Protein Generator with Ribosomal Origin and Folding

MinGyu Choi [1] **Derek Chen** [2] **Tommi Jaakkola** [2] **Regina Barzilay** [2]

Abstract

Current protein generative models often prioritize computational efficiency over fidelity to biological mechanisms, leading to artifacts such as mode collapse into helical structures that are difficult to diagnose and correct. We hypothesize that generative processes more closely aligned with authentic biological pathways can produce more diverse and unbiased outputs. To this end, we propose a generative model that combines internal coordinate parameterization with a novel trans-dimensional diffusion process inspired by ribosomal protein synthesis and co-translational folding. The model incrementally elongates the polypeptide chain while allowing nascent residues to fold, enabling early segments to explore diverse substructures and later segments to condition on partially folded contexts. In addition, our model supports length-independent flexible generation, allowing protein size to emerge dynamically during sampling and removing the inherent bias introduced by prespecified lengths. Empirically, our approach achieves superior in-distribution coverage and secondary structure balance without finetuning compared to state-of-the-art baselines.

1. Introduction

Proteins exert their biological functions through complex three-dimensional conformations, rendering structural characterization fundamental to both mechanistic understanding and rational molecular design. Early *in silico* structure prediction efforts approached this challenge through methodologies grounded in biochemical intuition and experimental observations, including fragment-assembly techniques (Simons et al., 1997; Xu & Zhang, 2012), internal-coordinate representations (Coutsias et al., 2004; Kolodny et al., 2005), and physics-informed potential functions (Rohl et al., 2004).

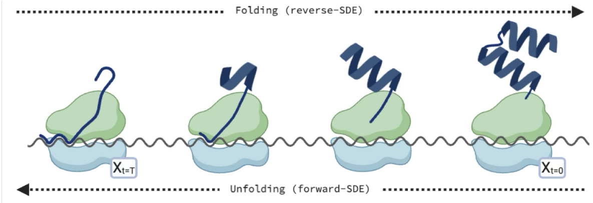


Figure 1. Illustration of biological protein synthesis via ribosome and co-translation folding. `RiboFold` mimics this mechanism with trans-dimensional autoregressive generative process.

The advent of AlphaFold2 (Jumper et al., 2021) marked a pivotal shift in the field, introducing a neural architecture capable of learning spatial constraints directly from multiple sequence alignments and structural databases. Its unprecedented accuracy catalyzed a new generation of models that cast aside biology-based components in favor of large-scale representation learning.

Building upon this paradigm, recent advances in generative modeling—particularly score-based (Ho et al., 2020) and flow-based frameworks (Lipman et al., 2023)—have enabled the direct generation of protein structures *de novo*. These include a growing body of work leveraging denoising diffusion models (Yim et al., 2023; Watson et al., 2023; Ingraham et al., 2023; Lin & AlQuraishi, 2023; Wang et al., 2024; Wu et al., 2024) and their flow-matching extensions (Bose et al., 2024; Campbell et al., 2024; Geffner et al., 2025), which aim to model the high-dimensional manifold of plausible protein conformations.

Despite significant advancements, current generative frameworks remain constrained in their ability to elucidate the fundamental mechanisms underlying protein folding. To optimize computational efficiency, many models abstract chemical continuity by treating atoms or residues as independent entities, breaking covalent bonds. This simplification results in chemically invalid intermediate structures that fail to capture the cooperative and sequential nature inherent in biological protein *folding* processes. This absence of valid chemical intermediates also obscures the identification and correction of structural biases, such as the overrepresentation of compact helical topologies (Lu et al., 2025). Given the critical importance of loop-rich structures for enzymatic activity and antibody design, addressing this bias is a pressing challenge.

[1] Department of Chemistry, [2] Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology. Correspondence to: MinGyu Choi <chemgyu@mit.edu>.

Proceedings of the Workshop on Generative AI for Biology at the 42nd International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

To address these challenges, we begin with the hypothesis that generative processes more closely aligned with biological mechanisms yield more accurate and less biased structural outputs. We first revisit internal-coordinate parameterizations and incorporate them into a generative modeling framework. In addition, we propose a trans-dimensional, soft autoregressive generative process that mimics biological protein synthesis and co-translational folding in cells.

Here, we introduce the **Protein Generator with Ribosomal Origin and Folding (RiboFold)**, a model that bridges biological intuitions with computational advancements. Our main contributions are:

1. **Genuine *Folding* Process**: Extending prior work (Wu et al., 2024), we introduce a more efficient parameterization and an optimized noise schedule to enhance generation stability and fidelity (Section 3.1). We also propose a novel autoregressive diffusion process and its trans-dimensional extension that emulates *in vivo* co-translational folding (Section 3.2). Our generative process exhibits a characteristic of *clustered folding*, wherein local structures form first and subsequently assemble into global architectures (Section 3.3).

2. **Efficient All-Atom Generation**: We propose a branched, trans-dimensional generative process operating over the side-chain dimension (Section 3.3). This incrementally grows each side chain while exposing only physically interpretable features, significantly reducing the degrees of freedom compared to previous all-atom approaches that inflate the input space with placeholder atoms (Chu et al., 2024; Qu et al., 2024).

3. **Unbiased Controllable Structural Diversity**: Leveraging internal coordinate representation, our model provides explicit control over secondary structure formation, effectively mitigating the common helical bias seen in prior works (Section 4.3). This also facilitates conditional generation based on desired secondary structure content (Section 4.4).

4. **Flexible Length Generation**: Our framework supports open-ended generation with dynamically inferred protein length. The trans-dimensional generation removes length-dependent biases (Section 4.3), enhancing conformational diversity and enabling applications where protein size is unknown or dictated by functional or structural requirements.

To the best of our knowledge, this is the first generative framework to explicitly incorporate mechanisms from biological protein synthesis and co-translational folding.

2. Related Works

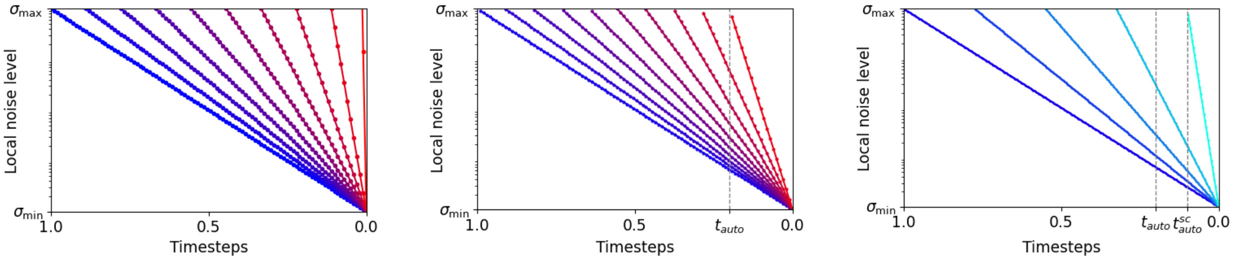
Parameterization in Protein Generative Models. Following the success of AlphaFold2 (Jumper et al., 2021), frame cloud parameterization emerged as a dominant approach in protein generative modeling. Here, each residue is represented by a local coordinate frame—defined by a translation and rotation. Models like FrameDiff (Yim et al., 2023) perform Riemannian manifold diffusion, while FoldFlow (Bose et al., 2024) and MultiFlow (Campbell et al., 2024) extend this with Riemannian flow matching. In contrast, point cloud parameterizations operate directly on Cα or all-atom positions, offering flexibility and simplicity. Recent models such as AlphaFold3 (Abramson et al., 2024), Protpardelle (Chu et al., 2024), and Pallatom (Qu et al., 2024) adopt this strategy for all-atom generation using unified coordinate spaces. Most relevant to our work, however, are AlphaFold1 (Senior et al., 2020) and Folding Diffusion (Wu et al., 2024), which model proteins using internal coordinates—bond / dihedral angles and interatomic distances—providing a compact, invariant representation that aligns well with the natural folding process.

Diffusion Process in Torsion Space. Diffusion process in torsion space was studied in (Jing et al., 2022). The perturbation kernel $p_{t|0}(\mathbf{X}_t|\mathbf{X}_0)$ for rescaled Brownian motion on $\mathbb{T}^{n\times d}$ manifests as the wrapped normal distribution on $\mathbb{R}^{n\times d}$, given by: $p_{t|0}(\mathbf{X}_t|\mathbf{X}_0) \propto \sum_{\mathbf{d}\in\mathbb{Z}^{d\times n}} \exp(-||\mathbf{X}_0 - \mathbf{X}_t + 2\pi\mathbf{d}||^2/2\sigma_t^2)$. The score $\nabla_{\mathbf{X}_t} \log p_{t|0}(\mathbf{X}_t|\mathbf{X}_0)$ then is approximated practically with $N = 1000$ as follows:

$$\nabla_{\mathbf{X}_t} \log p_{t|0}(\mathbf{X}_t|\mathbf{X}_0) \simeq \frac{\sum_i -\Delta_i \exp(-\Delta_i^2/2\sigma_t^2)}{\sigma_t^2 \sum_i \exp(-\Delta_i^2/2\sigma_t^2)} \quad (1)$$

where $\Delta_i = \mathbf{X}_0 - \mathbf{X}_t + 2\pi i$ and $i \in \{-N, .., N\}$.

Autoregressive Diffusion Model. Compared to fully autoregressive methods (Billera et al., 2024), which fix residue features once added, autoregressive diffusion allows bidirectional adjustments between early and late features. Among diffusion models with per-index noise schedules, Rolling Diffusion (Ruhe et al., 2024) applies a uniform denoising speed across all video frames, while AR-Diffusion (Wu et al., 2023) uses decreasing denoising speeds for natural language sequences. Diffusion Forcing (Chen et al., 2024) proposes a general framework supporting diverse autoregressive schedules across video, planning, and robotics. To our knowledge, RiboFold is the first work to employ per-index autoregressive diffusion specifically for protein generation, with a schedule carefully designed to mimic the dynamics of natural protein folding.



(a) Backbone noise schedule without t_{auto} (b) Backbone noise schedule with t_{auto} (c) All-atom noise schedule of $i = 0$

Figure 2. Trans-dimensional autoregressive noise schedule and their sampling steps. Feature dimensions having $\sigma > \sigma_{\max}$ are masked out, and the model decides whether to keep adding or halt during the sampling process. The first residue (blue) starts denoising from the first step. Middle residues (purple) are initialized and start denoising from the middle of the sampling. Early dimensions are already partially denoised when these middle dimensions are added; later residues require less sampling steps. (b) t_{auto} helps to provide sufficient sampling steps for the last residues. (c) Branched autoregressive schedule adds new sidechain dimensions with uniform rate.

3. Ribosomal Protein Generator

3.1. Homogeneous Protein Generator (HomoFold)

We first draw inspiration from the *in vitro* protein folding process, where denatured proteins **simultaneously** fold into their native state. Starting from randomly perturbed bond and dihedral angles of the protein backbone, our base model employs torsional diffusion (Jing et al., 2022) to iteratively denoise these angles and recover the native conformation. While closely related to prior work (Wu et al., 2024), we introduce two key innovations: (i) a rigorously designed noise schedule that allocates perturbations in proportion to their impact on local and global structure, and (ii) improved parameterization and architecture for stability and fidelity.

Structure Parameterization. We represent the individual amino acid conformations using a vector in d-dimensional hypertorus, $\mathbf{x} \in \mathbb{T}^d$, where each dimension signifies the bond and dihedral angles. A protein structure with $n \in \mathbb{N}$ amino acids therefore consists of $n \times d$ variables, denoted as $\mathbf{X} \triangleq \mathbf{x}^{0:n-1} \in \mathbb{T}^{n \times d}$.[1] Backbone structures are represented with $d = 6$, with three bond angles and three dihedral angles.[2] Sidechain structures are parameterized by up to four additional dihedral angles (Jumper et al., 2021). In total, we use $d = 10$ for all-atom protein structure.

Noise Schedule. We *design* the forward diffusion process such that the most significant global and local structural changes occur around the midpoint of the sampling timesteps $t \simeq 0.5$, as measured by TM-score and LDDT (Mariani et al., 2013), respectively. To achieve this, we introduce an exponential noise schedule (Song & Ermon, 2019) defined as $\sigma_t = \sigma_{\min}^{1-t}\sigma_{\max}^{t}$, where $(\sigma_{\max}, \sigma_{\min}) = (2\pi, 2\pi/e^{10})$ for both bond and dihedral angles (Figure S2).

Network and Training Objectives. The network approximates the true score by first predicting the trigonometric representation of the injected noise per angle, $\text{Trig}^{\theta}(\boldsymbol{\epsilon_t}) := [\sin^{\theta}(\boldsymbol{\epsilon_t}), \cos^{\theta}(\boldsymbol{\epsilon_t})] \in \mathbb{R}^{n \times 2d}$. The noise in radians is then reconstructed via $\boldsymbol{\epsilon}_t^{\theta} = \arctan(\sin^{\theta}(\boldsymbol{\epsilon}_t)/\cos^{\theta}(\boldsymbol{\epsilon}_t))$, which is then used to approximate the score via Equation (1). The denoising score matching objective is as follows with residue index i and feature index d:

$$\mathcal{L}_{\text{DSM}}^{\theta,i,d} = \mathbb{E}_t\left[\frac{1}{\sigma_t}||\text{Trig}^{\theta}(\boldsymbol{\epsilon}_t^{i,d}) - \text{Trig}(\boldsymbol{\epsilon}_t^{i,d})||^2\right] \qquad (2)$$

We use the Diffusion Transformer (Peebles & Xie, 2023) for improved scalability, and incorporate Rotary Positional Encoding (RoPE) (Su et al., 2024) to ensure equivariance with respect to length-dependent positional shifts.

3.2. Ribosomal Protein Generator

While homogeneous protein generators produce physically plausible backbones, they often suffer from mode collapse due to the highly concentrated distribution of dihedral angles around helical regions (Figure 5(d)). To mitigate this, we introduce an autoregressive component along with its trans-dimensional extension. This design encourages early segments to explore a broader conformational space, while avoiding biases associated with predefined length.

Soft Autoregressive Diffusion Schedule. We design an index-dependent noise schedule so that later residues are conditioned on earlier ones. We assign each residue i a linear target time $t_i = 1 - i/n$ at which it should reach the maximum noise level $\sigma_{\max}$ (Figure 2(a)). Under an exponential schedule, we obtain $\sigma_{t_i}^{i} = (\sigma_{\max}^{i})^{t_i} \cdot (\sigma_{\min})^{1-t_i} = \sigma_{\max}$ with fixed $\sigma_{\min}$, which eventually gives $\log \sigma_{\max}^{i} = \frac{1}{t_i}(\log \sigma_{\max} - (1 - t_i) \log \sigma_{\min}) = \frac{1}{n-i}(n \log \sigma_{\max} - i \log \sigma_{\min})$. Practically, to ensure that final residues have sufficient denoising steps, we introduce a time threshold $t_{\text{auto}} \in [0, 1]$, and redefine $\sigma_{\max}^{i}$ using $t_i = t_{\text{auto}} + (1 - t_{\text{auto}})(n - i)/n = 1 - i(1 - t_{\text{auto}})/n$ (Figure 2(b)).

[1] Throughout paper, superscripts and subscripts denote residue indices $i \in [0{:}n{-}1]$ and diffusion timesteps $t \in [0, 1]$, respectively.

[2] Internal coordinate parameterization generally requires three additional degrees of freedom by bond lengths; however, as shown in Figure S1, single-chain protein structures could be reconstructed with only six angles with (pre-calculated) mean bond lengths.

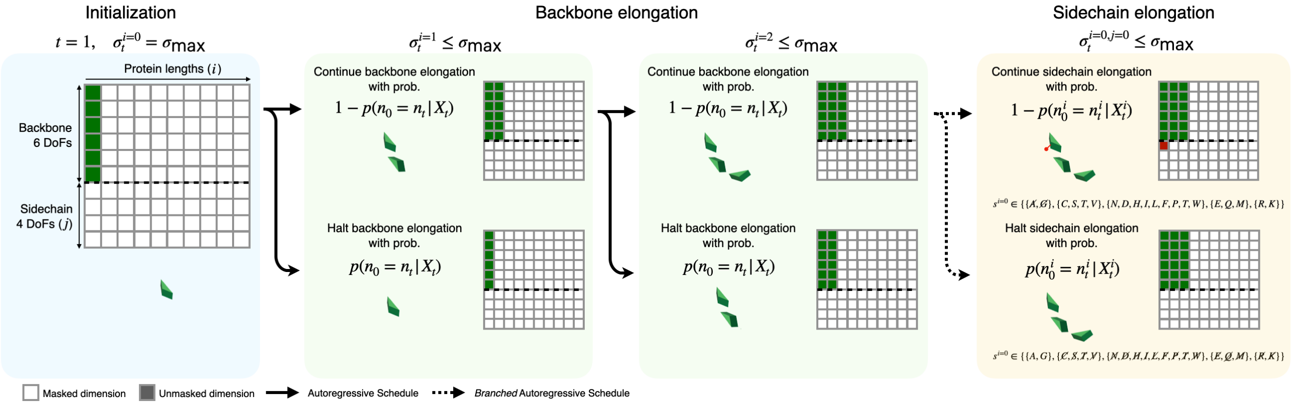

Initialization
$t = 1, \quad \sigma_t^{i=0} = \sigma_{\max}$
Protein lengths (i)
Backbone 6 DoFs
Sidechain 4 DoFs (j)
Backbone elongation
$\sigma_t^{i=1} \le \sigma_{\max}$
Continue backbone elongation with prob.
$1 - p(n_0 = n_t | X_t)$
Halt backbone elongation with prob.
$p(n_0 = n_t | X_t)$
$\sigma_t^{i=2} \le \sigma_{\max}$
Continue backbone elongation with prob.
$1 - p(n_0 = n_t | X_t)$
Halt backbone elongation with prob.
$p(n_0 = n_t | X_t)$
Sidechain elongation
$\sigma_t^{i=0,j=0} \le \sigma_{\max}$
Continue sidechain elongation with prob.
$1 - p(n_0^i = n_t^i | X_t^i)$
$s^{i=0} \in \{\{\cancel{A}, \cancel{G}\}, \{C, S, T, V\}, \{N, D, H, I, L, F, P, T, W\}, \{E, Q, M\}, \{R, K\}\}$
Halt sidechain elongation with prob.
$p(n_0^i = n_t^i | X_t^i)$
$s^{i=0} \in \{\{A, G\}, \{\cancel{C}, \cancel{S}, \cancel{T}, \cancel{V}\}, \{\cancel{N}, \cancel{D}, \cancel{H}, I, \cancel{L}, F, P, \cancel{T}, \cancel{W}\}, \{\cancel{E}, \cancel{Q}, \cancel{M}\}, \{\cancel{R}, \cancel{K}\}\}$
□ Masked dimension
■ Unmasked dimension
Autoregressive Schedule
Branched Autoregressive Schedule

Figure 3. Illustration of our generative process. (Left, blue) We initialize from a single amino acid backbone with six degrees of freedom. (Center, green) Following the autoregressive schedule, the model determines whether to add a new backbone residue. At this stage, each residue can represent any amino acid type. (Right, yellow) Following a branched schedule, the model decides whether to add side-chain dimensions. Halting sidechain elongation implies that the residue will ultimately be alanine or glycine, which have no additional torsion angles; otherwise, it will be one of the remaining 18 residue types. Likewise, the model sequentially prunes out possible amino acid types.

Sequential Backbone Elongation. We observe that when the diffusion noise level $\sigma_t > 2\pi$, Brownian perturbations dominate, and the corresponding features no longer carry meaningful structural information—only the chain length $n_{t=0}$ remains encoded. These features in fact impose a fixed-length prior, constraining the structural diversity.

We therefore remove this implicit length prior and enable flexible-length generation by masking out features $\mathbf{x}_{t>t_{2\pi}}$, where $\sigma_{t_{2\pi}} = 2\pi$. This ensures the model operates only on meaningful structural information, enabling exploration of the earlier residues on more diverse substructures.

Network and Traing Objective. The generative process begins from a single amino acid backbone (Figure 3, left). As sampling progresses, the model determines whether to append a new amino acid backbone by predicting the probability of extension versus termination (Figure 3, center) from mean-pooled representation from noisy structures. Formally, we parameterize $y^\theta_{\text{stop}} = p^\theta(n_0 = n_t \mid \mathbf{X}_t)$ and halt chain growth with probability $p^\theta(n_0 = n_t \mid \mathbf{X}_t)$. The training objective is then as follows:

$$\mathcal{L}^\theta_{\text{BB}} = \mathbb{E}_t \left[BCE(y^\theta_{\text{stop},t}, y_{\text{stop}}) \right] \tag{3}$$

where $BCE(y^\theta, y) = -\left[y^\theta \log y + (1 - y^\theta) \log(1 - y) \right]$ is a binary cross entropy, $y_{\text{stop}} = 1$ if $n_0 = n_t$, 0 otherwise.

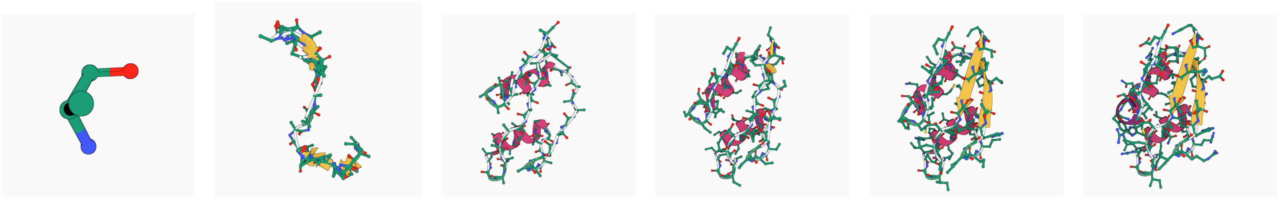
Biological Implications. This framework is motivated by the broader class of biological processes that exhibit autoregressive characteristics during protein folding. While co-translational folding during ribosomal synthesis is a primary inspiration, similar mechanisms arise in protein translocation across membranes and in chaperone-assisted refolding during protein quality control (Choi et al., 2012).

3.3. All-atom Ribosomal Protein Generators

Generating all-atom protein structures is inherently challenging because the dimensionality of the generative space depends on the amino acid sequence. The total number of variables is unknown *a priori*, as each residue can contribute up to ten heavy side-chain atoms and four side-chain dihedral angles. Conventional generative models cannot accommodate this variability; they require a fixed-dimensional representation before sampling. Some models address this by introducing additional degrees of freedom—i.e., “fake” features—that are discarded during post-processing (Chu et al., 2024; Qu et al., 2024). While workable, this approach increases computational cost and may hinder learning by exposing the model to operate in an artificially inflated space.

To overcome this limitation, we introduce an additional residue-wise trans-dimensional diffusion framework. Building on a growing backbone representation, our model dynamically adds up to four angular dimensions per residue (corresponding to χ_0–χ_3 dihedral angles). This design enables the model to determine—on the fly—how many side-chain degrees of freedom to include, allowing for end-to-end, variable-length all-atom generation.

Branched Autoregressive Diffusion Schedule. Our branched diffusion schedule can work either in homogeneous or ribosomal settings. Given t_i (1 for homogeneous, $1 - i/N$ for ribosomal), the timepoint where i-th backbone is added, we assign each side chain dimension $j \in [0, 3]$ a linear target time $t_{ij} = t_i(3 - j)/4$ to reach the maximum noise level $\sigma_{\max}$. Similar to the backbone schedule, we add threshold $t^{\text{sc}}_{\text{auto}} \in [0, 1]$ to ensure that the final sidechain dihedral angle has sufficient denoising time (Figure 2(c)).



(a) Initialization (b) BB elongation (c) Helix formation (d) SC elongation (e) Strand formation (f) Final structure

Figure 4. Example generative process from RiboFold.

Branched Side-chain Elongation. Similar to backbone elongation, we mask out features $\mathbf{x}_{t>t_{2\pi}}$ where $\sigma_{t_{2\pi}} = 2\pi$. This introduces a natural side chain packing algorithm, allowing the model to defer amino acid type selection until sufficient structural context is available. Rather than committing to a residue identity upfront, the model begins with the full set of 20 amino acid candidates and progressively narrows the candidate space as generation proceeds. For instance, once an additional side-chain dimension is added, residues like alanine or glycine—which lack such dimensions—can be pruned from consideration (Figure 3, right). This sequential pruning prioritizes chemically plausible choices and enables delayed decisions for bulkier residues, more closely mirroring rational side-chain packing.

Network and Training Objective. The network predicts the sequence identity $y^{\theta,i}_{\text{aatype},t} = p^{\theta}(s^i_0 \mid \mathbf{x}^i_t)$ per residue, then deterministically converts it to binary probability $p^{\theta}\left(n^i_0 = n^i_t \mid \mathbf{x}^i_t\right)$ where n^i_0 is the ground truth number of sidechain dimensions and n^i_t is current number of dimensions. The model then halts sidechain elongation with probability $p^{\theta}\left(n^i_0 = n^i_t \mid \mathbf{x}^i_t\right)$. The training objective is then as follows:

$$\mathcal{L}^{\theta,i}_{\text{SC}} = \mathbb{E}_t \left[CE(y^{\theta,i}_{\text{aatype},t}, y^i_{\text{aatype}}) \right] \tag{4}$$

where $CE(y, \hat{y}) = -\sum_k y_k \log \hat{y}_k$ is cross entropy over 20 canonical amino acid types.

Folding Trajectory. The generation trajectory of RiboFold is shown in Figure 4. We initialize from a single amino acid backbone, where bond and torsion angles are sampled from uniform distribution over $[0, 2\pi)$ (Figure 2(a)). Following the autoregressive diffusion schedule, each new residue is appended to the C-terminus of growing chain while the existing backbone features concurrently fold (Figure 2(b)). Our process shows *clustered folding* where local clusters of motifs are formed before the global structure is formed though it is not guided by folding simulations. Specifically, local structures (α helices) are easy to form, and they emerge in early stages (Figure 4(c)). Global structures (β strands) appear later in the process (Figure 4(e)). Following the branched sidechain elongation schedule, side chain dimensions start to grow from existing backbones (Figure 4(d)). The procedure concludes with a complete all-atom protein structure (Figure 4(f)).

Algorithm 1 Main Inference Loop

```
def RiboFoldSampling($N_{\text{steps}}$)
  $\mathbf{m} \leftarrow zeros(n_{max}, 10)$
  # Initialize with single amino acid backbone:
  $\mathbf{m}[0, :6] = 1$
  $\mathbf{x}_{\text{random}} \sim \mathcal{U}[0, 2\pi)^{(n_{\max}, 10)}$
  $\mathbf{x}_1 = \mathbf{x}_{\text{random}} \cdot \mathbf{m}$
  $dt = 1/N_{\text{steps}}$
  for all $i \in [0, ..., N_{\text{steps}} - 1]$ do
      $t = 1 - i \cdot dt$
      # Make masked feature with autoregressive schedule:
      $\sigma_t \leftarrow \text{DiffusionSchedule}(t)$
      $\mathbf{m}_t = \sigma_t < 2\pi$
      $\epsilon_t, p(n_0|\mathbf{x}_t), \{p(s^i_0|\mathbf{x}^i_t)\}^{n_t-1}_{i=0} = \text{Network}(\mathbf{x}_t, \mathbf{m}_t, t)$
      $\{p(n^i_0|\mathbf{x}^i_t)\}_i \leftarrow \{\text{Seq2Dim}(p(s^i_0|\mathbf{x}^i_t))\}_i$ in parallel
      # Feature update (folding):
      $\mathbf{x}_t = \text{DiffusionSampling}(\mathbf{x}_t, \epsilon_t, \mathbf{m}_t, t)$
      # Backbone elongation:
      if $u \sim \mathcal{U}(0, 1) > p(n_0|\mathbf{x}_t)$ then
          $\mathbf{x}_t = \text{where}(\mathbf{m}, \mathbf{x}_t, \mathbf{x}_{\text{random}} \cdot \mathbf{m}_t)$
          $\mathbf{m} = \mathbf{m}_t$
      end if
      # Sidechain elongation:
      if $u^i \sim \mathcal{U}(0, 1) > p(n^i_0 = n^i_t|\mathbf{x}^i_t)$ in parallel then
          $\mathbf{x}^i_t = \text{where}(\mathbf{m}^i, \mathbf{x}^i_t, \mathbf{x}^i_{\text{random}} \cdot \mathbf{m}^i_t)$
          $\mathbf{m}^i = \mathbf{m}^i_t$
      end if
  end for
Return: $\mathbf{x}_0, \mathbf{s}_0$
```

3.4. Training

During training, we optimize the following with the timestep sampled from the uniform distribution $t \sim \mathcal{U}[0, 1]$:

$$\mathcal{L}^{\theta} = \mathcal{L}^{\theta}_{\text{BB}} + \sum_{i=0}^{n_t-1} \left[\mathcal{L}^{\theta,i}_{\text{SC}} + \sum_{d=0}^{n^i_t-1} \mathcal{L}^{\theta,i,d}_{\text{DSM}} \right] \tag{5}$$

3.5. Inference

The inference algorithm is shown in Algorithm 1. `Seq2Dim` in algorithm is a deterministic mapping between sequence prediction with sidechain dimensions.

4. Experiments

4.1. Evaluation Metric: In-distribution Coverage

Our primary evaluation criterion is *in-distribution coverage* with respect to the training dataset—i.e., how faithfully each model reproduces the structural vocabulary present in the training data. Protein structure generators have traditionally been evaluated based on *designability*, and subsequently by *diversity* and *novelty* with the subset of designable structures (Watson et al., 2023; Yim et al., 2023). This emphasis on designability was driven by practical considerations: ensuring that a corresponding sequence can be identified for a generated backbone (enabling downstream protein expression experiments), and that the refolded all-atom structure retains the original backbone conformation.

However, designability-centric evaluation is inherently biased. It tends to favor helix-rich compact structures, and paradoxically, models exhibiting extremely high designability—surpassing that of the training data—often overfit to a narrow subset. For context, the average designability of natural structures is approximately 78% in the PDB (Faltings et al., 2025) and 56% in CATH (Lu et al., 2025). In contrast, state-of-the-art methods report designability approaching 99% (Campbell et al., 2024; Geffner et al., 2025), suggesting these models do not faithfully capture the structural diversity of the training data but are instead overoptimized.

To assess in-distribution coverage without such bias, we instead employ two complementary metrics: the *Fréchet Protein Distance* (FPD) and *CAT(H) Diversity*. These evaluate the distributional and categorical coverage, respectively.

Fréchet Protein Distance (FPD) Inspired by the Fréchet Inception Distance (FID) (Heusel et al., 2017), and similarly with (Lu et al., 2025; Geffner et al., 2025; Faltings et al., 2025), we compute the Fréchet Protein Distance by embedding both generated and reference protein structures into the representation space of ESM3 (Hayes et al., 2025), a large protein language model. We then calculate the Wasserstein-2 distance (Fréchet distance) between the resulting multivariate Gaussian distributions. This provides a holistic measure of distributional similarity, capturing both structural quality and diversity without requiring designable sequences.

CAT(H) Diversity. We count the number of unique structural classes using the CATH protein structure classification (Orengo et al., 1997; Sillitoe et al., 2021), which organizes protein domains into a four-level hierarchical scheme. We specifically focus on C (Class) label, where C=1,2,3 represent mainly α, mainly β, and mixture α/β, respectively. To assign CATH labels to generated structures, we filter out any structure with a maximum target length–normalized TM-score (Zhang & Skolnick, 2004) below 0.5 compared to the training set. Each remaining structure is then assigned the label of its most similar training structure based on TM-score. We report the number of unique labels discovered.

4.2. Experimental Details

Other Metrics. In addition to two in-distribution coverage metrics, we also assess following auxiliary statistics: (1) # Match: number of samples having in-distribution TM-score $\geq$ 0.5 (2) % helix/strand: ratio of secondary structures predicted using P-SEA algorithm (Labesse et al., 1997) implemented in biotite (Kunzmann et al., 2023) (3) # Clash: number of samples having steric clashes.

Baselines. We compare our model with two sequence-structure codesign generative models, each representing point cloud–based and frame cloud–based parameterizations. `Protpardelle` (Chu et al., 2024) is an all-atom protein generative model that applies a Euclidean diffusion to the superposition of all-atom coordinates. `MultiFlow` (Campbell et al., 2024) is a sequence-structure codesign model that performs flow matching on SE(3) backbone frames and discrete sequences.[3]

Training. We use the single-chain CATH S20 dataset—a curated subset of CATH filtered to $\leq 20\%$ sequence identity—containing 3,141 structures with lengths ranging from 50 to 128 residues. Training was performed over two days on 4 Nvidia A100 GPUs using the AdamW optimizer (Loshchilov & Hutter, 2019) with a learning rate of 0.0003. The baselines were trained on the same dataset under the same conditions until convergence for two days.

4.3. Unconditional Generation

Setup. To evaluate in-distribution coverage, we generated 3,000 structures per model without condition. For the baselines and the HomoFold, protein lengths were randomly sampled from the training dataset distribution. For RiboFold, structures were generated without explicit length specification. To ensure a fair comparison, low-temperature sampling was disabled across all models.

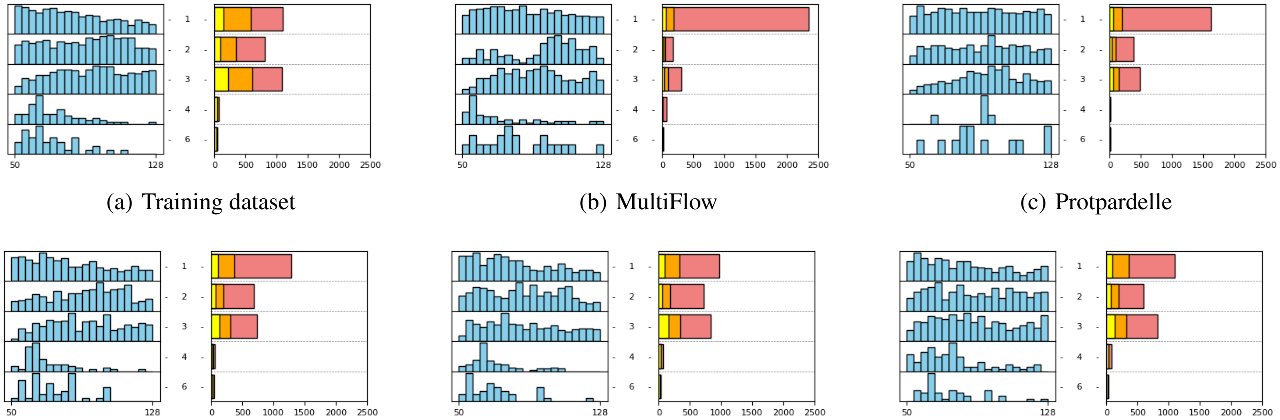
In-distribution Coverage. Table 1 presents the quantitative results. All models achieved a high match rate, indicating effective training on the dataset. In terms of in-distribution coverage, measured by Fréchet Protein Distance (FPD), RiboFold achieves a 96.6 % improvement over MultiFlow and 89.5 % over Protpardelle. For structural diversity, assessed via CAT(H) classification, our models discover twice as many unique classes as the baselines.

Figure 5 shows the distribution of predicted CATH classes among successful samples and Figure 6 shows PCA projection of ESM3 representations. Both MultiFlow and Protpardelle predominantly generate Class 1 structures, reflecting an inherent bias toward helical topologies. HomoFold produces significantly fewer Class 1 samples, while RiboFold most accurately reproduces the class distribution observed in the training set.

[3] MultiFlow does not output side-chain coordinates.

Table 1. In-distribution coverage of generated proteins. Fréchet Protein Distance (FPD), computed on ESM-3 encoder features, measures distributional similarity to the training set (lower is better). CAT(H) diversity evaluates categorical coverage based on structural class assignments from the most similar training structure (closer to the training distribution is better). While all models show similar success rates, RiboFold achieves the best coverage across both metrics.

	FPD(↓)	# CAT(↑) $[\alpha/\beta/\alpha+\beta]$	# CATH(↑) $[\alpha/\beta/\alpha+\beta]$	# Match $[\alpha/\beta/\alpha+\beta]$
Training Dataset	-	530 (— %) [141 / 97 / 221]	1,631 (— %) [580 / 349 / 611]	3,137 (— %) [1,091 / 810 / 1,085]
MultiFlow	127.	110 (21 %) [51 / 17 / 36]	326 (20 %) [183 / 46 / 90]	2,911 (98 %) [2,347 / 164 / 304]
Protpardelle	44.5	139 (26 %) [55 / 21 / 57]	412 (25 %) [213 / 65 / 128]	2,632 (88 %) [1,974 / 237 / 366]
HomoFold	6.39	350 (66 %) [106 / 66 / 135]	929 (57 %) [372 / 191 / 313]	2,829 (97 %) [1,281 / 685 / 733]
RiboFold-BB	2.11	357 (67 %) [98 / 60 / 156]	912 (56 %) [334 / 183 / 347]	2,658 (91 %) [969 / 725 / 829]
RiboFold	4.38	345 (65 %) [96 / 69 / 134]	932 (57 %) [356 / 197 / 325]	2,660 (91 %) [1,100 / 596 / 819]



(a) Training dataset (b) MultiFlow (c) Protpardelle

(d) HomoFold (e) RiboFold-BB (f) RiboFold

Figure 5. Distributions by CATH class. (Left) length distribution. RiboFold nearly reproduces length distribution of each class without prespecification. (Right) Sample distribution - (red) # samples, (orange) # unique CATH labels, (yellow) # unique CAT labels. Baselines show a strong bias toward class 1 helical structures. RiboFold better reflects training distribution.

A comparison between HomoFold and RiboFold shows that both models discover a similar number of unique CAT(H) classes. However, the sample distribution in HomoFold is heavily skewed toward Class 1, reflecting a strong structural bias. In contrast, RiboFold produces a distribution that more closely matches the training set, indicating improved coverage. This suggests that flexible-length generation in RiboFold allows for more faithful sampling from the underlying data distribution, avoiding the limitations imposed by fixed-length inputs.

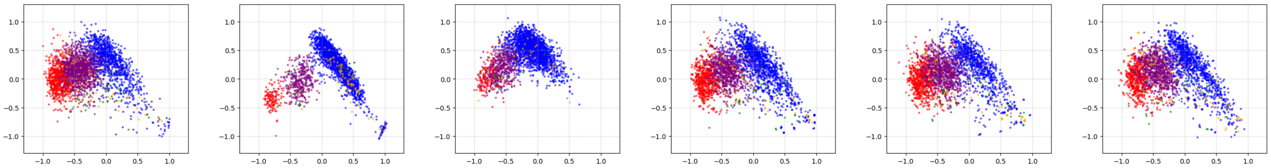
Unbiased Length Distribution. RiboFold supports unbiased length generation, as evidenced in Figures 5(e), 5(f). The length distribution within each CAT(H) class closely mirrors that of the training dataset, indicating that the model captures class-conditional length variation without explicit constraints. This result demonstrates that RiboFold is not only unbiased in terms of structural vocabulary, but also with respect to sequence length, enabling more faithful sampling across the full spectrum of training data characteristics.

Secondary Structure. In addition to global structural coverage, RiboFold supports unbiased secondary structure generation at the residue level. When compared to the training distribution, baseline models exhibit a strong bias toward α-helices, underrepresenting other structural elements. In contrast, RiboFold-BB closely matches the per-residue secondary structure distribution of the training data, while RiboFold slightly underrepresents β-strands but still achieves significantly better balance than the baselines. These results highlight RiboFold's ability to model local structural diversity more faithfully.

Table 2. Secondary structure and clash statistics.

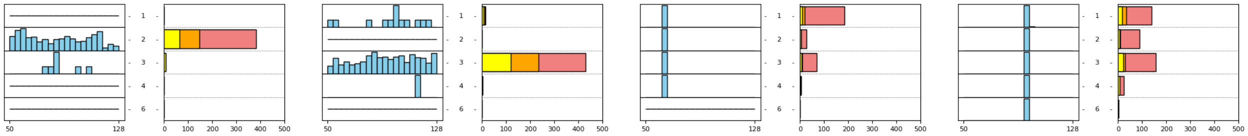
	α %	β %	# Clashes
Training Dataset	31.8	16.1	0
MultiFlow	72.1	0.7	8
HomoFold	37.9	14.4	26
RiboFold-BB	33.4	16.8	42
Protpardelle	51.3	0.6	132
RiboFold	37.0	14.9	205

Steric Clashes. However, internal coordinate representations have inherent limitations—most notably the lever-arm effect, where small angular deviations propagate along the chain and accumulate into large positional displacements. This effect directly contributes to steric clashes, as downstream atoms may violate spatial constraints despite locally valid torsion angles. In our model, this issue becomes increasingly pronounced with larger numbers of β-strands, where ensuring global geometric coherence is more critical.



(a) Training dataset (b) MultiFlow (c) Protpardelle (d) HomoFold (e) RiboFold-BB (f) RiboFold

Figure 6. PCA projections of ESM3 mean-pooled encoder embeddings. RiboFold generates protein structures that are most similar to the training distribution. Blue: class 1(mainly α); Red: class 2 (mainly β); Purple: class 3 (mixed α, β).



(a) C=2 (b) C=3 (c) n=64 (d) n=96

Figure 7. Length / CATH distributions after conditional generation. Conditional generation effectively steers the distribution, in both secondary structure conditioning (7(a), 7(b)) and length conditioning (7(c), 7(d)).

Table 3. Conditional Generation on CATH labels. Values represent unique $[\alpha/\beta/\alpha+\beta]$ classes. $w = 0.1$ was used for sampling.

Condition	# CAT	# CATH
Uncond.	43 / 33 / 60	98 / 70 / 111
C=1	71 / 0 / 2	196 / 0 / 2
C=2	0 / 68 / 2	0 / 159 / 2
C=3	9 / 0 / 132	14 / 0 / 249

4.4. Conditional generation

Setup. To evaluate the conditioning capabilities of our model, we train RiboFold-BB with both *secondary structure* and *final length* conditioning. For secondary structure conditioning, we experiment with and without Class (C) condition. For length conditioning, we trained with and without explicit length input. At inference time, we generate 500 structures using classifier-free guidance (Ho & Salimans, 2021) to enable flexible control during sampling.

Secondary Structure Conditioning. Figure 7(a) and 7(b) illustrate qualitative statistics of conditional generation. Since the internal coordinate parameterization offers explicit control over local geometry and secondary structure, conditioning with the C-label is particularly effective in RiboFold.

Length Conditioning. Figure 7(c) and Figure 7(d) illustrate sample statistics under length conditioning. While RiboFold naturally supports flexible-length generation, cases with specific target lengths can be effectively handled by setting a preconditioned length during sampling.

5. Discussion and Conclusion

In this work, we introduced **RiboFold**, a generative framework inspired by biological protein synthesis and co-translational folding. By leveraging internal coordinate parameterization and a trans-dimensional autoregressive diffusion process, RiboFold emulates the ribosome's incremental elongation mechanism, enabling more faithful and interpretable structure generation.

RiboFold achieves superior *in-distribution coverage*, significantly reducing structural bias—particularly the overrepresentation of helical topologies—and more accurately reproduces the class and length distributions of natural proteins. It also supports *flexible-length generation*, allowing protein size to emerge from structural or functional constraints, and offers fine-grained *conditional control* over structural features such as secondary structure and length.

However, our method does not fully resolve the *lever-arm effect*, where small angular perturbations compound over long chains, leading to global distortions and steric clashes—particularly in β-rich topologies. Moreover, RiboFold works only on single-chain structures.

Overall, RiboFold demonstrates that biologically grounded generative processes can improve structural realism and controllability. This ribosomally inspired formulation offers a strong foundation for future work in *de novo* protein design. Promising directions include refining global structure via hybrid structure parameterizations, incorporating experimental constraints, and extending the framework to support multi-chain and protein–ligand complex generation.

Acknowledgements

The authors thank Juno Nam for careful review of the manuscript and valuable suggestions. We also thank Jason Yim for insightful discussions. We acknowledge support from the Machine Learning for Pharmaceutical Discovery and Synthesis (MLPDS) consortium, the DTRA Discovery of Medical Countermeasures Against New and Emerging (DOMANE) threats program, the NSF Expeditions grant (award 1918839) Understanding the World Through Code, and the DSO Singapore grant on next generation techniques for protein ligand binding.

References

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630 (8016):493–500, 2024.

Billera, L., Oresten, A., Stålmarck, A., Sato, K., Kaduk, M., and Murrell, B. The continuous language of protein structure. *bioRxiv*, pp. 2024–05, 2024.

Bose, J., Akhound-Sadegh, T., Huguet, G., FATRAS, K., Rector-Brooks, J., Liu, C.-H., Nica, A. C., Korablyov, M., Bronstein, M. M., and Tong, A. SE(3)-stochastic flow matching for protein backbone generation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=kJFIH23hXb.

Campbell, A., Yim, J., Barzilay, R., Rainforth, T., and Jaakkola, T. Generative Flows on Discrete State-Spaces: Enabling Multimodal Flows with Applications to Protein Co-Design. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=kQwSbv0BR4.

Chen, B., Martí Monsó, D., Du, Y., Simchowitz, M., Tedrake, R., and Sitzmann, V. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *Advances in Neural Information Processing Systems*, 37:24081–24125, 2024.

Choi, S. I., Son, A., Lim, K.-H., Jeong, H., and Seong, B. L. Macromolecule-assisted de novo protein folding. *International journal of molecular sciences*, 13(8):10368–10386, 2012.

Chu, A. E., Kim, J., Cheng, L., El Nesr, G., Xu, M., Shuai, R. W., and Huang, P.-S. An all-atom protein generative model. *Proceedings of the National Academy of Sciences*, 121(27):e2311500121, 2024.

Coutsias, E. A., Seok, C., Jacobson, M. P., and Dill, K. A. A kinematic view of loop closure. *Journal of computational chemistry*, 25(4):510–528, 2004.

Faltings, F., Stark, H., Jaakkola, T., and Barzilay, R. Protein fid: Improved evaluation of protein structure generative models, 2025. URL https://arxiv.org/abs/2505.08041.

Geffner, T., Didi, K., Zhang, Z., Reidenbach, D., Cao, Z., Yim, J., Geiger, M., Dallago, C., Kucukbenli, E., Vahdat, A., and Kreis, K. Proteina: Scaling flow-based protein structure generative models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=TVQLu34bdw.

Hayes, T., Rao, R., Akin, H., Sofroniew, N. J., Oktay, D., Lin, Z., Verkuil, R., Tran, V. Q., Deaton, J., Wiggert, M., et al. Simulating 500 million years of evolution with a language model. *Science*, pp. eads0018, 2025.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Ho, J. and Salimans, T. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. URL https://openreview.net/forum?id=qw8AKxfYbI.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Ingraham, J. B., Baranov, M., Costello, Z., Barber, K. W., Wang, W., Ismail, A., Frappier, V., Lord, D. M., Ng-Thow-Hing, C., Van Vlack, E. R., et al. Illuminating protein space with a programmable generative model. *Nature*, 623(7989):1070–1078, 2023.

Jing, B., Corso, G., Chang, J., Barzilay, R., and Jaakkola, T. Torsional Diffusion for Molecular Conformer Generation. *Advances in Neural Information Processing Systems*, 35: 24240–24253, 2022.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

Kolodny, R., Guibas, L., Levitt, M., and Koehl, P. Inverse kinematics in biology: the protein loop closure problem. *The International Journal of Robotics Research*, 24(2-3): 151–163, 2005.

Kunzmann, P., Müller, T. D., Greil, M., Krumbach, J. H., Anter, J. M., Bauer, D., Islam, F., and Hamacher, K. Biotite: new tools for a versatile python bioinformatics library. *BMC bioinformatics*, 24(1):236, 2023.

Labesse, G., Colloc'h, N., Pothier, J., and Mornon, J.-P. P-sea: a new efficient assignment of secondary structure from cα trace of proteins. *Bioinformatics*, 13(3):291–295, 1997.

Lin, Y. and AlQuraishi, M. Generating novel, designable, and diverse protein structures by equivariantly diffusing oriented residue clouds. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 20978–21002, 2023.

Lipman, Y., Chen, R. T. Q., Ben-Hamu, H., Nickel, M., and Le, M. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Lu, T., Liu, M., Chen, Y., Kim, J., and Huang, P.-S. Assessing generative model coverage of protein structures with shapes. *bioRxiv*, pp. 2025–01, 2025.

Mariani, V., Biasini, M., Barbato, A., and Schwede, T. lddt: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, 29(21):2722–2728, 2013.

Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B., and Thornton, J. M. Cath–a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1109, 1997.

Peebles, W. and Xie, S. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Qu, W., Guan, J., Ma, R., Zhai, K., Wu, W., and Wang, H. P (all-atom) is unlocking new path for protein design. *bioRxiv*, pp. 2024–08, 2024.

Rohl, C. A., Strauss, C. E., Misura, K. M., and Baker, D. Protein structure prediction using rosetta. In *Methods in enzymology*, volume 383, pp. 66–93. Elsevier, 2004.

Ruhe, D., Heek, J., Salimans, T., and Hoogeboom, E. Rolling diffusion models. In *International Conference on Machine Learning*, pp. 42818–42835. PMLR, 2024.

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W., Bridgland, A., et al. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792): 706–710, 2020.

Sillitoe, I., Bordin, N., Dawson, N., Waman, V. P., Ashford, P., Scholes, H. M., Pang, C. S., Woodridge, L., Rauer, C., Sen, N., et al. Cath: increased structural coverage of functional space. *Nucleic acids research*, 49(D1):D266–D273, 2021.

Simons, K. T., Kooperberg, C., Huang, E., and Baker, D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *Journal of molecular biology*, 268(1):209–225, 1997.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., and Liu, Y. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Wang, C., Qu, Y., Peng, Z., Wang, Y., Zhu, H., Chen, D., and Cao, L. Proteus: Exploring protein structure generation for enhanced designability and efficiency. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=IckJCzsGVS.

Watson, J. L., Juergens, D., Bennett, N. R., Trippe, B. L., Yim, J., Eisenach, H. E., Ahern, W., Borst, A. J., Ragotte, R. J., Milles, L. F., et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976): 1089–1100, 2023.

Wu, K. E., Yang, K. K., van den Berg, R., Alamdari, S., Zou, J. Y., Lu, A. X., and Amini, A. P. Protein Structure Generation via Folding Diffusion. *Nature communications*, 15(1):1059, 2024.

Wu, T., Fan, Z., Liu, X., Zheng, H.-T., Gong, Y., Jiao, J., Li, J., Guo, J., Duan, N., Chen, W., et al. Ar-diffusion: Autoregressive diffusion model for text generation. *Advances in Neural Information Processing Systems*, 36:39957–39974, 2023.

Xu, D. and Zhang, Y. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins: Structure, Function, and Bioinformatics*, 80(7):1715–1735, 2012.

Yim, J., Trippe, B. L., De Bortoli, V., Mathieu, E., Doucet, A., Barzilay, R., and Jaakkola, T. Se (3) diffusion model with application to protein backbone generation. In *International Conference on Machine Learning*, pp. 40001–40039. PMLR, 2023.

Zhang, Y. and Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57(4):702–710, 2004.

A. Minimal Internal Coordinate Parameterization.

Figure S1 shows backbone reconstruction with subset of degrees of freedom. Parameterizing dihedral angles with bond angles reproduce 100% input protein structures across all lengths, with and without bond distance parameterization. Omitting bond angle parameterization and replacing with mean bond angles, however, cannot reconstruct input backbone structures.

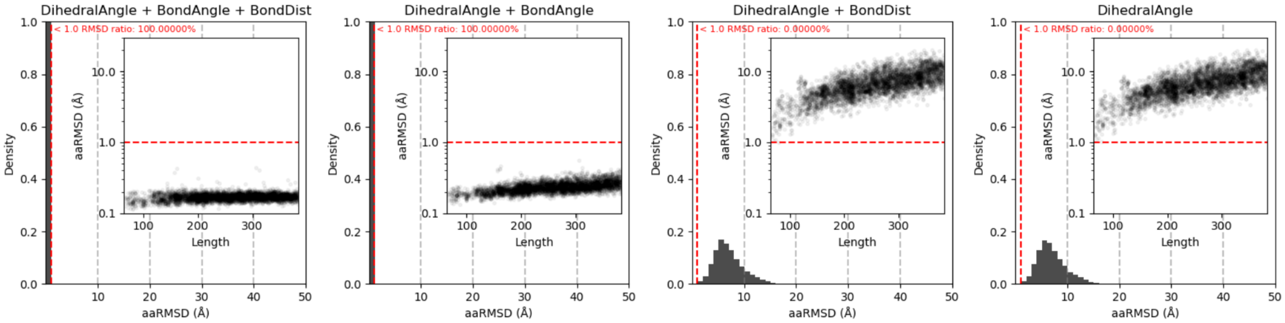
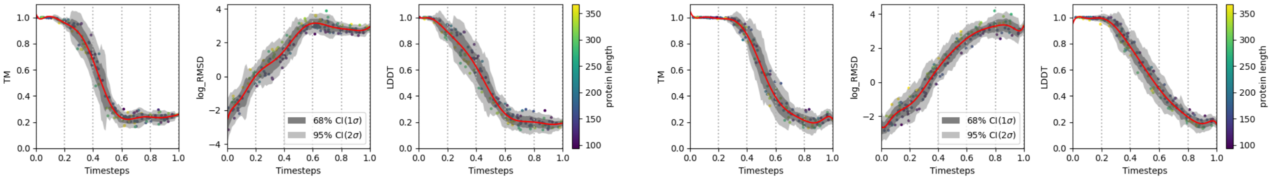


Figure S1. Reconstruction success rate with each degree of freedoms. Bond- and torsion-angle parameterization gives minimum degrees of freedom with 100 % reconstruction.

B. Noise Schedule Details.

Figure S2 shows local and global structural similarity after imposing noise following forward noise schedule. Across diverse protein lengths, our forward noise schedule imposes the steepest local and global structural change at the midpoint.



(a) bond angle

(b) torsion angle

Figure S2. Structure similarity along the forward diffusion process, (a) on torsion angles and (b) on bond angles. (left) measured by TM-score, (middle) measured by log-RMSD, and (right) measured by LDDT.