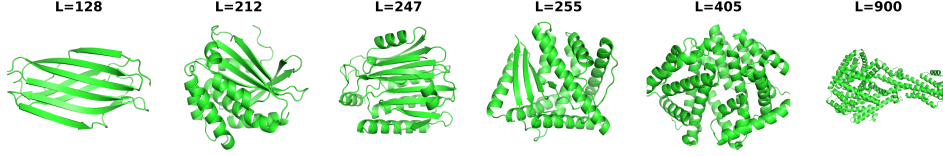


Figure 2: **Secondary Structure Conditioned Samples.** We generate proteins using a model variant trained with secondary structure conditioning. To guide generation, we extract secondary structure strings from existing proteins and use this coarse-grained structural representation as input. This conditioning enables the model to produce diverse and designable protein structures.

proteins in Figure 2. These results demonstrate that, even with coarse-grained secondary structure conditioning, our model can generate long, diverse proteins exhibiting a wide range of folds.

C Model and Training Hyperparameters

Table 1 includes a more thorough list of the hyperparameters used for our experiments.

D Full Motif Scaffolding Results

Table 2 and table 3 represent the numbers of unique successful scaffolds generated by Genie2 [3], RFDiffusion [6], Proteína [2] and *Ambient Protein Diffusion* for each motif in the benchmark dataset in Genie2.

E Full Training Algorithm and Implementation Details

E.1 Additional Implementation Details

Loss buffer. The loss rescaling introduced in the main paper ensures balanced weighting across noise levels. At the same time, it also introduces a potential instability: the loss explodes as $\sigma(t)$ approaches $\sigma(t_i)$. To mitigate this instability, we define a buffer zone around each protein’s assigned noise level. Specifically, given a protein’s assigned noise level t_i , it is only used during training at timesteps $t + \tau$, where τ is a buffer hyperparameter that controls the exclusion margin. This constraint prevents the model from encountering degenerate gradient behavior near the rescaling boundaries and is only applied to medium and low confidence structures (pLDDT < 90). We underline that is similar to how in normal diffusion there is a buffer time zone around $t = 0$ that is never sampled.

Ambient in high-noise regime. As explained in the main paper, each protein is only used for a subset of diffusion times according to its average pLDDT value. The proteins that have super high PLDDT (> 90) are considered clean data and can be used with the normal training objective. However, as found in [4], using the Ambient training objective for high-noise might be useful even if clean data is available. Intuitively, this objective prevents memorization and promotes diversity in the outputs. We ablated this design choice, and we found a slight increase in diversity for the same designability by using this. Hence, we used this tool from [4] for all our Ambient Protein Diffusion trainings.

E.2 Algorithm

We provide the full algorithm in Algorithm 1. We commit to open-sourcing our code and models to facilitate the broader adoption of our method from the community.

Hyperparameter	Genie2 (repro)	<i>Ambient</i> (Stage 1)	Stage 2	Stage 3
<i>Diffusion</i>				
Number of timesteps	1,000	-	-	-
Noise schedule	Cosine	-	-	-
Ambient walls	-	(600,900)	(600,900)	(600,900)
<i>Model Architecture</i>				
Single feature dimension	384	-	-	-
Pair feature dimension	128	-	-	-
Pair transform layers	5	8	8	8
Pair drop path rate	0.0	0.1	0.1	0.1
Triangle dropout	0.25	-	-	-
Structure layers	8	-	-	-
<i>Training</i>				
Optimizer	AdamW	-	-	-
Number of training proteins	586k	196k	269k	291k
Number epochs	40	200	50	20
Warmup iterations	10,000	1,000	500	100
Total batch size	384	384	96	48
Learning rate	1.0×10^{-4}	1.0×10^{-4}	1.0×10^{-5}	1.0×10^{-5}
Weight decay	0.05	-	-	-
Minimum protein length	20	20	50	50
Maximum protein length	256	256	512	768
Minimum mean pLDDT	80	70	70	70
<i>Compute Resources</i>				
Number of GPUs	48	48	48	48
Training time	18 hr	18hr	48hr	48hr

Table 1: **Hyperparameters of the diffusion protein model.** Dashes (-) indicate that the value is the same as the previous column. The Ambient walls correspond to the assigned diffusion times based on the protein’s PLDDT (times are from 1 to 1000). Proteins with PLDDT > 90 are used everywhere. Proteins with PLDDT > 80 are used for times in [600, 1000] and proteins with PLDDT > 70 are used for times in [900, 1000]. We underline that these hyperparameters were not particularly optimized, and even more benefits might be observed by properly tuning these values.

Motif Name	Genie 2	RFDiffusion	Proteína	<i>Ambient Protein Diffusion</i>
6E6R_long	415	381	713	601
6EXZ_long	326	167	290	432
6E6R_med	272	151	417	406
1YCR	134	7	249	146
5TRV_long	97	23	179	119
6EXZ_med	54	25	43	69
7MRX_128	27	66	51	44
6E6R_short	26	23	56	27
5TRV_med	23	10	22	23
7MRX_85	23	13	31	17
3IXT	14	3	8	4
5TPN	8	5	4	11
7MRX_60	5	1	2	1
1QJG	5	1	3	5
5TRV_short	3	1	1	3
5YUI	3	1	5	4
4ZYP	3	6	11	3
6EXZ_short	2	1	3	3
1PRW	1	1	1	1
5IUS	1	1	1	1
1BCF	1	1	1	1
5WN9	1	0	2	1
2KL8	1	1	1	1
4JHW	0	0	0	0
Total	1445	889	2094	1923

Table 2: **Detailed single motif scaffolding results.** Ambient Protein Diffusion achieves superior results to Genie 2 and RFDiffusion and performs on par with Proteína. Crucially, our model achieves these results zero-shot, i.e., unlike Proteína, it is not optimized for motif scaffolding and still achieves comparable performance while being an order of magnitude smaller.

Motif Name	Genie 2	<i>Ambient Protein Diffusion</i>
3BIK+3BP5	17	23
1PRW_four	11	38
1PRW_two	8	15
4JHW+5WN9	4	12
2B5I	0	1
3NTN	0	0
Total	40	89

Table 3: **Multi-motif scaffolding results.** Ambient Protein Diffusion achieves consistently superior results to the predecessor Genie-2 model, despite using the same architecture, i.e. the benefit comes from better use of the data. The motif 2B5I is only solved by Ambient Protein Diffusion.

Algorithm 1 Ambient Protein Diffusion: Training Algorithm.

Require: untrained network h_θ , dataset $\mathcal{D} = \{(x_0^{(i)}, \text{pLDDT}^{(i)})\}_{i=1}^N$, pLDDT to diffusion time mapping function $f : [0, 100] \mapsto \mathbb{R}^+$, noise scheduling $\sigma(t)$, batch size B , diffusion time T , buffer τ .

- 1: $\tilde{\mathcal{D}} \leftarrow \left\{ \left(x_0^{(i)} + f(\text{pLDDT}^{(i)})\epsilon^{(i)}, f(\text{pLDDT}^{(i)}) \right) \mid (x_0^{(i)}, \text{pLDDT}^{(i)}) \in \mathcal{D}, \epsilon^{(i)} \sim \mathcal{N}(0, I_d) \right\} \triangleright$
Noise each point in the training set according to its pLDDT and get (noisy, noise level) pairs.
- 2: **while** not converged **do**
- 3: $t_s^{(1)}, \dots, t_s^{(B)} \leftarrow \text{Sample uniformly } B \text{ times in } [0, T] \triangleright \text{Sample diffusion times for this batch.}$
- 4: $\tilde{\mathcal{D}}_p \leftarrow \text{shuffle}(\tilde{\mathcal{D}}) \triangleright \text{Shuffle dataset.}$
- 5: loss $\leftarrow 0 \triangleright \text{Initialize loss.}$
- 6: pos $\leftarrow 0 \triangleright \text{Initialize index at shuffled dataset.}$
- 7: **for** $i \in [1, B]$ **do**
- 8: **while** True **do** $\triangleright \text{find the first eligible point}$
- 9: $y, t_y \leftarrow \tilde{\mathcal{D}}_p[\text{pos}]$
- 10: **if** $t_y \geq t_s^i + \tau$ **then**
- 11: break
- 12: **else**
- 13: pos $\leftarrow \text{pos} + 1 \triangleright \text{Move to the next point in the dataset.}$
- 14: **end if**
- 15: **end while**
- 16: $\epsilon \sim \mathcal{N}(0, I) \triangleright \text{Sample noise.}$
- 17: $t \leftarrow t_s^{(i)} \triangleright \text{Time to be used in this training update.}$
- 18: $t_i \leftarrow t_y \triangleright \text{Assigned time based on the PLDDT value}$
- 19: $x_{t_i} \leftarrow y \triangleright \text{Noised point to the assigned time.}$
- 20: $x_t \leftarrow x_{t_i} + \sqrt{\sigma^2(t) - \sigma^2(t_i)}\epsilon \triangleright \text{Add additional noise.}$
- 21: $\alpha(t, t_i) \leftarrow \frac{\sigma^2(t) - \sigma^2(t_i)}{\sigma^2(t)}$
- 22: $w(t, t_i) \leftarrow \frac{\sigma^4(t)}{(\sigma^2(t) - \sigma^2(t_i))^2} \triangleright \text{Loss reweighting.}$
- 23: loss $\leftarrow \text{loss} + w(t, t_i) \|\alpha(t, t_i)h_\theta(x_t, t) + (1 - \alpha(t, t_i))x_t - x_{t_i}\|^2 \triangleright \text{Ambient loss}$
- 24: **end for**
- 25: loss $\leftarrow \frac{\text{loss}}{B} \triangleright \text{Compute average loss.}$
- 26: $\theta \leftarrow \theta - \eta \nabla_\theta \text{loss} \triangleright \text{Update network parameters via backpropagation.}$
- 27: **end while**

References

- 55 [1] Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J Ragotte, Lukas F
56 Milles, Basile IM Wicky, Alexis Courbet, Rob J de Haas, Neville Bethel, et al. Robust deep
57 learning-based protein sequence design using proteinmpnn. *Science*, 378(6615):49–56, 2022.
- 58 [2] Tomas Geffner, Kieran Didi, Zuobai Zhang, Danny Reidenbach, Zhonglin Cao, Jason Yim, Mario
59 Geiger, Christian Dallago, Emine Kucukbenli, Arash Vahdat, et al. Proteina: Scaling flow-based
60 protein structure generative models. *arXiv preprint arXiv:2503.00710*, 2025.
- 61 [3] Yeqing Lin, Minji Lee, Zhao Zhang, and Mohammed AlQuraishi. Out of many, one: Designing
62 and scaffolding proteins at the scale of the structural universe with genie 2. *arXiv preprint*
63 *arXiv:2405.15489*, 2024.
- 64 [4] Kulin Shah, Alkis Kalavasis, Adam R. Klivans, and Giannis Daras. Does generation require
65 memorization? creative diffusion models using ambient diffusion, 2025.
- 66 [5] Ian Sillitoe, Nicola Bordin, Natalie Dawson, Vaishali P Waman, Paul Ashford, Harry M Scholes,
67 Camilla SM Pang, Laurel Woodridge, Clemens Rauer, Neeladri Sen, et al. Cath: increased
68 structural coverage of functional space. *Nucleic acids research*, 49(D1):D266–D273, 2021.
- 69 [6] Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E
70 Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De novo
71 design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.
- 72