Proteus: pioneering protein structure generation for enhanced designability and efficiency

Chentong Wang [1 2] **Yannan Qu** [2] **Zhangzhi Peng** [3] **Yukai Wang** [2] **Hongli Zhu** [2] **Dachuan Chen** [2] **Longxing Cao** [2]

Abstract

The development of *de novo* protein design method is crucial for widespread applications in biology and chemistry. Protein backbone diffusion aims to generate designable protein structures with high efficiency. Although there have been great strides in protein structure prediction, applying these methods to protein diffusion has been challenging and inefficient. We introduce Proteus, an innovative approach that uses graph-based triangle methods and a multi-track interaction network. Proteus demonstrated cutting-edge designability and efficiency in computational evaluations. We tested the reliability of the model by experimental characterization. Our analysis indicates that from both computational and experimental perspectives, it is able to generate proteins with a remarkably high success rate. We believe Proteus's capacity to rapidly create highly designable protein backbone without the need for pre-training techniques will greatly enhance our understanding of protein structure diffusion and contribute to advances in protein design. The source code of Proteus will be released once accepted.

1. Introduction

The biological function of a protein is often directly determined by its tertiary structure, which makes it attractive to design protein backbones directly. De novo protein design methods focus on designing proteins with the desired structure or function. With the recent advancement of protein structure prediction models, hallucination(Anishchenko et al., 2021; Wang et al., 2022) directly exploits these models for the protein generation task. By leveraging the generative ability of the diffusion model(Ho et al., 2020), RFdiffusion(Watson et al., 2023) has shown superior performance on a wide range of protein design challenges, including protein binder design, motif scaffolding, and symmetric design. Though RFdiffusion shows promising results, it needs to be pretrained from Rosettafold2(Baek et al., 2023), which makes it very uninterpretable and impossible to modify the model architecture for performance optimization.

To tackle this challenge, researchers have proposed various diffusion strategies and model architectures to generate designable protein backbones without pretraining. An approach is FoldingDiff(Wu et al., 2022), which performs diffusion using internal coordinates with a bidirectional transformer that iteratively denoises a sequence of torsion angles. Though FoldingDiff yields protein-like backbones, the majority of generated structures are predicted to not be designable. Two concurrent works directly diffusion on residue coordinates(Lin & AlQuraishi, 2023) or tangent space of coordinate and rotation(Yim et al., 2023), shows more promising backbone designability. Previous works(Anand & Achim, 2022; Lee et al., 2023; Trippe et al., 2022) explored multiple model architectures, including U-Net(Ronneberger et al., 2015), Equivariant Graph Neural Networks (EGNNs)(Satorras et al., 2022), Invariant Point Attention(IPA)(Jumper et al., 2021), which succeed in Computer Vision or dynamic system modeling.

While these efforts have relatively improved the designability of existing protein structure diffusion models that do not rely on pretraining, there remains a significant gap compared to RFdiffusion. This disparity results in limitations for these models, making them either ineffective or challenging to apply in real-world protein design tasks. RFdiffusion, with its advanced capabilities, sets a higher benchmark in the field, highlighting the need for further advancements in model development to bridge this gap.

To bridge the performance gap between unpretrained and pretrained models, we developed Proteus. Proteus surpasses the designability of RFdiffusion by utilizing a graph-based triangle technique and a multi-track interaction network with great enhancement of the dataset. Also our model sets a new standard in efficiency, driven by two main factors: reduc-

[1]Zhejiang University, Hangzhou, Zhejiang, China [2]School of Life Sciences, Westlake University, Hangzhou, Zhejiang, China [3]Duke University, Durham, North Carolina, USA. Correspondence to: Longxing Cao <caolongxing@westlake.edu.cn>.

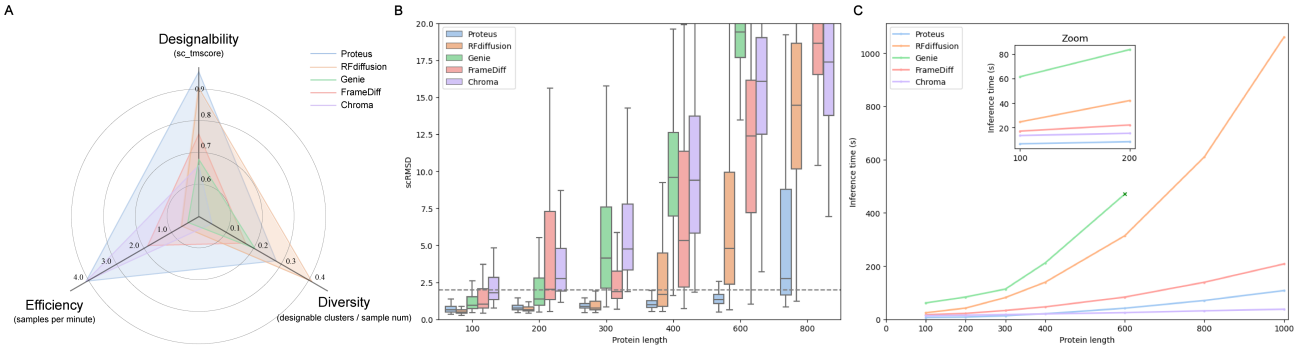

Figure 1: Benchmark illustration of Proteus and other backbone diffusion models on designability, efficiency, and diversity. All metrics are averaged from 200 backbones of each length 100,200,300,400,600,800. 8 sequences of each backbone are designed by ProteinMPNN except for Chroma, which uses ChromaDesign. (A) The radar plot of model evaluation on three dimensions. (B) Self-consistency RMSD between the generated backbone and best prediction of ESMfold. (C) Sample time of backbone generation, evaluated on A40. Genie fails to generate backbones larger than 600 residues due to being out of memory.

ing necessary sampling steps by enhanced representational capacity and graph modeling, which lowers computational complexity. Consequently, Proteus achieves protein generation speeds on par with Chroma(Ingraham et al., 2022).

In summary, our model accomplishes multiple optimization goals within the realm of protein design: it operates without the need for pretraining, achieves high designability, and maintains rapid generation speeds. The introduction of Proteus represents a significant advancement in protein design methodology, offering a solution that balances efficiency with the complexity of protein structure generation.

2. Preliminaries

Protein backbone representation Following the approach of AlphaFold2(Jumper et al., 2021), the backbone of each residue is parameterized as a series of rigid transformations, also known as frames. These frames, denoted by T = (R, t), $R \in \mathbb{R}^{3\times3}$, are defined within the special Euclidean group $SE(3)$ and represent orientation-preserving transformations. Each frame T_n maps fixed, chemically idealized main atom coordinates $[N_n, C_n, C_{\alpha_n}]$ of the residue. Here, $R \in SO(3)$ is a rotation matrix derived from the Gram-Schmidt process applied to the three backbone atoms N, CA, and C, and $t \in \mathbb{R}^3$ represents the translation corresponding to the coordinates of atom C_α.

Diffusion modeling on protein backbone Though multiple procedures exist on protein backbone diffusion, including diffusion on inter-residue geometry or torsion angles. We describe here the most promising approach - SE(3) diffusion, which processes each residue independently, and the corresponding probability estimation on SO(3) for score matching calculation proposed by Yim et al..

Briefly, the protein backbone's forward diffusion process is driven by Brownian motion on SO(3) and R(3) individually as shown in Equation (1)

$$d\mathbf{T}^{(t)} = \left[\mathbf{0}, -\frac{1}{2}\mathbf{X}^{(t)}\right] dt + \left[d\mathbf{B}^{(t)}_{\mathrm{SO}(3)}, d\mathbf{B}^{(t)}_{\mathbb{R}^3}\right]. \quad (1)$$

By defining probability estimation on R(3) forward diffusion, $p_{t|0}(x^{(t)}|x^{(0)}) = \mathcal{N}(x^{(t)}; e^{-t/2}x^{(0)}, (1-e^{-t})Id_3)$, the corresponding conditional score can be computed explicitly

$$\nabla \log p_{t|0}(x^{(t)}|x^{(0)}) = (1-e^{-t})^{-1}\left(e^{-t/2}x^{(0)} - x^{(t)}\right).$$

For the proper estimation of probability on SO(3), Brownian motion on SO(3) is defined as $p_{t|0}(r^{(t)}|r^{(0)}) = f\left(\omega\left(r^{(0)^T} r^{(t)}\right), t\right)$, where $\omega(r)$ is the rotation angle in radians for any r $\in SO(3)$.The final probability estimation can be described as

$$f(\omega, t) = \sum_{\ell\in\mathbb{N}} (2\ell+1)\mathrm{e}^{-\ell(\ell+1)t/2} \frac{\sin((\ell+1/2)\omega)}{\sin(\omega/2)} \quad (2)$$

With the probability estimation, the corresponding conditional score can be computed as

$$\nabla \log p_{t_0}\left(r^{(t)}|r^{(0)}\right) = \frac{r^{(t)}}{\omega(t)} \log\left(r^{(0,t)}\right) \frac{\partial\omega f(\omega(t), t)}{f(\omega(t), t)},$$

With a well-trained denoising score matching(DSM) network on tangent space of R(3) and SO(3), a protein backbone can be sampled from noise with n steps by Euler–Maruyama discretization(Bortoli et al., 2022).

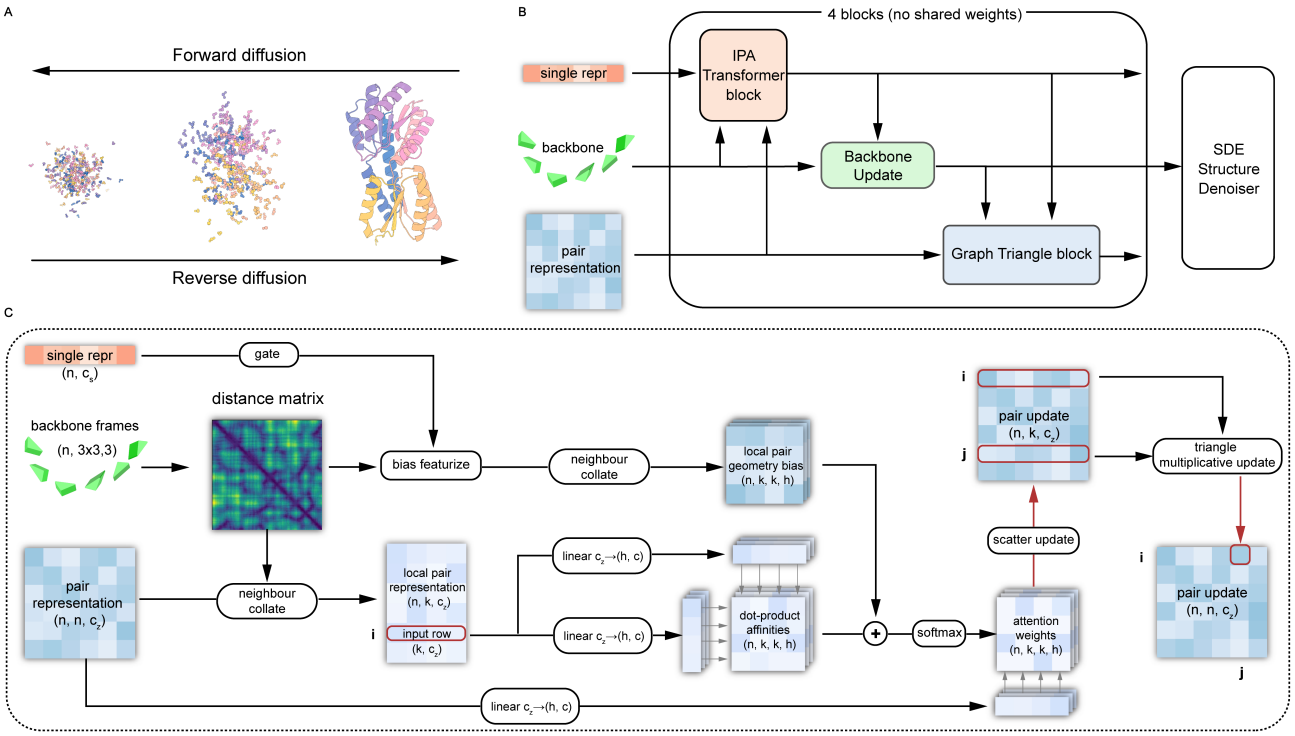

Figure 2: Model illustration (A) The Protein backbone diffusion model is trained to recover noised structures and generate new structures by reversing the forward process. (B) The overall architecture of Proteus. (C) The detailed model architecture of graph triangle block.

In summary, the goal of the network is to predict a denoised protein backbone at timestep 0 given noisy backbones of any timesteps, in which the probability estimation can be calculated, and the stochastic differential equation(SDE) can be applied to sample a denoised protein backbone. We recommend readers check Yim et al.'s work for a more profound understanding of SE(3) diffusion.

Deep learning network architectures for protein structure modeling The model architectures of Alphafold2(Jumper et al., 2021) and Rosettafold(Baek et al., 2021) have significantly advanced the development of protein backbone generation models. We will explain the details of these networks here to better clarify the motivation behind our work. Alphafold2 utilizes Evoformer, a module that translates multiple sequence alignment (MSA) and structure template information into sequence and edge representations of the query sequence. Then, the Structure module translates these latent representations into the protein structure by iterating the protein structure initiated from the 'black hole.' The property of the Structure module takes a structure as input and updates it through iteration, making it a suitable backbone network for protein backbone diffusion(Anand & Achim, 2022; Yim et al., 2023; Lin & AlQuraishi, 2023). Additionally, the SE(3)-transformer(Fuchs et al., 2020) adopted by Rosettafold also possesses the same capability. In addition to the aforementioned networks that update protein structure, Alphafold2 improves prediction accuracy, mainly thanks to the triangle attention layer, a transformer-like architecture that integrates the edge representation of proteins biased by the third edge. Due to its O(N^3) complexity, the current protein backbone diffusion model mostly adopts a standard message passing layer or Unet(Ronneberger et al., 2015) to update the edge representation, except for RFdiffusion(Watson et al., 2023), which adopted axial attention(Ho et al., 2019) on the pair and has a more representative ability for the protein backbone denoising task.

3. Methods

3.1. Model Architecture

Overview In this section, we will introduce the model architecture of Proteus. Proteus iteratively updates the structural frames of proteins through a series of L layers of folding blocks. As shown in Figure 2B, a single layer of the folding block takes input from three different tracks: node representation, edge representation, and structure frames. A folding block comprises an IPA-Transformer block, a backbone update layer, and a graph triangle block, each designed to model and update one of the three tracks while being aware of the representations in the other tracks.

Algorithm 1 Proteus Model Inference

function SAMPLE($N_{res}, N_{step} = 100, N_{layer} = 4, t_{min} = 0.005,$ $noise_scale = 0.1$)
  $t = 1.0, \delta t = (1 - t_{min})/N_{step}$
  for $i \in [1, \ldots, N_{res}]$ **do**
    $\vec{x}_i^{(t)} \sim \mathcal{N}(0, Id_3)$,$r_i^{(t)} \sim \mathcal{N}(0, Id)$
    $T_i^{(t)}$ = $(\vec{x}_i^{(t)}, r_i^{(t)})$, T_i^{prev} = $(\vec{0}, I)$
  end for
  z_{rel} = Relative_Position(N_{res})
  if is oligomer **then**
    z_{rel} = ChainBreak(z_{rel})
  end if
  # denoising iterations:
  for $n \in [N_{step}, \ldots, 1]$ **do**
    $s^{(t)}, z^{(t)}$ = InputEmbedder(t, z_{rel})
    $s^{(t)}, z^{(t)}$ += ConditionEmbedder(T^{prev})
    $\hat{T}^{(0)} = T^{(t)}$
    for $l \in [1, \ldots, N_{layer}]$ **do**
      $s^{(t)}$ = IPATransformer($s^{(t)}, z^{(t)}, \hat{T}^{(0)}$)
      $\hat{T}^{(0)}$ = BackboneUpdate($s^{(t)}$)
      $z^{(t)}$ = TriangleGraph($s^{(t)}, z^{(t)}, \hat{T}^{(0)}$)
    end for
    T^{prev} = $\hat{T}^{(0)}$
    $t = t - \delta t$
    $T^{(t)}$ = $\text{SDE}_{(SE3)}(\hat{T}^{(0)}, T^{(t)}, t, \delta t, noise_scale)$
  end for
end function

The IPA-Transformer block consists of an Invariant Point Attention (IPA) followed by a standard transformer. IPA performs standard attention with a bias based on spatial inter-residue atom distance and edge representation. The Backbone Update is adopted from Alphafold2, where a linear layer is employed to predict translation and rotation updates for each residue's frame based on the updated sequence representation.

The graph triangle block is applied to update the edge representation and employs a graph-based attention mechanism on edge representation with a sequence representation-gated structure bias. The entire network consists of L layers of folding blocks, and no weights are shared between them. Feature initialization is performed using the diffusion timestep as sequence representation. Edge representation is initialized using a relative sequential distance map introduced in AF2. If self-condition is available, the Ca distance map of the previous prediction's residues and relative rotational features are added as additional features, with further details provided in Table 5. Our main focus here is on describing the graph triangle block, which is where most of the designability and efficiency improvements originate.

Graph triangle block The graph triangle block is designed to update the edge representation of the protein. We innovated by the original concept of triangle attention and the multiplication approach from Alphafold2's network, which is the primary module within Alphafold2's Evoformer. Triangle attention is crucial for capturing protein geometry by implicitly maintaining and updating edge representations between each residue. While it is a powerful module and a key contributor to Alphafold2's outstanding performance, it faces challenges related to its $O(N^3)$ complexity and memory requirements. A single layer of Evoformer can consume up to 20GB of memory when training on a 384-residue protein, necessitating the use of gradient checkpoint techniques, particularly because there are 48 layers of Evoformer blocks. In terms of diffusion modeling, Alphafold2 can take at least 10 minutes on a V100 GPU to generate a protein with 50 steps. Another inconsistency when applying the triangle technique to the protein backbone diffusion model is the absence of direct constraints from the structure. Evoformer is designed to integrate Multiple Sequence Alignment (MSA) information into single and edge representations, with no structural information available prior to the first block of the Structure Module.

Considering the two major limitations mentioned above when applying the triangle technique to protein backbone diffusion, we introduce the graph triangle block, where triangle attention of edge representation is calculated with $O(NK^2)$ complexity. Based on the noisy input structure, we define k nearest neighbor residues for each residue and then collate the $N*K$ edge from the entire $N*N$ edge representation. Each residue's K edges calculate attention logits with each other. To incorporate 3D spatial information, we calculate the third edge's inter-atom distances and extract their RBF features as structure bias instead of directly utilizing the representation of the third edge. Furthermore, the structure bias is gated by a feedforward network using the node representations of the edge's starting and ending residues' sequence representation, facilitating full integration of input from all three tracks. Once triangle attention and updates are performed for $N * K$ edge representations, we apply a triangle multiplicative update, as used in Evoformer, to update all $N * N$ edge representations.

This technique provides three advantages compared to the previous axial attention layer used in RFdiffusion and the simple message-passing layer used in FrameDiff and Genie. First, triangle attention has been successfully introduced for edge representation updates, significantly enhancing the model's ability to model the protein backbone diffusion process, where the model achieves superior results even with significantly fewer steps. This results in at least a threefold speedup and significantly outperforms the baseline model. Second, due to the substantial reduction in memory demand, we can now train on much longer proteins without the need

Table 1: Performance comparison on unconditional monomer generation task. The metric is average from length 100, 200, 300, 400, 600 as in Figure 1

Method	N_{params}	Designability	Sample time(s)	Diversity	Timesteps
Proteus	19.8M	0.921	18.20	0.235	100
RFDiffusion	59.8M	0.705	120.24	0.328	50
GENIE(SwissProt)	4.1M	0.349	188.07	0.163	1000
FrameDiff	17.4M	0.405	40.47	0.136	500
Chroma	18.5M	0.174	18.31	0.038	500

to crop the input. We can handle proteins up to 512 residues in length, compared to the maximum of 384 residues in the training of Alphafold2 and RFdiffusion, leading to improved performance in generating longer protein structures. The last advantage is that we consider input from all three tracks to update edge representation and protein structure, resulting in a more compact model architecture. This architecture is best suited for structure-to-structure tasks, such as protein structure generation and predicting the protein's apo-to-holo conformational transition(Hou et al., 2023).

3.2. Training

Dataset We curated a dataset from the Protein Data Bank (PDB)(Berman et al., 2000) with a cutoff date of August 1st, 2023. Instead of training solely on monomers, we split all protein oligomers into single chains. To eliminate duplicate protein chains, we mapped protein sequences to UniProt IDs and retained only the highest resolution structure among chains sharing the same UniProt ID and having an 80 percent sequence overlap. Subsequently, we filtered the remaining protein chains, ensuring a maximum chain length of 512 and a minimum length of 60. We kept proteins that contained a maximum of 50 percent of the loop regions identified by the DSSP program(Kabsch & Sander, 1983). This curation process resulted in 50,773 single-chain proteins for training. Our findings indicate that augmenting the dataset with additional single-chain data derived from oligomers significantly enhances model performance compared to using only monomers. Ablation data is presented in Table 4.

Training losses We have adopted the training loss used in FrameDiff, which can be divided into two main components: denoising translation and rotation score-matching losses, along with auxiliary losses involving the pair-wise distance matrix and coordinate loss on backbone atoms, as depicted in Equation 3. Oxygen atoms are calculated based on other backbone atoms

$$\mathcal{L} = \underbrace{\mathcal{L}_{\text{trans}} + 0.5\mathcal{L}_{\text{rot}}}_{\text{dsm loss}} + \underbrace{0.25\mathcal{L}_{\text{coord}}^{t<0.25} + 0.25\mathcal{L}_{\text{dm}}^{t<0.25}}_{\text{auxiliary loss}} \quad (3)$$

Here, $\mathcal{L}\text{trans}$ computes the L2 loss between predicted translations and native translations, while $\mathcal{L}\text{rot}$ calculates the L2 loss on rotation matrices weighted by $\lambda_t^r = 1/\mathbb{E}[\left|\nabla \log p_{t|0}\left(R_n^{(t)}|R^{(0)}\right)\right| \text{SO}(3)]$, following the approach in (Song et al., 2020). $\mathcal{L}\text{coord}$ represents the coordinate loss between predicted backbones and native backbones. Lastly, $\mathcal{L}_{\text{dm}}$ computes the pair-wise distance loss between predicted atom positions and native positions. We apply auxiliary losses when $t < 0.25$. Empirically, we have observed that including auxiliary losses at a small timestep can aid convergence and enhance the model's performance, which aligns with the findings of FrameDiff.

3.3. Sampling

The sampling procedure is described in Algorithm 1. The initial frame is sampled from rotation and translation separately. We simply adopt Gaussian distribution on $\mathbb{R}^3$ for initial translation. For rotation, we first sample the coefficients of orthonormal basis vectors of the Lie algebra $\mathfrak{so}(3)$ and translate them into the rotation matrix. Proteus takes the input noisy structure and outputs the prediction structure at a timestep equal to 0. Then, we use Euler–Maruyama discretization(Bortoli et al., 2022) as the SDE solver with N_{steps} of iteration. Additionally, we add chain breaks first introduced by Minkyung Baek on the relative position index for oligomer generation.

4. Experiments

Proteus undergoes evaluation through both in-silico validation and in vitro experiments. In Section 4.1, we analyze Proteus' performance in the unconditional monomer generation task. We compare Proteus with several promising protein backbone diffusion models, including Chroma(Ingraham et al., 2022), RFdiffusion(Watson et al., 2023), FrameDiff(Yim et al., 2023), and Genie(Lin & AlQuraishi, 2023). Additionally, we assess our model's performance in complex generation, comparing it to Chroma for dimer, trimer, and tetramer configurations.

In Section 4.2, we conduct in vitro experiments to validate the performance of 16 long protein design models generated by Proteus. These experiments aim to assess our model's capabilities in real-world scenarios.

4.1. in-silico protein generation and evaluation

Monomer backbone generation and evaluation To fairly benchmark the performance of a protein backbone diffusion model, three major concerns must be assessed: designability, efficiency, and diversity. Results are shown in Table 1

The first concern is **designability**, which refers to whether any protein sequence exists that can be folded into the generated backbone. Designability stands as the most significant criterion for evaluating a diffusion model's performance, as it directly influences the model's ability to create proteins that can exist in the real world.

The second concern is **efficiency**. Efficiency plays a dominant role in determining the success rate of design models, especially in scenarios where computational resources are limited – a common situation in real-world design tasks. With finite computational resources, efficiency becomes equally important alongside designability.

The third concern is **diversity**, which measures the structural similarity of the generated protein backbones. A higher level of diversity implies the provision of a greater number of solutions to protein design tasks.

Table 2: *Success efficiency benchmark of various lengths.* Success efficiency is computed as the computation resource(how many seconds on a single A40) needed to generate a designable backbone with scRMSD < 2 Å. * means the model does not predict any designable backbone in 200 samples.

Protein length	100	200	300	400	600	800
Proteus	7.02	9.11	13.47	23.09	50.62	196.11
RFDiffusion	24.84	43.97	97.14	262.64	1494.28	30500.00
GENIE(SwissProt)	72.66	127.55	482.97	21169.0	*	*
FrameDiff	23.12	45.12	62.11	216.88	1851.11	*
Chroma	24.88	56.66	352.00	4020.00	*	*

To evaluate designability, a generated protein backbone is input into ProteinMPNN (Dauparas et al., 2022) at temperature 0.1, except for Chroma's backbone. We utilize Chroma's inverse folding model, ChromaDesign (Ingraham et al., 2022) at temperature 0.1 and diffusion augmentation 0.5, to replace ProteinMPNN, as we have observed a higher success rate of Chroma's backbone by using ChromaDesign, as reported in its paper. Then the inverse folding model generates multiple sequences corresponding to the input backbone. Subsequently, the structure corresponding to each sequence is predicted using ESMFold(Lin et al., 2023). The designability of the backbone is represented by the highest TM-score (scTM, where higher values are better) between the backbone and the predicted structure. We use a C_α-RMSD threshold of $< 2, \text{Å}$ as the criterion for successful design. A generated protein is considered non-functional and likely to fail experimental validation if the C_α-RMSD between the designed and predicted models exceeds this threshold.

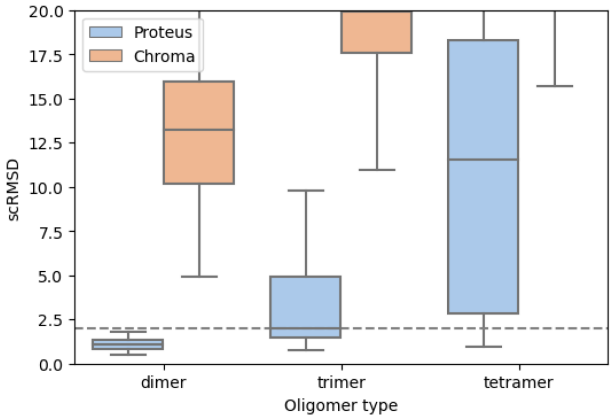

Figure 3: Complexes generation benchmark. scRMSD is based on 200 backbone samples of each oligomer type, and each chain is fixed to 200 residues.

Efficiency is estimated based on the time taken to generate a protein backbone using an A40 GPU. We further calculate **success efficiency** in Table 2, which indicates how much computation time is required to generate a designable backbone. This metric is particularly important in real protein design tasks, as it considers both designability and model speed, providing valuable insights when computational resources are limited.

Diversity is evaluated by calculating the ratio of diverse designable structure clusters to the total number of generated backbones. Specifically, we cluster the backbones using MaxCluster(Herbert & Sternberg, 2008) with a TM-score cutoff of 0.6. Notably, we only consider designable backbones defined as those with a C_α-RMSD of $< 2, \text{Å}$. This is because sub-optimal models often produce unrealistic structures that cannot be folded by any sequence and are highly non-physical. Including such backbones in the assessment would skew the metric towards models generating random, unconstrained backbones rather than models that genuinely capture the realistic distribution while adhering to the physical principles of the protein universe.

Complexes backbone generation and evaluation We compare Proteus' ability to generate protein complexes to Chroma on a similar setting of monomer design benchmark with various oligomer types. We test scRMSD on two chains(dimer), three chains(trimer), and four chains(tetramer). Each chain is fixed to 200 residues.

Results We analyze Proteus's performance on monomer generation tasks with all available leading-edge protein backbone diffusion models, including RFDiffusion, Genie, FrameDiff, and Chroma. We sample 200 backbones for every length [100, 200, 300, 400, 600] and take the average score of each length as reported in Table 1. We surpass

all current diffusion models in designability, even RFdiffusion, which uses a more computation-intensive model architecture and is pretrained from the structure prediction model. Figure 1B shows the scRMSD distribution of each length. Proteus is comparable with RFdiffusion at shorter lengths(100-300) and shows significantly better performance at long proteins(¿300). We contribute this to our model's novel architecture that uses graph-level triangle techniques and allows the model to be efficiently trained on longer sequences of up to 512 amino acids. The outstanding performance of Proteus in complex assembly generation, far surpassing Chroma, demonstrates that Proteus's generalization ability is sufficient to extend to complex multi-domain protein design and the design of large nanoparticles.

Table 3: Proteus sample parameter benchmark

noise level ζ	0.1	0.5	1.0	0.1	0.1	0.1
N_{timestep}	100	100	100	50	200	500
< 2Å scRMSD (↓)	92.0%	87.8%	37.3%	89.8%	91.0%	92.2%
DIVERSITY (↑)	0.23	0.27	0.16	0.26	0.22	0.22

Notably, while maintaining the highest designability, Proteus achieves a similar computational time as Chroma. By enhancing network representation capabilities, Proteus only requires 100 sampling steps without compromising designability, as indicated in Table 3. This is in contrast to Genie, which requires 1000 steps, and FrameDiff and Chroma, which require 500 steps. In Figure 1C, the computation time for various lengths is depicted. Proteus is faster than Chroma for lengths shorter than 400. However, Chroma becomes more time-efficient when generating proteins longer than 400. Nevertheless, its success efficiency lags behind all diffusion models due to its graph-based model architecture. With the best designability and efficient performance at the graph-based level, Proteus clearly outperforms all other diffusion models, as demonstrated in the success efficiency benchmark presented in Table 2.

Table 4: Ablation Study

Dataset	Monomer	✓	✓	✓	✓	✓
	Oligomer		✓	✓	✓	✓
Self-Condition	C_α distance matrix	✓	✓			
	Structure featurizer			✓	✓	✓
Model	Graph triangle attention				✓	✓
	Structure bias					✓
	Triangle multiplication				✓	✓
Results	Success rate	36.9%	59.2%	63.2%	87.5%	92.1%

ablation We make detailed ablations in Table 4 by dividing the improvement into three parts: dataset augmentation, self-condition enhancement, and model architecture. Previous protein backbone diffusion methods had proposed various datasets: FoldingDiff uses a total of 30395 protein chains from the CATH database(Sillitoe et al., 2014) with a crop length of 128. Genie adopted AlphaFold-Swissprot database(Varadi et al., 2021) containing 195214 protein chains with a cutoff of pLDDT ¿ 80 and at most 256 residues. FrameDiff(Yim et al., 2023) curated a dataset from PDB with only monomer entry up to 512 residues, resulting in a dataset containing 20312 chains. Dissimilar from all the above datasets, we expand the dataset by adding oligomer protein and split into single chains, resulting in a dataset of 50773 chains and a significant improvement in designability. The ablation benchmark also shows adding a roration-based pair-wise self-condition feature similar to the template featurization introduced by Alphafold2 makes the model better capture the predicted structure of previous steps.

Though these modifications greatly improve the model's performance, it still lags behind RFDiffusion. Proteus surpasses RFdiffusion by introducing the local triangle graph. The triangle technique greatly increases the model's designability. Due to attention calculation on the graph level, the sample speed is kept at a similar level to Chroma. The structure bias used in the calculation of triangle attention also obviously improves designability, indicating structure level information helps the update of edge representation, which is a new finding that is unexplored by Alphafold2.

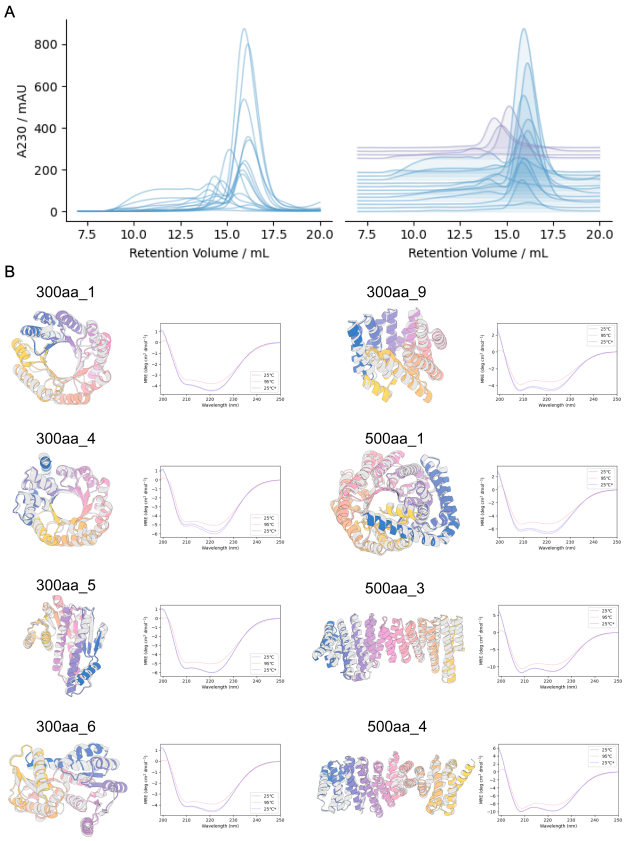

Figure 4: Generated designs are well-folded monomer and thermostable. (A) SEC elution profiles of experimental designs. (B) Examples of designed proteins that expressed as soluble monomers. Designs (grey) overlaid with AF2 predictions (colours) are shown on the left, alongside 25°C and 95°C circular dichroism (CD) spectra on the right.

4.2. experimental validation

We ordered 16 designs, including 12 designs of 300 amino acid proteins and 4 designs of 500 amino acid proteins. All designs were expressed in E.coli. Of these, size exclusion chromatography (SEC) indicated that 9 of the 300 amino acid designs and 3 of the 500 amino acid designs showed monodisperse peaks corresponding to the expected molecular weight. Circular dichroism (CD) spectroscopy confirmed that these designs had secondary structures consistent with the design models. And these designs were highly thermostable, maintaining a stable secondary structure across a temperature range of 25°C to 95°C. Experimental data is shown in Figure 4.

5. Related work

Structure diffusion models on proteins Inspired by the considerable success of diffusion modes(Ho et al., 2020; Song et al., 2020; Bortoli et al., 2022) in many fields, past works had developed diffusion models on protein modality of structure and sequence or include both(Anand & Achim, 2022). Anand & Achim first proposed a backbone, sequence, and sidechain co-diffusion model utilizing the Structure Module of Alphafold2. Then, various methods diffuse on inter-residue geometry(Lee et al., 2023) and backbone dihedral angle(Wu et al., 2022) had been proposed. Current protein structure diffusion models mainly focus on diffusion in SE3 or R3 space to generate structure in an end-to-end fashion(Yim et al., 2023; Lin & AlQuraishi, 2023) and further expand to function motif scaffolding(Trippe et al., 2022; Yim et al., 2024), more general condition and efficiency is achieved by Chroma(Ingraham et al., 2022) by the geometric constrained harmonic constraint. Beyond above methods, RFdiffusion(Watson et al., 2023) gets state-of-the-art designability with experimental characterization by finetuning on Rosettafold2(Baek et al., 2023). Differing from RFdiffusion, we accomplish higher designability and efficiency without pretraining.

Other diffusion models on proteins Based on the success of protein structure diffusion, much research focuses on the diffusion of protein sequence or integrating sequence information into protein structure diffusion. EvoDiff(Alamdari et al., 2023) utilize evolutionary-scale protein sequence data to generate protein sequence. Though promising analysis proves EvoDiff's diverse generation space, the structure prediction metrics indicate rather low stability and experimental success rates. (Chu et al., 2023) tries to integrate protein sequence features during backbone diffusion schedule by using a timestep-dependent ProteinMPNN model to design sequences on noisy structures, potentially expanding backbone motif scaffolding ability to side chain scaffolding, but decreases the model's designability, which against intuition that integrate sequence can improve structure quality. In the specific task of antibody diffusion, a lot of work(Luo et al., 2022; Kong et al., 2023; Peng et al., 2023) has been made to simultaneously generate the sequence and structure of complementarity-determining regions(CDR).

Protein-ligand complexes prediction In previous works, prediction of protein-ligand complexes are considered as a regression fashion prediction of rigid docking of ligands to holo proteins(Stärk et al., 2022; Lu et al., 2022). To exploit the generation ability of the diffusion model, Diffdock has been proposed with great improvement on docking accuracy(Corso et al., 2022). With the diffusion model showing potential for modeling protein structure, protein-ligand co-folding models(Qiao et al., 2023) have been proposed with a more accurate prediction on the task of holo protein structures not known, which is a critical challenging application in drug discovery. Additionally, RFdiffusion All Atom(Krishna et al., 2023) are trained to design novel small molecule binders.

6. Discussion

In this paper, we propose a novel model architecture allied with protein backbone diffusion and show superior designability and efficiency far beyond all available backbone diffusion models without pre-train. By transferring triangle techniques into the integration of latent representation of residue edges by the construction of KNN graph and building multi-track interaction networks, we even largely surpass RF diffusion on longer monomer(>=400 amino acids) generation. Further analysis of the complex generation benchmark shows Proteus's fascinating generalizability from monomer to oligomer, which far exceeds Chroma, the current best model to generate complex structures.

There are several research directions in which we believe Proteus can have a broad impact. First, there is the area of large nanoparticle generation. Proteus has demonstrated exceptional performance in generating long proteins and complexes with high accuracy and experimental success rates. Many native protein functions arise from complex interactions involving multi-domains or multi-chains, necessitating a model capable of accurately modeling a large number of residues while maintaining precision.

The second direction involves protein-ligand co-folding. Recent advancements, such as Rosettafold All Atom(Krishna et al., 2023) and the latest version of Alphafold2, have shown significant improvements compared to Diffdock, the current state-of-the-art in blind rigid docking. By incorporating Proteus into the diffusion modeling of ligands and proteins, we aim to bring accurate and efficient generative capabilities to co-folding tasks.

Impact Statements

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Alamdari, S., Thakkar, N., van den Berg, R., Lu, A. X., Fusi, N., Amini, A. P., and Yang, K. K. Protein generation with evolutionary diffusion: sequence is all you need. *bioRxiv*, 2023. doi: 10.1101/2023.09.11.556673.

Anand, N. and Achim, T. Protein structure and sequence generation with equivariant denoising diffusion probabilistic models. *ArXiv*, abs/2205.15019, 2022. URL https://api.semanticscholar.org/CorpusID:249192041.

Anishchenko, I., Pellock, S. J., Chidyausiku, T. M., Ramelot, T. A., Ovchinnikov, S., Hao, J., Bafna, K., Norn, C., Kang, A., Bera, A. K., DiMaio, F., Carter, L., Chow, C. M., Montelione, G. T., and Baker, D. De novo protein design by deep network hallucination. *Nature*, 600(7889):547–552, 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-04184-w. URL https://doi.org/10.1038/s41586-021-04184-w.

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., Wang, J., Cong, Q., Kinch, L. N., Schaeffer, R. D., Milln, C., Park, H., Adams, C., Glassman, C. R., DeGiovanni, A., Pereira, J. H., Rodrigues, A. V., van Dijk, A. A., Ebrecht, A. C., Opperman, D. J., Sagmeister, T., Buhlheller, C., Pavkov-Keller, T., Rathinaswamy, M. K., Dalwadi, U., Yip, C. K., Burke, J. E., Garcia, K. C., Grishin, N. V., Adams, P. D., Read, R. J., and Baker, D. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, 2021. doi: 10.1126/science.abj8754. URL https://www.science.org/doi/abs/10.1126/science.abj8754.

Baek, M., Anishchenko, I., Humphreys, I. R., Cong, Q., Baker, D., and DiMaio, F. Efficient and accurate prediction of protein structure using rosettafold2. *bioRxiv*, 2023. doi: 10.1101/2023.05.24.542179. URL https://www.biorxiv.org/content/early/2023/05/25/2023.05.24.542179.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. The Protein Data Bank. *Nucleic Acids Research*, 28(1): 235–242, 01 2000. doi: 10.1093/nar/28.1.235. URL https://doi.org/10.1093/nar/28.1.235.

Bortoli, V. D., Mathieu, E., Hutchinson, M., Thornton, J., Teh, Y. W., and Doucet, A. Riemannian score-based generative modelling, 2022.

Chu, A. E., Cheng, L., Nesr, G. E., Xu, M., and Huang, P.-S. An all-atom protein generative model. *bioRxiv*, 2023. doi: 10.1101/2023.05.24.542194. URL https://www.biorxiv.org/content/early/2023/05/25/2023.05.24.542194.

Corso, G., Stärk, H., Jing, B., Barzilay, R., and Jaakkola, T. Diffdock: Diffusion steps, twists, and turns for molecular docking. *ArXiv*, abs/2210.01776, 2022. URL https://api.semanticscholar.org/CorpusID:252693198.

Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R. J., Milles, L. F., Wicky, B. I. M., Courbet, A., de Haas, R. J., Bethel, N., Leung, P. J. Y., Huddy, T. F., Pellock, S., Tischer, D., Chan, F., Koepnick, B., Nguyen, H., Kang, A., Sankaran, B., Bera, A. K., King, N. P., and Baker, D. Robust deep learningased protein sequence design using proteinmpnn. *Science*, 378(6615):49–56, 2022. doi: 10.1126/science.add2187. URL https://www.science.org/doi/abs/10.1126/science.add2187.

Fuchs, F. B., Worrall, D. E., Fischer, V., and Welling, M. Se(3)-transformers: 3d roto-translation equivariant attention networks, 2020.

Herbert, A. and Sternberg, M. MaxCluster: a tool for protein structure comparison and clustering, 2008.

Ho, J., Kalchbrenner, N., Weissenborn, D., and Salimans, T. Axial attention in multidimensional transformers, 2019.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *ArXiv*, abs/2006.11239, 2020. URL https://api.semanticscholar.org/CorpusID:219955663.

Hou, M.-H., Jin, S.-R., Cui, X.-Y., Peng, C.-X., Zhao, K.-L., Song, L., and Zhang, G.-J. Protein multiple conformations prediction using multi-objective evolution algorithm. *bioRxiv*, 2023. doi: 10.1101/2023.04.21.537776.

Ingraham, J., Baranov, M., Costello, Z., Frappier, V., Ismail, A., Tie, S., Wang, W., Xue, V., Obermeyer, F., Beam, A., and Grigoryan, G. Illuminating protein space with a programmable generative model. *Nature*, 623:1070 – 1078, 2022. URL https://api.semanticscholar.org/CorpusID:254222077.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B.,

Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021. doi: 10.1038/s41586-021-03819-2.

Kabsch, W. and Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, 1983. doi: https://doi.org/10.1002/bip.360221211. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/bip.360221211.

Kong, X., Huang, W., and Liu, Y. Conditional antibody design as 3d equivariant graph translation, 2023.

Krishna, R., Wang, J., Ahern, W., Sturmfels, P., Venkatesh, P., Kalvet, I., Lee, G. R., Morey-Burrows, F. S., Anishchenko, I., Humphreys, I. R., McHugh, R., Vafeados, D., Li, X., Sutherland, G. A., Hitchcock, A., Hunter, C. N., Baek, M., DiMaio, F., and Baker, D. Generalized biomolecular modeling and design with rosettafold all-atom. *bioRxiv*, 2023. doi: 10.1101/2023.10.09.561603. URL https://www.biorxiv.org/content/early/2023/10/09/2023.10.09.561603.

Lee, J. S., Kim, J., and Kim, P. M. Score-based generative modeling for de novo protein design. *Nature Computational Science*, 3:382 – 392, 2023. URL https://api.semanticscholar.org/CorpusID:250645859.

Lin, Y. and AlQuraishi, M. Generating novel, designable, and diverse protein structures by equivariantly diffusing oriented residue clouds, 2023.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., and Rives, A. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574. URL https://www.science.org/doi/abs/10.1126/science.ade2574.

Lu, W., Wu, Q., Zhang, J., Rao, J., Li, C., and Zheng, S. Tankbind: Trigonometry-aware neural networks for drug-protein binding structure prediction. *bioRxiv*, 2022. doi: 10.1101/2022.06.06.495043. URL https://www.biorxiv.org/content/early/2022/06/06/2022.06.06.495043.

Luo, S., Su, Y., Peng, X., Wang, S., Peng, J., and Ma, J. Antigen-specific antibody design and optimization with diffusion-based generative models. *bioRxiv*, 2022. doi: 10.1101/2022.07.10.499510. URL https://www.biorxiv.org/content/early/2022/07/25/2022.07.10.499510.

Peng, Z., Han, C., Wang, X., Li, D., and Yuan, F. Generative diffusion models for antibody design, docking, and optimization. *bioRxiv*, 2023. URL https://api.semanticscholar.org/CorpusID:263225383.

Qiao, Z., Nie, W., Vahdat, A., au2, T. F. M. I., and Anandkumar, A. State-specific protein-ligand complex structure prediction with a multi-scale deep generative model, 2023.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation, 2015.

Satorras, V. G., Hoogeboom, E., and Welling, M. E(n) equivariant graph neural networks, 2022.

Sillitoe, I., Lewis, T. E., Cuff, A., Das, S., Ashford, P., Dawson, N. L., Furnham, N., Laskowski, R. A., Lee, D., Lees, J. G., Lehtinen, S., Studer, R. A., Thornton, J., and Orengo, C. A. CATH: comprehensive structural and functional annotations for genome sequences. *Nucleic Acids Research*, 43(D1):D376–D381, 10 2014. ISSN 0305-1048. doi: 10.1093/nar/gku947. URL https://doi.org/10.1093/nar/gku947.

Song, Y., Sohl-Dickstein, J. N., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *ArXiv*, abs/2011.13456, 2020. URL https://api.semanticscholar.org/CorpusID:227209335.

Stärk, H., Ganea, O.-E., Pattanaik, L., Barzilay, R., and Jaakkola, T. Equibind: Geometric deep learning for drug binding structure prediction, 2022.

Trippe, B. L., Yim, J., Tischer, D. K., Broderick, T., Baker, D., Barzilay, R., and Jaakkola, T. Diffusion probabilistic modeling of protein backbones in 3d for the motif-scaffolding problem. *ArXiv*, abs/2206.04119, 2022. URL https://api.semanticscholar.org/CorpusID:249538415.

Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., dek, A., Green, T., Tunyasuvunakool, K., Petersen, S., Jumper, J., Clancy, E., Green, R., Vora, A., Lutfi, M., Figurnov, M., Cowie, A., Hobbs, N., Kohli, P., Kleywegt, G., Birney, E., Hassabis, D., and Velankar, S. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic*

Acids Research, 50(D1):D439–D444, 11 2021. ISSN 0305-1048. doi: 10.1093/nar/gkab1061. URL https://doi.org/10.1093/nar/gkab1061.

Wang, J., Lisanza, S., Juergens, D., Tischer, D., Watson, J. L., Castro, K. M., Ragotte, R., Saragovi, A., Milles, L. F., Baek, M., Anishchenko, I., Yang, W., Hicks, D. R., Expsit, M., Schlichthaerle, T., Chun, J.-H., Dauparas, J., Bennett, N., Wicky, B. I. M., Muenks, A., DiMaio, F., Correia, B., Ovchinnikov, S., and Baker, D. Scaffolding protein functional sites using deep learning. *Science*, 377(6604):387–394, 2022. doi: 10.1126/science.abn2100. URL https://www.science.org/doi/abs/10.1126/science.abn2100.

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., Wicky, B. I. M., Hanikel, N., Pellock, S. J., Courbet, A., Sheffler, W., Wang, J., Venkatesh, P., Sappington, I., Torres, S. V., Lauko, A., Bortoli, V. D., Mathieu, E., Ovchinnikov, S., Barzilay, R., Jaakkola, T., DiMaio, F., Baek, M., and Baker, D. De novo design of protein structure and function with rfdiffusion. *Nature*, 620:1089 – 1100, 2023. URL https://api.semanticscholar.org/CorpusID:266694644.

Wu, K. E., Yang, K. K., van den Berg, R., Zou, J. Y., Lu, A. X., and Amini, A. P. Protein structure generation via folding diffusion, 2022.

Yim, J., Trippe, B. L., Bortoli, V. D., Mathieu, E., Doucet, A., Barzilay, R., and Jaakkola, T. Se(3) diffusion model with application to protein backbone generation, 2023.

Yim, J., Campbell, A., Mathieu, E., Foong, A. Y. K., Gastegger, M., Jiménez-Luna, J., Lewis, S., Satorras, V. G., Veeling, B. S., Noé, F., Barzilay, R., and Jaakkola, T. S. Improved motif-scaffolding with se(3) flow matching, 2024.

A. Model architectural details

Table 5: Input feature of Proteus

Input Feature (shape)	**Description**
timestep (*N_res, 32*)	The timestep embedding from 0 to 1 by sinusoidal positional encoding.
rigid_frame (*N_res, 7*)	Input noisy structure, 3 dim of translation and 4 dim of rotation represented by a quaternion.
rel_pos (*N_res, N_res, 65*)	Relative position index embedding in [-32,32.]
torsion_angle_prev(*N_res, 14*)	Self-condition input structure's residue torsion angle of prev prediction represented by sin and cos formulation, 4 chi angles will be masked out in the network.
distogram_prev(*N_res, N_res, 39*)	Self-condition input structure's pair distance histogram of pseudo $C\beta$, bin_min=3.25Å, bin_max=50.75Å.
unit_vector_prev(*N_res, N_res, 1*)	Self-condition input structure's unit vector of each residue's $C\alpha$ coordinate in another residue's local frame.

Table 6: Model Training Parameters

Model Parameters	
Dimension of sequence track C_s	256
Dimension of edge track C_z	128
K nearest neighbour	32
Heads of graph triangle attention	4
Hidden dimension of triangle attention head	128
Dimension of structure bias feature	64
Dimension of sequence track gate	16
Training Setting	
Max protein length N_{res}	512
Max batch size	16
Max squared residues	3×10^4
Training steps	2×10^6
Training time	$\approx$ 20 days
Device	2× A40

B. In vitro experiment protocol

Protein Purification and Expression Synthetic genes encoding designed protein were purchased from Universe Gene Technology. These sequences were cloned into the pET28a vector, included N-terminal or C-terminal histidine tags and an HRV 3C protease cleavage site. These plasmids were transformed into BL21 (DE3) E. coli competent cells. All transformants were cultured into 50 ml of LB medium with 50 mg/ml kanamycin. Protein expression was induced with 1 mM isopropyl 1-thio--d-galactopyranoside at 37 °C overnight or at 20 °C overnight after initial growth for 6 to 8 h at 37 °C. The cell were harvested by centrifugation and lysed by sonication after resuspension of the cells in lysis buffer (25 mM Tris pH 7.0, 150 mM NaCl). The cell lysate was cleared by centrifugation (12,000 × rpm). The supernatant was purified by 1 ml Ni2+ immobilized metal affinity chromatography with Ni-NTA Superflow resin (Qiagen). Resins with bound cell lysate were washed five times with 5 mL of washing buffer (comprising lysis buffer and 30 mM imidazile) and eluted with 6 mL of

elution buffer (comprising lysis buffer and 300 mM imidazile). Both eluates were analyzed using 15% SDS-PAGE gel to assess purity. The histidine tags were cleaved using histidine-tagged HRV 3C protease during dialysis against lysis buffer overnight. A second IMAC purification was performed for HRV 3C cleaved samples to capture uncleaved protein and HRV 3C protease. Designs were finally purified using Superdex 200 Increase 10/300GL (GE Healthcare) with lysis buffer.

Circular dichroism experiments Circular dichroism spectra were recorded on a Chirascan V100 circular dichroism spectrometer (Applied Photophysics) using protein concentrations ranging from 0.6 to 0.9 mg/ml. Thermal melt analyses were conducted over a temperature range of 25°C to 95°C, measuring CD at 222 nm. Wavelength scans (190 to 260 nm) were recorded at both 25°C and 95°C. All reported measurements were obtained within the linear range of the instrument.