Demystifying Protein Generation with Hierarchical Conditional Diffusion Models

Zinan Ling[1], Yi Shi[2], Brett McKinney[1], Da Yan[3], Yang Zhou[4], Bo Hui[1]
[1]University of Tulsa [2]Johns Hopkins University
[3]Indiana University Bloomington [4]Auburn University
bo-hui@utulsa.edu

Abstract

Generating novel and functional protein sequences is critical to a wide range of applications in biology. Recent advancements in conditional diffusion models have shown impressive empirical performance in protein generation tasks. However, reliable generation of proteins remains an open research question in *de novo* protein design, especially when it comes to conditional diffusion models. Considering the biological function of a protein is determined by multi-level structures, we propose a novel multi-level conditional diffusion model that integrates both sequence-based and structure-based information for efficient end-to-end protein design guided by specified functions. By generating representations at different levels simultaneously, our framework can effectively model the inherent hierarchical relations between different levels, resulting in an *informative and discriminative* representation of the generated protein. We also propose Protein-MMD (Maximum Mean Discrepancy), a new reliable evaluation metric, to evaluate the quality of generated protein with conditional diffusion models. Our new metric is able to capture both distributional and functional similarities between real and generated protein sequences while ensuring conditional consistency. Using conditional protein generation tasks with benchmark datasets, we demonstrate the efficacy of the proposed protein generation framework and evaluation metric.

1 Introduction

Designing proteins with specific biological functions is a fundamental yet formidable challenge in biotechnology. It benefits wide-ranging applications from synthetic biology to drug discovery [1–5, 5, 6]. The challenge arises from the intricate interplay between protein sequence, structure, and function, which has not yet been fully understood [7]. Traditional methods, such as directed evolution, rely on labor-intensive trial-and-error approaches involving random mutations and selective pressures, making the process time-consuming and costly [8]. Recently, generative models have emerged as promising tools for protein design, enabling the exploration of vast sequence-structure-function landscapes [9–12]. However, existing generative models—including those focused on enzyme engineering, antibody creation, and therapeutic protein development—are typically task-specific and require retraining for new design objectives [10, 11]. These limitations impede their adaptability and scalability across different protein families.

While conditional generative models offer an end-to-end solution by directly linking the design process to the guidance, these models have been applied to protein generation [13–15]. In conditional protein generation tasks, maintaining conditional consistency across diverse contexts and ensuring functional relevance are critical [16, 17]. Specifically, the generated proteins should fully adhere to the specified functional constraints [18]. At the same time, achieving diversity and novelty in generated proteins is essential for successful design. In the literature, structural novelty can be assessed using Foldseek [19], which performs rapid protein structure searches against databases

Published in the AI for Science workshop (NeurIPS 2025).

like PDB [20] and AlphaFold [21] to ensure the generated proteins are novel compared to known structures. Diversity is measured using TM-score [22], which calculates structural variation between the generated proteins themselves and between the generated and wild-type proteins [17].

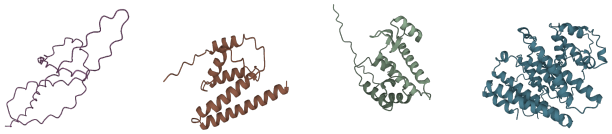

Figure 1: Protein visualization.

Despite the success of existing diffusion models in protein generation, these models only generate the protein representation at a single level and ignore hierarchical relations among different levels of representations. Choosing the level of granularity at which representing the comprehensive information of the protein raises significant concerns about the reliability of generated proteins in real-world applications. Motivated by the need to capture both the structural and functional nuances of protein sequences, we propose a novel multi-level conditional generative diffusion model for protein design that integrates both sequence-based [6] and structure-based [23] hierarchical information. Specifically, our proposed method generates the protein at three different levels: the amino acid level, the backbone level, and the all-atom level. Generation at multi-levels enables efficient end-to-end generation of proteins with specified functions and modeling the inherent hierarchical relations between different representations, resulting in an informative and discriminative representation of the protein. Also, the conditional diffusion flow in the architecture preserves the hierarchical relations between different levels. Intuitively, a representation at the lower level (e.g., the atom level) can decide the potential representation space at the higher level (e.g., the amino acid level). Modeling such hierarchical relations can guarantee consistency at different levels. Our model incorporates a rigid-body 3D rotation-invariant preprocessing step combined with an autoregressive decoder to maintain SE(3)-invariance, ensuring accurate modeling of protein structures in 3D space. Figure 1 shows the proteins generated by different methods with the same input. The thin line indicates that the sequence is unlikely to undergo meaningful folding into a stable 3D structure. Compared with the baselines, our method can generate discriminative and functional proteins.

We remark that it is still unknown how to assess the *conditional consistency* [18] in *de novo* protein design. Specifically, the fundamental problem of properly evaluating conditional consistency is quantifying to what extent the generated protein adheres to the specified functional constraints. Unlike computer vision, where metrics such as FID [24] have become a standard for assessing generated images, it is unclear whether such metrics are suitable for protein generation tasks. In protein design, the generated output cannot be as easily visualized or assessed as in images, making the choice of evaluation metrics even more critical. Therefore, how to adapt metrics like FID or Maximum Mean Discrepancy [18] presents challenges. To address the challenges of evaluating the *conditional consistency*, we propose *Protein-MMD*, a metric based on Maximum Mean Discrepancy (MMD), to better capture both distributional and functional similarities between real and generated protein sequences, while ensuring conditional consistency. We prove that our Protein-MMD provides a more accurate measure that reflects the given condition. Experiments demonstrate that our proposed model outperforms existing approaches in generating diverse, novel, and functionally relevant proteins. Our main contributions are summarized as follows:

- We design a novel multi-level conditional generative diffusion model that integrates sequence-based and structure-based information for efficient end-to-end protein design.
- We highlight the limitations of current evaluation metrics in protein generation and propose *Protein-MMD*, a novel metric to evaluate conditional consistency for protein generation.
- We experiment with standard datasets to verify the effectiveness of the proposed model. Our evaluation metric paves the way for reliable protein design with given conditions.

2 Methodology

2.1 Multi-level Diffusion

Motivated by the need to capture both the structural and functional nuances of protein sequences, we propose a multi-level diffusion model to generate information about a protein at three levels: the amino acid level, the backbone level, and the all-atom level. By constructing representations at different levels, our framework effectively integrates the inherent hierarchical relations of proteins,

resulting in a more rational protein generative model. We remark that there are hierarchical relations among different levels. To the best of our knowledge, this work is the first diffusion model to generate information at three levels and leverage the hierarchical relation between different levels.

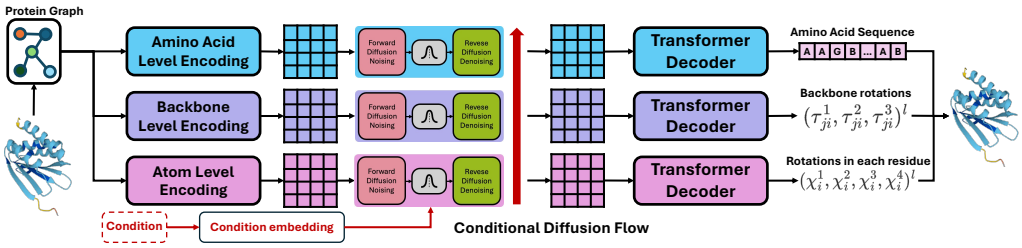

Figure 2: The architecture of the multi-level diffusion model.

Figure 2 shows the architecture of our model. At each level, the information will be encoded with its own set of embeddings and processed through a conditional diffusion flow where the condition comes from a lower level. With decoders, the sequence, backbone rotations, and residue rotations will be combined to indicate the complete information of a generated protein.

Amino Acid Level Representation. As the 3D conformation dictates biochemical interactions [3, 7], we first represent a protein's structure as a graph $\mathcal{G}_a = (\mathcal{V}_a, \mathcal{E}_a)$, where $\mathcal{V}_a$ is the set of nodes corresponding to residues (amino acids), and $\mathcal{E}_a$ is the set of edges representing interactions between two residues. Specifically, an edge between two nodes v_i and v_j is established if the Euclidean distance between their Cα atoms in 3D space is below a certain threshold, indicating a potential biochemical or structural interaction. At the amino acid level, each node $v_i \in \mathcal{V}_a$ corresponds to an amino acid and is represented by a vector $v_i = (\phi_i; h_i)$, where $\phi_i \in \mathbb{R}^3$ denotes the spatial coordinates of the amino acid's Cα atom in three-dimensional space, and h_i abstracts biochemical or structural properties. Each edge is represented as an embedding of the sequential distance [25].

Backbone Level Representation. An amino acid consists of backbone atoms and side chain atoms. Similarly, we use backbone atom (C, N, C_α) coordinates as the feature of in node of the backbone $\mathcal{V}_b$. We follow [25] to compute three Euler angles $\tau^1_{i,j}$, $\tau^2_{i,j}$, $\tau^3_{i,j}$ between two backbone atoms i and j. The angles will be integrated with the sequential distance as the edge feature. Backbone-level representation derives finer-grained protein information. With the three angles, the orientation between any two backbone planes can be determined to capture the backbone structures.

Atom Level Representation. Atom-level representation considers all atoms in the protein and provides the most fine-grained information. There are several methods to treat an atom as a node in the representation [26, 27]. Side chain torsion angles are important properties of protein structures [21]. In this paper, we also consider geometric representation at the atom level by incorporating the first four torsion angles: χ^1_i, χ^2_i, χ^3_i, and χ^4_i. With the complete geometric representation at the atom level, the diffusion model can capture 3D information about all atoms in a protein and distinguish any two distinct protein structures in nature.

Encoding. We adopt a graph neural networks model [28] to encode the representing at different levels by leveraging the message-passing mechanism. In many models dealing with the spatial positions of amino acids, SE(3)-equivariance is often leveraged to ensure the invariance of operations such as translation and rotation [2, 29]. We also introduce a novel method to ensure SE(3)-invariance by transforming each amino acid's coordinates ϕ. This step is crucial for facilitating the subsequent autoregressive decoding.

Given a protein chain, we first translate the coordinates such that the position of the first amino acid is moved to the origin, i.e., $(0, 0, 0)$. Then, we apply a rotation matrix to align the position of the second amino acid onto the positive x-axis:

$$R_1 = I + \sin(\theta)K + (1 - \cos(\theta))K^2, \tag{1}$$

where θ is the rotation angle between a node v and the x-axis, and K is the skew-symmetric matrix derived from the cross-product of v and the unit vector along the x-axis. The third amino acid is rotated around the x-axis to place it in the positive xy-plane:

$$R_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos(\psi) & -\sin(\psi) \\ 0 & \sin(\psi) & \cos(\psi) \end{pmatrix}, \tag{2}$$

where ψ is the angle that brings the third amino acid into the xy-plane. This process is iteratively applied to all amino acids in the protein chain.

In each iteration, the next amino acid is positioned relative to the previous ones, aligning the structure step by step while preserving the overall 3D conformation. The method ensures that all residues maintain SE(3)-invariance, making the transformations consistent across the entire protein chain.

The decoder at each level is an autoregressive Transformer [30] model that reconstructs the protein at each respective level. The autoregressive decoder can then use these transformed embeddings to reconstruct the information of a protein. At the sequence level, the decoder predicts the next amino acid token in the sequence. At the backbone level and the atom level, the decoder predicts geometric features (e.g., bond angles and distances) in an autoregressive fashion of each amino acid in the protein chain. Our method facilitates the use of SE(3)-invariant embeddings within an autoregressive framework. The decoder's autoregressive nature allows it to progressively predict amino acid positions by leveraging the SE(3)-invariant representation.

Proof of SE(3)- invariance of the Transformation. Let $\{\phi_i\}_{i=1}^n$ be the original coordinates of the amino acids in the protein chain. Consider an arbitrary rotation $R \in \mathrm{SO}(3)$ and translation $\Gamma \in \mathbb{R}^3$ applied to the protein, resulting in transformed coordinates:

$$\phi_i' = R\phi_i + \Gamma. \tag{3}$$

Our goal is to show that after applying the transformation method to both $\{\phi_i\}$ and $\{\phi_i'\}$, the resulting representations are identical.

Proof: For any transformation T in $\mathrm{SO}(3)$ and any vector $v \in \mathbb{R}^3$, we have:

$$T(v) = Rv. \tag{4}$$

Since rotations preserve vector norms, we can express $T(v)$ in terms of the norm of v and its unit vector $v' = v/\|v\|$:

$$T(v) = \|v\|Rv' = \|v\|T(v'). \tag{5}$$

This implies that the effect of T on v can be decomposed into scaling by $\|v\|$ and transforming its direction via rotation and translation. To simplify the expression and subsequent calculations, we denote all vectors ϕ_i as unit vectors (i.e., their norms are equal to 1).

Step 1: Translation to Origin Compute the relative positions with respect to the first amino acid:

$$\xi_i = \phi_i - \phi_1, \tag{6}$$

$$\xi_i' = \phi_i' - \phi_1' = (R\phi_i + \Gamma) - (R\phi_1 + \Gamma) = R(\phi_i - \phi_1) = R\xi_i. \tag{7}$$

Thus, we have $\xi_i' = R\xi_i$.

Step 2: Rotation to Align Second Amino Acid Along Positive x-Axis: since $\|\xi_2\| = \|\xi_2'\| = 1$, we have:

$$R_1\xi_2 = e_x, \tag{8}$$

$$R_1'\xi_2' = e_x, \tag{9}$$

where $e_x = [1, 0, 0]^\top$. Since $\xi_2' = R\xi_2$, we have:

$$R_1'R\xi_2 = e_x. \tag{10}$$

Let $R_1' = R_1R^{-1}$, then:

$$R_1'\phi_i' = R_1R^{-1}R\phi_i = R_1\phi_i. \tag{11}$$

Step 3: Rotation Around x-Axis to Place Third Amino Acid in xy-Plane. Find rotation matrices R_2 and R_2' (rotations around the x-axis) such that:

$$R_2R_1\phi_3 \in \mathrm{span}\{e_x, e_y\}, \tag{12}$$

$$R_2'R_1'\phi_3' \in \mathrm{span}\{e_x, e_y\}. \tag{13}$$

Since $R_1'd_3' = R_1d_3$, we have:

$$R_2'R_1\phi_3 = R_2R_1\phi_3. \tag{14}$$

Thus, $R_2' = R_2$. After applying the sequence of transformations, the final coordinates are:

$$\tilde{\phi}_i = R_2 R_1 \phi_i, \tag{15}$$

$$\tilde{\phi}_i' = R_2' R_1' d_i' = R_2 R_1 \phi_i = \tilde{\phi}_i. \tag{16}$$

Thus, $\tilde{\phi}_i' = \tilde{\phi}_i$, proving that the transformed coordinates are invariant under any initial rotation R and translation Γ. This confirms that the method achieves SE(3)-invariance.

Hierarchical Diffusion with Conditional Flow: To achieve control over the conditional generation of proteins at multiple levels, we employ a novel hierarchical diffusion model with a conditional flow mechanism. This design enables fine-grained manipulation of protein structure generation under specific conditions, such as targeted functional attributes. The diffusion process is split into three distinct levels: all-atom, backbone, and amino acid (sequence). Conditional information is injected from a lower level to ensure conditional consistency.

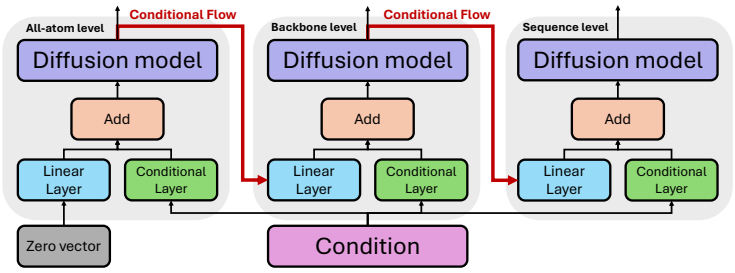

Figure 3: Consistency in the latent space.

Our conditional flow mechanism facilitates the transfer of information from lower levels (atom) to higher levels (backbone and amino acid) during the generation process. After denoising at each level, the latent representation is passed upward through a linear projection. Figure 3 shows the conditional flow (red lines). Specifically, for each level, the conditional flow integrates the latent vector from the lower level through a projection operation, which aligns the latent vector of the lower level to the higher level's embedding space via a learned linear transformation. This ensures that the structural information from the previous level is preserved and effectively conditions the next level's generation. The input at the atom level starts as a zero vector $z_t^0 = 0 \in \mathbb{R}^{L \times d}$. At the higher levels, the latent vector from the previous level, after removing noise, is linearly projected and combined with the current level's conditional embedding and time step embedding to ensure that the generative process is guided by both the condition and the structural information from the lower levels. The update at level i is given by:

Algorithm 1 Training Diffusion Models with Conditional Flow

while epoch $<$ epochs **do**
  Sample a random timestep t
  for all levels $i \in \{1, 2, 3\}$ **in parallel do**
    if $i = 1$ **then**
      Initialize zero vector z_t^0
    else
      Initialize z_0^{i-1} from ground truth data
    end if
    Sample noise vectors
    Diffuse latent vectors to get z_t^{i-1} and z_{t-1}^i
    Update latent vector:
      $z_t^i \leftarrow \epsilon^i(z_{t-1}^i; z_t^{i-1} W^i, c, \gamma_t)$
    Compute loss at ith level
    Update model parameters
  end for
  epoch$+ = 1$
end while

$$z_t^i = \epsilon^i(z_{t-1}^i; z_t^{i-1} W^i, c, \gamma_t), \tag{17}$$

where $z_t^i \in \mathbb{R}^{L \times d}$ is the latent vector at level i and time step t, z_t^{i-1} is the latent vector from the previous level, $W^i \in \mathbb{R}^{d \times d}$ is a learned linear projection matrix, c represents the conditional embedding (e.g., the protein's functional target), and γ_t is the time step embedding. Denote ϵ^i as the diffusion model at level i, which predicts the noise added during the forward process.

Training with Teacher Forcing: To enable efficient parallel training, we use the teacher forcing method during training. In this setup, for each level, the input z_0^{i-1} to the conditional flow is the ground truth data from the previous level, rather than the model's own generated output. This allows

us to decouple the training of the three levels, enabling them to be trained independently and in parallel. The training process for the diffusion model at each level follows the typical DDPM framework but with the conditional flow incorporated to introduce additional control over the generative process. The training procedure is outlined in Algorithm 1.

2.2 Evaluation of Conditional Consistency

Evaluating the quality and consistency of protein generation models requires a well-defined framework, particularly in the context of conditional generation. In this section, we define the theoretical basis for assessing conditional consistency in multi-class generation tasks and propose a novel framework to assess the suitability of different evaluations.

Denote $\{C^1, C^2, \ldots, C^K\}$ as a set of target classes, where each class C^k corresponds to an independent and mutually exclusive category (e.g., different protein functions or classes). Let x represent a data sample, and $d(x, C_k)$ be a conditional consistency metric that measures the consistency between a sample x and the target class C_k. Given a model exhibiting strong conditional consistency, it should generate samples such that as we progress through a sequence of generated samples $\{x^i | i = 0, 1, 2, \cdots, \infty\}$ ordered by increasing quality (i.e., this sequence is assumed to exist with each sample x^i becoming more consistent with the target class C_k as i increases), the consistency distance between each sample and samples in C_k should decrease. Mathematically, a good evaluation metric d satisfies:

$$\lim_{i \to \infty} d(x^i, C^k) \to 0. \tag{18}$$

It implies that as the sample quality improves, the consistency to the correct target class decreases asymptotically towards zero. We can further derive the following theorem.

Theorem: $\exists N \in \mathbb{N}^+, \forall i > N,\ d(x^i, C^k) < \min_{j \neq k} d(x^i, C^j)$ where C^j is any other class.

Proof: see Appendix C.2.

Given that test samples exhibit strong conditional consistency, the theorem suggests that if we measure $d(\cdot)$ between test samples and all target classes, the majority will be classified into the correct target class C^k. However, relying solely on spatial distance may be too rigid for general evaluation, especially in conditional settings. In Figure 4, the green points represent generated samples, and darker shades indicate better sample quality. A well-defined metric should indicate that the green points are closer to their correct target class (i.e., Class 2) rather than the blue or pink classes.

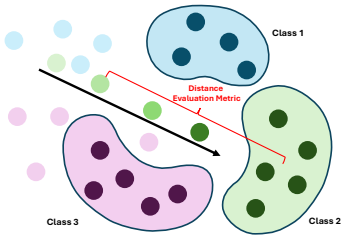

Figure 4: Consistency in the latent space

Besides the accuracy (which class the generated belongs to), Mean Reciprocal Rank (MRR) and Normalized Mean Rank (NMR) are widely used to assess how well the evaluation metric ranks generated samples based on their correct target classes. Specifically:

$$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i}, \text{NMR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{\text{rank}_i - 1}{N - 1}. \tag{19}$$

where Q is the set of test queries, and rank_i is the rank of the correct target class for the i-th test query. These metrics, in combination with accuracy, provide a more comprehensive evaluation framework for assessing conditional consistency evaluation metrics in generative models.

In this paper, we propose *Protein-MMD*, a new evaluation metric that calculates the Maximum Mean Discrepancy (MMD) based on protein embeddings. Specifically, both real and generated protein sequences are encoded using the ESM2 language model [6], which provides biologically informed embeddings. ESM2 was chosen due to its ability to capture both structural and functional properties of proteins, thanks to its pretraining on a large protein corpus. This makes ESM2 particularly effective for evaluating distributional and functional similarities between real and generated proteins, aligning with the goals of *de novo* protein design:

$$\text{Protein-MMD}(p_r, p_g) = \left\| \frac{1}{n} \sum_{i=1}^{n} \varphi(x_i) - \frac{1}{m} \sum_{j=1}^{m} \varphi(y_j) \right\|_{\mathcal{H}}^2, \tag{20}$$

where $\varphi(\cdot)$ denotes the embeddings extracted from the language model. These embeddings represent both sequence and functional information, making them particularly well-suited for comparing real and generated protein distributions.

To validate the effectiveness of Protein-MMD and other metrics, we apply the evaluation metrics on the Enzyme Commission (EC) dataset, which categorizes proteins based on the reactions they catalyze using EC numbers. We focus on seven classes from the first EC number category for our conditional generation task. In Table 1, we compare three evaluation metrics: MMD (considering only sequence statistics as presented in [31]), Protein-FID (using ESM2 in place of Inception for protein generation), and Protein-MMD. All metrics are used to compute the Accuracy, Mean Reciprocal Rank (MRR), and Normalized Mean Rank (NMR) scores to evaluate how it performs in evaluating the conditional consistency. As observed, *Protein-MMD* outperforms both MMD and Protein-FID across all evaluation metrics. The higher accuracy and MRR scores indicate that *Protein-MMD* better captures the conditional consistency of the proteins in the test set. The lower NMR score further demonstrates that *Protein-MMD* ranks the correct target class higher in comparison to other metrics, validating its effectiveness in conditional protein generation tasks. While *Protein-MMD* proves to be the most effective metric according to our framework, we acknowledge the widespread use of FID in generative modeling tasks. Therefore, we will continue to report Protein-FID results alongside Protein-MMD in subsequent experiments.

Table 1: Evaluation on the EC dataset.

Metric	**Accuracy** ↑	**MRR** ↑	**NMR** ↓
MMD	0.0687	0.3101	0.5506
Protein-FID	0.2988	0.4825	0.3920
Protein-MMD	**0.5487**	**0.6629**	**0.2524**

3 Experiments

3.1 Experimental Setup

We compared our model against several baselines, each representing distinct approaches to protein generation. ProteoGAN [31] is a GAN-based method, while ESM2 [6] and ProstT5 [32] are Transformer-based language models specifically designed for protein sequence modeling. ProteinMPNN [11] and LatentDiff [10], on the other hand, are graph-based models, with LatentDiff also incorporating a diffusion-based framework, specifically using a latent diffusion approach. For each model, we evaluate the performance using both diversity metrics (TM-score, RMSD, and Seq.ID) and conditional consistency metrics (Protein-MMD and Protein-FID). Higher RMSD, lower TM-score, and lower Seq.ID indicate higher diversity, while lower Protein-MMD and Protein-FID values signify higher conditional consistency between the generated and real protein distributions. More detailed settings can be found in Appendix C.1.

Table 2: Results on EC and GO datasets.

	EC Dataset					GO Dataset				
	Diversity			Conditional Consistency		Diversity			Conditional Consistency	
	TM-score↓	RMSD↑	Seq.ID↓	Protein-MMD↓	Protein-FID↓	TM-score↓	RMSD↑	Seq.ID↓	Protein-MMD↓	Protein-FID↓
ProteGAN	0.26	<u>5.35</u>	6.71	13.99	260.31	0.23	5.96	**6.33**	**10.89**	**256.31**
ESM2	0.29	4.25	**6.57**	<u>13.35</u>	<u>238.46</u>	<u>0.22</u>	**7.33**	<u>6.39</u>	11.86	290.31
ProstT5	0.28	4.25	<u>6.61</u>	13.76	248.32	0.26	6.81	6.73	11.93	292.58
ProteinMPNN	**0.24**	4.24	67.43	22.31	587.72	**0.14**	<u>7.10</u>	77.96	15.94	410.43
LatentDiff	0.37	2.73	7.67	13.43	256.75	0.31	4.26	7.37	12.66	346.40
Ours(128)	**0.24**	4.7	7.56	13.74	250.2			—		
Ours(256)	0.27	4.40	6.88	13.67	248.31			—		
Ours(512)	<u>0.25</u>	**5.39**	6.79	**13.28**	**237.46**	0.26	6.09	7.13	<u>11.67</u>	<u>284.65</u>

The best performance for each metric is indicated in **bold**, while the second-best performance is <u>underlined</u>.

3.2 Results and Analysis

Table 2 presents the results of our model and the baselines on two datasets. Our model achieves the best performance in terms of most metrics on the EC dataset, indicating superior conditional consistency and diversity in generating proteins that adhere closely to the specified enzyme classes. On the EC dataset, our model (with sequence length 512) achieves the lowest Protein-MMD and Protein-FID scores, demonstrating effective modeling of the distributional and functional similarities between generated and real proteins. The RMSD and TM-score metrics indicate that our model generates structurally diverse proteins, with the highest RMSD and among the 2nd-lowest TM-scores, suggesting less topological similarity to templates. The sequence identity (Seq.ID) is also low, indicating higher sequence diversity. For the GO dataset, our model also performs competitively. However, in terms of conditional consistency metrics (Protein-MMD and Protein-FID), our model

ranks second, with ESM2 achieving the best Protein-MMD score and ProteoGAN achieving the best Protein-FID score. This suggests that our model generates diverse protein structures.

Ablation Study. To investigate the effectiveness of each of the three levels (amino acid, backbone, all-atom), we conducted an ablation study in the experiment. Specifically, the variant of our model removes either a specific level (the backbone or all-atom) or both two levels. Then we examine the performance of the conditional consistency metrics. Note that we can not remove the amino acid level because the amino acid is required for evaluation. The ablation study is conducted on the EC dataset. As shown in Table 3, if we remove any level (i.e., backbone and all-atom level) or both two levels, the performance will drop. It verifies the necessity of our multi-level conditional diffusion.

Table 3: Ablation study.

Method	Protein-MMD↓	Protein-FID↓
All	**13.50**	**241.82**
Removed backbone level	13.73	249.14
Removed all-atom level	13.94	251.83
Removed both	14.06	255.15

Impact of Maximum Sequence Length. In previous studies on protein *de novo* design, existing works usually employ a maximum sequence length of 128 [10]. However, through our experiments, we observed that for conditional generation tasks, shorter sequence lengths fail to fully leverage the conditional information, which in turn results in lower conditional consistency metrics. To address this, we constructed models with three different maximum sequence lengths: 128, 256, and 512, and investigated the impact of maximum length on the model's ability to maintain conditional consistency.

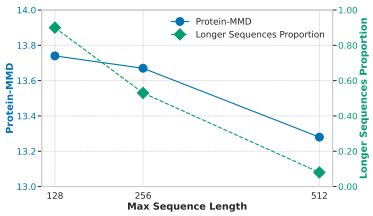

Figure 5: Consistency.

As shown in Figure 5, we observe a positive correlation between the Protein-MMD metric, which reflects conditional consistency, and the proportion of training data samples exceeding the current maximum sequence length. This indicates that longer sequences help the model better incorporate condition information during generation. Moreover, the results in Table 2 for our method with different lengths reveal that the maximum sequence length does not influence the model's performance on diversity metrics, which are independent of the quality of condition-guided generation. These findings underscore the importance of maximum sequence length in enhancing conditional consistency, offering valuable insights for the design of future protein conditional generation models.

Table 4: Comparison with ProteoGAN.

Method	IoU_{mean}↑	IoU_{max}↑
ProteoGAN	**0.2181**	0.4706
Ours (512)	0.2088	**0.5833**

Case Study. To further demonstrate the superiority of our model on the GO dataset, particularly regarding conditional consistency, we conducted a fine-grained case study comparing our method with the best baseline ProteoGAN. We utilized an *in-silico* evaluation to perform a fine-grained analysis of the generated protein sequences. By employing a trained ESM-MLP classifier on the GO dataset, we assessed each generated protein's adherence to the specified GO terms using the Intersection over Union (IoU) [33]. As shown in Table 4, our method exhibits a lower average IoU_{mean} compared to ProteoGAN, aligning with earlier results in Table 2. However, it achieves a higher IoU_{max}, indicating a greater potential for generating high-quality samples that closely match the desired GO annotations. Figure 6 illustrates the distribution of IoU scores. While ProteoGAN's samples are concentrated around medium quality, our method generates a broader range of samples, including those with higher IoU scores. This suggests that our model, despite a lower average performance, is more capable of producing proteins with superior conditional consistency.

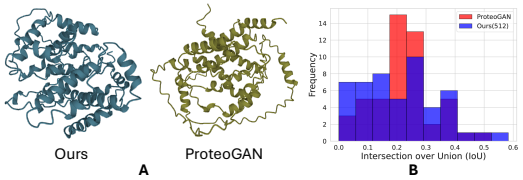

Figure 6: **A** shows the two highest generated protein results of Ours and ProteoGAN in terms of the IoU indicator. **B** shows the statistical frequency histogram.

References

[1] J. L. Watson, D. Juergens, N. R. Bennett, B. L. Trippe, J. Yim, H. E. Eisenach, W. Ahern, A. J. Borst, R. J. Ragotte, L. F. Milles *et al.*, "De novo design of protein structure and function with rfdiffusion," *Nature*, vol. 620, no. 7976, pp. 1089–1100, 2023.

[2] A. J. Bose, T. Akhound-Sadegh, G. Huguet, K. Fatras, J. Rector-Brooks, C. Liu, A. C. Nica, M. Korablyov, M. M. Bronstein, and A. Tong, "Se(3)-stochastic flow matching for protein backbone generation," vol. 1. MIT Press, 2016.

[3] P.-S. Huang, S. E. Boyken, and D. Baker, "The coming of age of de novo protein design," *Nature*, vol. 537, no. 7620, pp. 320–327, 2016.

[4] S. Feng, M. Li, Y. Jia, W. Ma, and Y. Lan, "Protein-ligand binding representation learning from fine-grained interactions," in *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

[5] Z. Huang, L. Yang, X. Zhou, Z. Zhang, W. Zhang, X. Zheng, J. Chen, Y. Wang, B. Cui, and W. Yang, "Protein-ligand interaction prior for binding-aware 3d molecule diffusion models," in *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

[6] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli *et al.*, "Evolutionary-scale prediction of atomic-level protein structure with a language model," *Science*, vol. 379, no. 6637, pp. 1123–1130, 2023.

[7] K. A. Dill, S. B. Ozkan, M. S. Shell, and T. R. Weikl, "The protein folding problem," *Annu. Rev. Biophys.*, vol. 37, no. 1, pp. 289–316, 2008.

[8] F. H. Arnold, "Design by directed evolution," *Accounts of chemical research*, vol. 31, no. 3, pp. 125–131, 1998.

[9] N. Anand and T. Achim, "Protein structure and sequence generation with equivariant denoising diffusion probabilistic models," *arXiv preprint arXiv:2205.15019*, 2022.

[10] C. Fu, K. Yan, L. Wang, W. Y. Au, M. C. McThrow, T. Komikado, K. Maruhashi, K. Uchino, X. Qian, and S. Ji, "A latent diffusion model for protein structure generation," in *Learning on Graphs Conference*. PMLR, 2024, pp. 29–1.

[11] J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. Wicky, A. Courbet, R. J. de Haas, N. Bethel *et al.*, "Robust deep learning–based protein sequence design using proteinmpnn," *Science*, vol. 378, no. 6615, pp. 49–56, 2022.

[12] B. L. Trippe, J. Yim, D. Tischer, D. Baker, T. Broderick, R. Barzilay, and T. S. Jaakkola, "Diffusion probabilistic modeling of protein backbones in 3d for the motif-scaffolding problem," in *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[13] U. J. Komorowska, S. V. Mathis, K. Didi, F. Vargas, P. Lio, and M. Jamnik, "Dynamics-informed protein design with structure conditioning," in *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. [Online]. Available: https://openreview.net/forum?id=jZPqf2G9Sw

[14] L. Klarner, T. G. J. Rudner, G. M. Morris, C. M. Deane, and Y. W. Teh, "Context-guided diffusion for out-of-distribution molecular and protein design," in *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. [Online]. Available: https://openreview.net/forum?id=8NfHmzo0Op

[15] N. Gruver, S. Stanton, N. C. Frey, T. G. J. Rudner, I. Hötzel, J. Lafrance-Vanasse, A. Rajpal, K. Cho, and A. G. Wilson, "Protein design with guided discrete diffusion," in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., 2023.

[16] B. L. Trippe, J. Yim, D. Tischer, D. Baker, T. Broderick, R. Barzilay, and T. Jaakkola, "Diffusion probabilistic modeling of protein backbones in 3d for the motif-scaffolding problem," *arXiv preprint arXiv:2206.04119*, 2022.

[17] Y. Hu, Y. Tan, A. Han, L. Zheng, L. Hong, and B. Zhou, "Secondary structure-guided novel protein sequence generation with latent graph diffusion," *arXiv preprint arXiv:2407.07443*, 2024.

[18] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 723–773, 2012.

[19] M. van Kempen, S. S. Kim, C. Tumescheit, M. Mirdita, C. L. Gilchrist, J. Söding, and M. Steinegger, "Foldseek: fast and accurate protein structure search," *Biorxiv*, pp. 2022–02, 2022.

[20] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The protein data bank," *Nucleic acids research*, vol. 28, no. 1, pp. 235–242, 2000.

[21] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko *et al.*, "Highly accurate protein structure prediction with alphafold," *nature*, vol. 596, no. 7873, pp. 583–589, 2021.

[22] Y. Zhang and J. Skolnick, "Tm-align: a protein structure alignment algorithm based on the tm-score," *Nucleic acids research*, vol. 33, no. 7, pp. 2302–2309, 2005.

[23] L. Wang, H. Liu, Y. Liu, J. Kurtin, and S. Ji, "Learning hierarchical protein representations via complete 3d graph networks," *arXiv preprint arXiv:2207.12600*, 2022.

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

[25] Z. Zhang, M. Xu, A. R. Jamasb, V. Chenthamarakshan, A. C. Lozano, P. Das, and J. Tang, "Protein representation learning by geometric structure pretraining," in *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. [Online]. Available: https://openreview.net/forum?id=to3qCB3tOh9

[26] P. Hermosilla, M. Schäfer, M. Lang, G. Fackelmann, P. Vázquez, B. Kozlíková, M. Krone, T. Ritschel, and T. Ropinski, "Intrinsic-extrinsic convolution and pooling for learning on 3d protein structures," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. [Online]. Available: https://openreview.net/forum?id=l0mSUROpwY

[27] B. Jing, S. Eismann, P. N. Soni, and R. O. Dror, "Equivariant graph neural networks for 3d macromolecular structure," *CoRR*, vol. abs/2106.03843, 2021. [Online]. Available: https://arxiv.org/abs/2106.03843

[28] L. Wang, Y. Liu, Y. Lin, H. Liu, and S. Ji, "Comenet: Towards complete and efficient message passing for 3d molecular graphs," in *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., 2022.

[29] J. Yim, B. L. Trippe, V. D. Bortoli, E. Mathieu, A. Doucet, R. Barzilay, and T. S. Jaakkola, "SE(3) diffusion model with application to protein backbone generation," in *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 2023, pp. 40 001–40 039. [Online]. Available: https://proceedings.mlr.press/v202/yim23a.html

[30] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[31] T. Kucera, M. Togninalli, and L. Meng-Papaxanthos, "Conditional generative modeling for de novo protein design with hierarchical functions," *Bioinformatics*, vol. 38, no. 13, pp. 3454–3461, 2022.

[32] M. Heinzinger, K. Weissenow, J. G. Sanchez, A. Henkel, M. Steinegger, and B. Rost, "Prostt5: Bilingual language model for protein sequence and structure. biorxiv," *bioRxiv preprint*, 2023.

[33] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. D. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 658–666.

[34] N. C. Frey, D. Berenberg, K. Zadorozhny, J. Kleinhenz, J. Lafrance-Vanasse, I. Hötzel, Y. Wu, S. Ra, R. Bonneau, K. Cho, A. Loukas, V. Gligorijevic, and S. Saremi, "Protein discovery with discrete walk-jump sampling," in *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. [Online]. Available: https://openreview.net/forum?id=zMPHKOmQNb

[35] W. Mao, M. Zhu, Z. Sun, S. Shen, L. Y. Wu, H. Chen, and C. Shen, "De novo protein design using geometric vector field networks," in *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. [Online]. Available: https://openreview.net/forum?id=9UIGyJJpay

[36] Z. Gao, X. Sun, Z. Liu, Y. Li, H. Cheng, and J. Li, "Protein multimer structure prediction via prompt learning," in *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. [Online]. Available: https://openreview.net/forum?id=OHpvivXrQr

[37] B. Gao, B. Qiang, H. Tan, Y. Jia, M. Ren, M. Lu, J. Liu, W. Ma, and Y. Lan, "Drugclip: Contrasive protein-molecule representation learning for virtual screening," in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., 2023.

[38] W. Lu, Q. Wu, J. Zhang, J. Rao, C. Li, and S. Zheng, "Tankbind: Trigonometry-aware neural networks for drug-protein binding structure prediction," in *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., 2022.

[39] Q. Pei, K. Gao, L. Wu, J. Zhu, Y. Xia, S. Xie, T. Qin, K. He, T. Liu, and R. Yan, "Fabind: Fast and accurate protein-ligand binding," in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., 2023.

[40] M. Liu, Y. Luo, K. Uchino, K. Maruhashi, and S. Ji, "Generating 3d molecules for target protein binding," in *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 2022, pp. 13 912–13 924. [Online]. Available: https://proceedings.mlr.press/v162/liu22m.html

[41] H. Zhou, Y. Fu, Z. Zhang, C. Bian, and Y. Yu, "Protein representation learning via knowledge enhanced primary structure reasoning," in *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. [Online]. Available: https://openreview.net/forum?id=VbCMhg7MRmj

[42] C. Gong, A. R. Klivans, J. Loy, T. Chen, Q. Liu, and D. J. Diaz, "Evolution-inspired loss functions for protein representation learning," in *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. [Online]. Available: https://openreview.net/forum?id=y5L8W0KRUX

[43] T. Chen, C. Gong, D. J. Diaz, X. Chen, J. T. Wells, Q. Liu, Z. Wang, A. D. Ellington, A. Dimakis, and A. R. Klivans, "Hotprotein: A novel framework for protein thermostability prediction and editing," in *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. [Online]. Available: https://openreview.net/forum?id=YDJRFWBMNby

[44] L. S. Moreta, O. Rønning, A. S. Al-Sibahi, J. Hein, D. L. Theobald, and T. Hamelryck, "Ancestral protein sequence reconstruction using a tree-structured ornstein-uhlenbeck variational autoencoder," in *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. [Online]. Available: https://openreview.net/forum?id=FZoZ7a31GCW

[45] Y. Lee, H. Yu, J. Lee, and J. Kim, "Pre-training sequence, structure, and surface features for comprehensive protein representation learning," in *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. [Online]. Available: https://openreview.net/forum?id=BEH4mGo7zP

[46] H. Fan, Z. Wang, Y. Yang, and M. S. Kankanhalli, "Continuous-discrete convolution for geometry-sequence modeling in proteins," in *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. [Online]. Available: https://openreview.net/forum?id=P5Z-Zl9XJ7

[47] J. Ingraham, V. K. Garg, R. Barzilay, and T. S. Jaakkola, "Generative models for graph-based protein design," in *Deep Generative Models for Highly Structured Data, ICLR 2019 Workshop, New Orleans, Louisiana, United States, May 6, 2019*. OpenReview.net, 2019. [Online]. Available: https://openreview.net/forum?id=SJgxrLLKOE

[48] Y. Huang, S. Li, L. Wu, J. Su, H. Lin, O. Zhang, Z. Liu, Z. Gao, J. Zheng, and S. Z. Li, "Protein 3d graph structure learning for robust structure-based protein property prediction," in *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, M. J. Wooldridge, J. G. Dy, and S. Natarajan, Eds. AAAI Press, 2024, pp. 12 662–12 670. [Online]. Available: https://doi.org/10.1609/aaai.v38i11.29161

[49] S. Aykent and T. Xia, "Gbpnet: Universal geometric representation learning on protein structures," in *KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 14 - 18, 2022*, A. Zhang and H. Rangwala, Eds. ACM, 2022, pp. 4–14. [Online]. Available: https://doi.org/10.1145/3534678.3539441

[50] T. Xia and W. Ku, "Geometric graph representation learning on protein structure prediction," in *KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021*, F. Zhu, B. C. Ooi, and C. Miao, Eds. ACM, 2021, pp. 1873–1883. [Online]. Available: https://doi.org/10.1145/3447548.3467323

[51] M. Hladis, M. Lalis, S. Fiorucci, and J. Topin, "Matching receptor to odorant with protein language and graph neural networks," in *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. [Online]. Available: https://openreview.net/forum?id=q9VherQJd8_

[52] J. Meier, R. Rao, R. Verkuil, J. Liu, T. Sercu, and A. Rives, "Language models enable zero-shot prediction of the effects of mutations on protein function," in *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021, pp. 29 287–29 303.

[53] N. Anand and P. Huang, "Generative modeling for protein structures," in *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., 2018, pp. 7505–7516. [Online]. Available: https://proceedings.neurips.cc/paper/2018/hash/afa299a4d1d8c52e75dd8a24c3ce534f-Abstract.html

[54] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=St1giarCHLP

[55] C. Jiang, B. Hui, B. Liu, and D. Yan, "Successfully applying lottery ticket hypothesis to diffusion model," *arXiv preprint arXiv:2310.18823*, 2023.

[56] S. Gao, B. Hui, and W. Li, "Image generation of egyptian hieroglyphs," in *Proceedings of the 2024 16th International Conference on Machine Learning and Computing*, 2024, pp. 389–397.

[57] Y. Xiao, G. Li, K. Deng, Y. Wu, Z. Zhan, Y. Wang, X. Ma, and B. Hui, "Lightcache: Memory-efficient, training-free acceleration for video generation," *arXiv preprint arXiv:2510.05367*, 2025.

[58] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4195–4205.

A Related Works

De novo protein design methods are dedicated to identify novel proteins with the desired structure and function properties [1, 3, 34, 35, 13, 36]. Recent advancements in machine learning have enabled a generative model to accelerate key steps in the discovery of novel molecular structures and drug design [37–40]. A prior step of generate models to the representation of proteins [25, 41, 42]. The majority of representation learning for protein is to represent a protein as a sequence of amino acids [43–46]. Considering the spatial information is important to the property of a protein, many works resort to a graph model for a comprehensive presentation with the structure information [47, 48]. In general, each node on the graph is an amino acid and the edge is decided by the distance between two nodes [49–51]. Despite the power of graph models, the relation information in a 3-dimensional space captures the multi-level structure such as the angle between two edges. A line of research works explore the protein structure in 3D space [26, 5]. Recently, large language models (LLMs) have also been introduced to model the sequence [52, 6] inspired by the success of natural language processing.

Capitalizing on the power of generative models such as Generative Adversarial Networks (GANs) and diffusion models, deep generative modeling has shown its potential for fast generation of new and viable protein structures. [53] has applied GANs to the task of generating protein structures by encoding protein structures in terms of pairwise distances on the protein backbone. Diffusion models [54–57] have emerged as a powerful tool for graph-structured diffusion processes [14]. FrameDiff has been proposed for monomer backbone generation and it can generate designable monomers up to 500 amino acids [29]. NOS is another diffusion model that generates protein sequences with high likelihood by taking many alternating steps in the continuous latent space of the model [15].

B Conclusions

In this paper, we introduce a novel multi-level conditional generative diffusion model that integrates sequence-based and structure-based information for efficient end-to-end protein design. Our model incorporates a 3D rotation-invariant preprocessing step to maintain SE(3)-invariance. To address the limitation of the existing evaluations, we propose a novel metric to evaluate conditional consistency.

C Appendices

C.1 Experiment settings

To verify the effectiveness of our proposed multi-level conditional diffusion model, we conducted comprehensive experiments on two standard datasets: the Enzyme Commission (EC) dataset and the Gene Ontology (GO) dataset. The EC dataset categorizes proteins based on the biochemical reactions they catalyze, while the GO dataset classifies proteins according to their associated biological

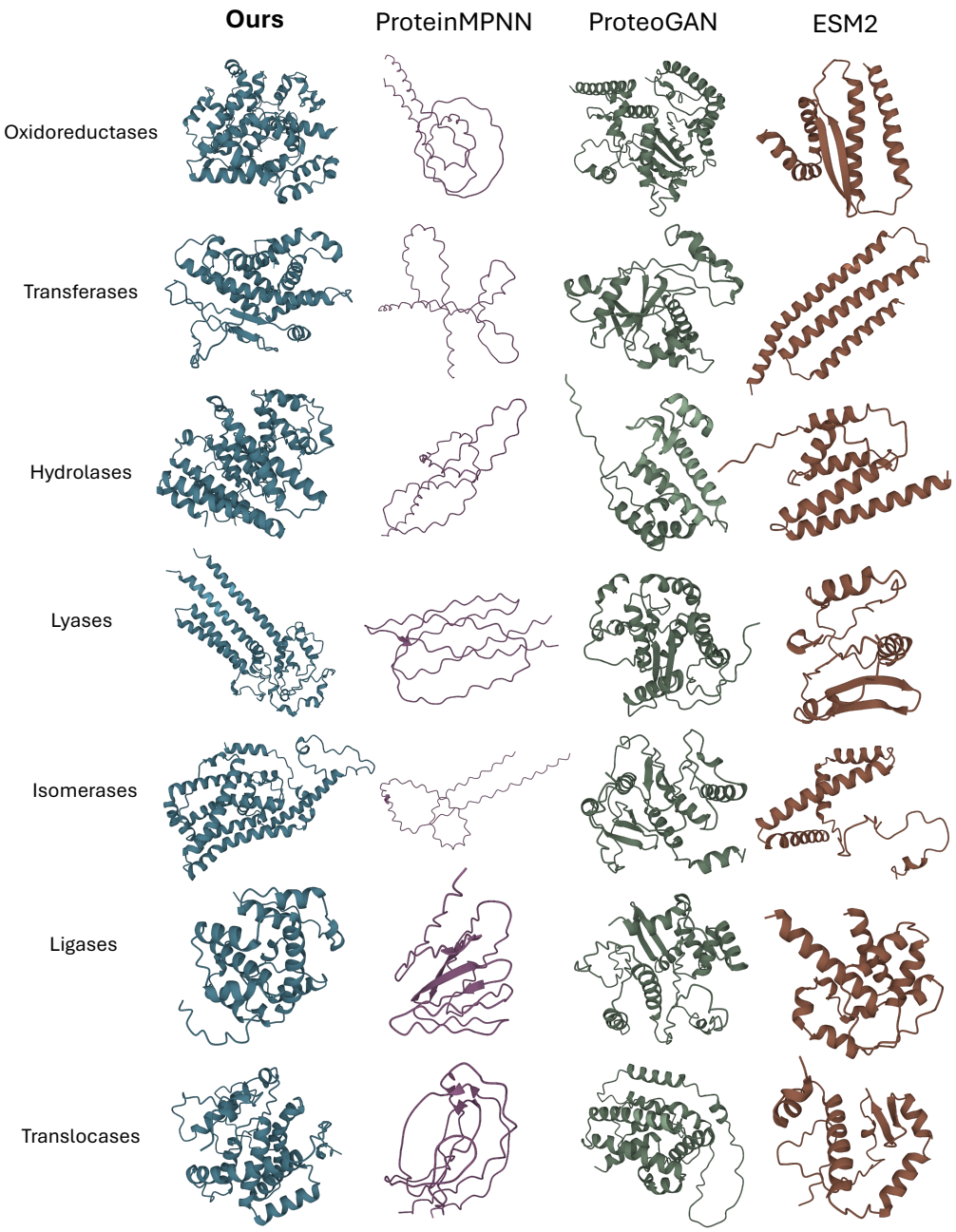

Figure 7: Protein Visualization Comparison on EC Dataset (Ours vs. ProteinMPNN, ProteoGAN, and ESM2).

processes, cellular components, and molecular functions. These datasets provide a robust benchmark for assessing both the diversity and conditional consistency of generated protein sequences.

Our model leverages the `esm2_t33_650M_UR50D` model from ESM2 [6] as the amino acid-level encoder. To construct the Protein Variational Auto-encoder model, we set the latent dimension to 384, and the decoder is composed of 8 Transformer [30] decoder blocks, each equipped with an 8-head self-attention mechanism. The Protein Variational Auto-encoder model is trained with a learning rate of 10^{-4}, using a combination of mean squared error (MSE) and cross-entropy as the loss functions. To regulate the latent vector distribution, we apply a KL divergence loss with a weight of 10^{-5}. We experimented with 128, 256, and 512 as the maximum sequence lengths. For the diffusion model, we

modify the DiT-B architecture from DiT [58], which consists of 12 DiT blocks and uses a hidden size of 768. The DiT model is trained from scratch with a learning rate of 10^{-4} and includes a weight decay of 10^{-5}.

C.2 Protein-MMD

Theorem: $\exists N \in \mathbb{N}^+, \forall i > N$, $d(x^i, C^k) < \min_{j \neq k} d(x^i, C^j)$ where C^j is any other class.

Proof: Assume that there exists a class $C^j (j \neq k)$ such that $d(x^i, C^j) \leq d(x^i, C^i)$ for $i > N$. Since C^k is defined as the correct target class and the quality of the generated sample improves with $i \to \infty$, the consistency $d(x_n^i, C^k)$ should approach zero. If $d(x^i, C^j) \leq d(x^i, C^i)$, we have $\lim_{n \to \infty} d(x^i, C^j) = 0$. It contradicts the assumption that C^k is the correct class for the generated data. Therefore, the assumption is false.

Visualization. In Figure 7 (Appendix), we present visualizations of proteins conditionally generated by our method and other baselines on the EC dataset. Specifically, we generate proteins with 7 different functions (e.g., Oxidoreductases). Compared with the baselines, our method can generate discriminative proteins given the same input. By modeling the hierarchical relation at different levels, our method can generate foldable and functional sequences in 3D space.