DIFFTOPO: FOLD EXPLORATION USING COARSE GRAINED PROTEIN TOPOLOGY REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

A major challenge in the field of computational de novo protein design is the exploration of uncharted areas within protein structural space, i.e., generating "designable" protein structures that nature has not explored. However, the large degrees of freedom of protein structural backbones complicate the sampling process during protein design. In this work, we propose a new coarse grained protein structure representation method DiffTopo - an E(3) Equivariant 3D conditional diffusion model, which greatly increases the sampling efficiency. Combined with the RFdiffusion framework, novel protein folds can be generated rapidly, allowing for efficient exploration of the designable topology space. This opens up possibilities to solve the problem of generating new folds as well to functionalize de novo proteins through motif scaffolding, where functional or enzymatic sites can be introduced into novel protein frameworks.

1 INTRODUCTION

Proteins govern vital biological functions, including enzyme catalysis, molecular transport, and cellular activity modulation. The intricate relationship between protein function and three-dimensional architecture is essential. One of the most important goals of the field of de novo protein design involves the computational generation of novel and realistic protein structures that fulfill specific structural and/or functional requirements (Pan & Kortemme, 2021; Kuhlman & Bradley, 2019).

Much work has been done to address the problem of computationally generating new protein structures, but has often encountered challenges in creating diverse and realistic folds. Traditional methods typically apply heuristics to assemble protein fragments into structures, which are limited by expert knowledge and available data (Simons et al., 1997; Minami et al., 2023; Mackenzie et al., 2016; Jacobs et al., 2016). Recently, deep generative models have been proposed to address these issues. Generative models rely on complex equivariant network architectures or loss functions to learn to generate 3D coordinates or internal torsion angles that describe protein structures (Anand et al., 2019; Anand & Achim, 2022; Lin & AlQuraishi, 2023; Luo et al., 2022; Trippe et al., 2023; Yim et al., 2023b;a; Eguchi et al., 2022; Watson et al., 2023). Equivariant architecture (Satorras et al., 2022; Jing et al., 2021; Fuchs et al., 2020) ensures that the probability density of protein structure sampling remains constant under translation and rotation. RFdiffusion (Watson et al., 2023) and Framediff (Yim et al., 2023b) have successfully learned the complex distribution of protein backbones and can generate designable protein backbones. However, these methods still encounter some obstacles in the goal of generating novel, previously unseen structures. For example, when generating backbones with a length less than 200 amino acids (AA), the resulting backbones are very similar to natural structures. The reason is when the length of the amino acid increases, the degree of freedom of the backbone increases exponentially. This is addressed in the work of Taylor et al. (2008) and Harteveld et al. (2022); Yang et al. (2020) where protein structure is conceptualized as a spatial stacking of standard secondary structures, significantly reducing the degree of freedom in the protein fold space.

Motivated by recent advances, we investigated the feasibility of utilizing deep learning models to explore the protein fold space condensed protein rerpesentations. In this work, we introduce a protein representation based on a coarse-grained (CG) topology representation (depicted in Figure 1 top row). We have established a pipeline (Figure 1 bottom row) that utilizes a simple secondary structure string description as conditional input to autonomously generate a plausible CG topology.

Our approach involves training a diffusion model named DiffTopo, which learns the relative spatial position distribution of plausible secondary structures. From this model, we sample and generate CG topology representations with reasonable secondary structures. We then construct the standard secondary structure at the positions specified by CG topology, resulting in what we term a "protein sketch." This Protein Sketch serves as direct input to RFdiffusion, to generate designable backbones consistent with the input. Notably, our method avoids direct sampling using explicit atomic representations, ensuring a fast and efficient sampling process. Difftopo's sampling is oriented towards exploring diverse protein folds, while RFdiffusion is subsequently employed to identify the corresponding designable backbone. This dual-stage approach contributes to the effectiveness and efficiency of our sampling methodology. The detailed description of the methodology is presented in the appendix A.1.

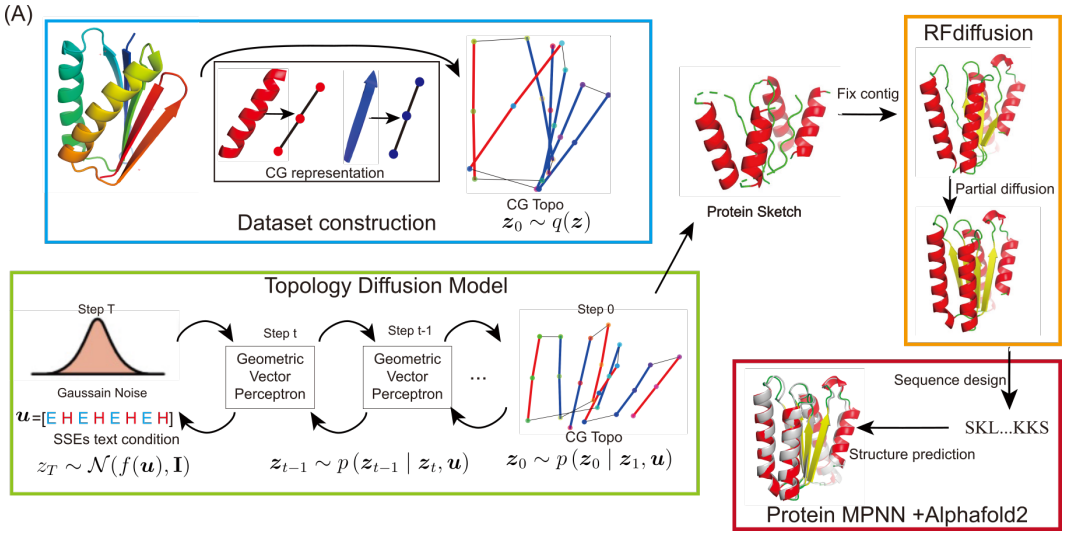

Figure 1: Overview of the DiffTopo de novo protein design pipeline. The first step is to convert the secondary structure into a CG topology representation, followed by transforming the non-redundant backbones in the entire CATH database into CG topology to create a dataset. Then we develop the DiffTopo-RFdiffusion framework for de novo protein design. DiffTopo is the diffusion model to generate CG topology, which can be converted to protein sketches guiding RFdiffusion in the generation of protein backbones.

2 Results

2.1 DiffTopo captures the distributions of native protein structures

Standardize metrics are not available to assess the quality of the generated topological sketches. In backbone generation tasks, evaluation often relies on geometric features like Ramachandran distribution and bond length, angle. Based on the sketch level representation, our evaluation focuses on the diffusion model's ability to estimate data probability density, checking alignment with authentic data distribution. We defined several geometric descriptors for the topological features of (Figure 2A). The DiffTopo generated CG topologies yield a similar distribution to the features observed in natural structures (Figure 2B, 2C). The results indicate that the diffusion model adequately approximates the distribution of real data, despite with a tendency to smoothing discrete high-density peaks.

2.2 Similarity of sketch and backbones

To confirm that the CG topology accurately represents the fold of the real protein, and to verify that the backbone generated by RFdiffusion has the same protein fold as the CG topology, we assess the similarity between protein sketches constructed from CG topology and backbones — both real and

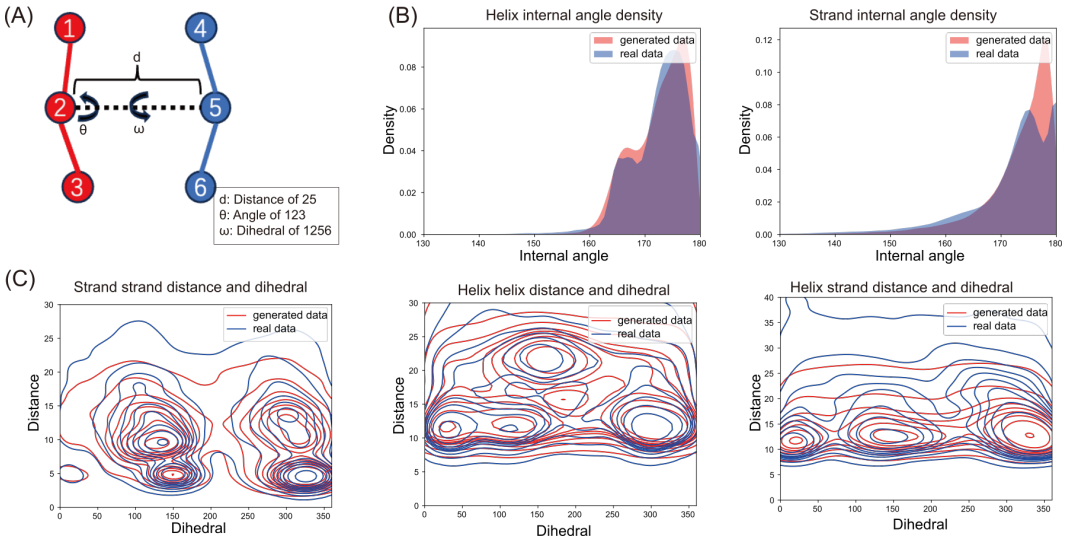

Figure 2: Structural representation and distribution of real and generated data. (A) Local and pairwise geometric features to describe spatial relative positions of SSEs, including internal angles within one SSE, distance between centroids of two SSEs and dihedrals between two SSEs. (B) Local geometric feature distributions of generated (red) and real data (blue). (C) Pairwise geometric feature distributions of generated (red) and real data (blue)

randomly generated by RFdiffusion. Here, we calculate TMscore (Zhang & Skolnick, 2004; Xu & Zhang, 2010) and RMSD metrics to compare native backbones with protein sketches derived from their CG topology. The TMscore and RMSD between sketches generated from CG topology and the corresponding backbones are also calculated. As depicted in Figure 3A and 3B, the arrangement of most Secondary Structure Elements (SSEs) in protein sketches is similar to their corresponding backbones, with TMscores surpassing 0.5 and RMSD below 3Å. The consistently high TMscores and low RMSD values confirm the agreement between the generated backbones and their sketch inputs. This supports the confident classification of these generated backbones and their associated protein sketches as representing the same protein fold.

2.3 Designability, Novelty and Diversity

To test the designability of the generated backbones backbones, we designed amino acid sequences for the backbone and predicted whether these sequences can fold back into the target topology. The designability of the backbone is evaluated through self-consistency assessment. Drawing inspiration from the research of FrameDiff (Yim et al., 2023b) and ProtDiff (Trippe et al., 2023), we quantify self-consistency using Cα RMSD (scRMSD, lower values are better) and predicted local distance difference test (pLDDT, higher values are better). To assess the novelty of the backbone, we pick generated backbones with high designability (ScRMSD $<$3 Åand pLDDT $>$90) and search them in the entire CATH database and report their highest TMscore, referred to as Max PDB TMscore. To quantify diversity, we utilize the clustering function of FoldSeek (van Kempen et al., 2023) to cluster the sampled backbones with a 0.5 TMscore threshold. The TMscore between backbones and the number of clusters obtained are reported. We used RFdiffusion random sampling of backbones with lengths ranging from 40 to 200 as a comparative benchmark. Two types of random sampling were performed: random sampling of CG topology generated backbones using random lengths of SSE strings as input, and random sampling of CG topology using SSE strings with low occurrence frequency in the database. Figures 3C and 3D illustrate that, while there was a trade-off in designability compared to using RFdiffusion alone, many designs still maintained scRMSD values below 3Å. However, Figure 3D highlights that, when employing SSE strings with low frequency in the database for random sampling, our approach yielded significantly higher novelty in obtaining highly designable backbones compared to RFdiffusion. Figure 3E shows the diversity in spatial arrangements of secondary structures sampled by our method. We sampled 500 highly designable

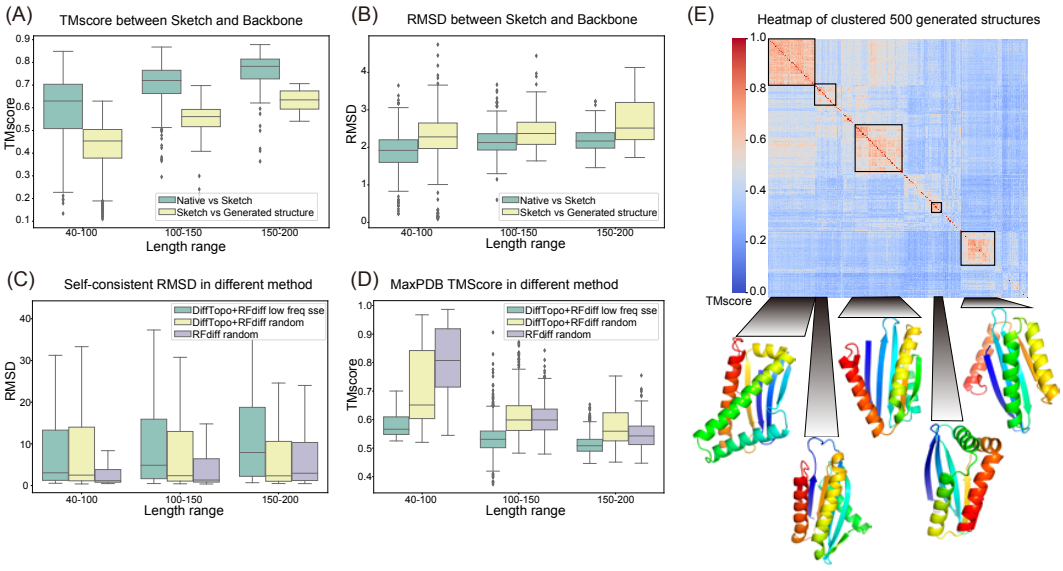

Figure 3: Features of protein sketches and generated backbones. (A) TM-score between sketches and native structures (green) and generated backbones at different lengths (yellow). (B) RMSD between sketches and native structures (green) and generated backbones at various lengths (yellow). (C) Designability of generated backbones: scRMSD based on 1000 samples for each length range (40-100, 100-150, 150-200) in different methods. Purple: RFdiffusion random sample (baseline), Yellow: our pipeline's random sampling of SSE strings, Green: our pipeline's sampling of low-probability SSE strings in the database. (D) Novelty of generated backbones: color of box represents the same as in (C). (E) Diversity of generated backbones: heatmap of TM scores between each backbone after hierarchical clustering. Displaying structures from the top 5 largest clusters.

backbones for the sequence "EEEHHEHE" and calculated the TM score between each structure, resulting in a total of 109 clusters. Here, we present the representative structures of the 5 largest clusters.

3 Application for Protein Design

3.1 Novel Fold Exploration and Scaffold Generation for Functional Motif

In order to showcase the model's exploration capabilities across different folds, we selected three distinct secondary structure compositions: α, all-β, and mixed α-β, for protein fold exploration. Figure 4 illustrates the novel folds with high designability that we discovered in each of these secondary structure compositions. Our pipeline demonstrates remarkable versatility by uncovering novel folds that do not exist in various secondary structure combinations, including all-α, all-β, and mixed α-β folds. We also present the most similar structures to these folds that can be found with FoldSeek (van Kempen et al., 2023). Interestingly, the closest structures are not of the same fold as tose that were generated, demonstrating the novelty of our structures. Notably, even in the presence of 394 distinct 7-helix folds within the CATH database, our approach successfully identifies new and unique folds. By evaluating the sequence-to-structure quality of these main chains, we observe that the DiffTopo-RFdiffusion framework can sample novel folds and generate high-quality main chain structures.

In addition to random sampling of main chain structures, we explored another application of DiffTopo—generating scaffolds for existing secondary structure motifs. Here, we attempted to design a new scaffold for the DBL1_03 protein binder that binds to PD-L1 (complex PDB id: 7XYQ) (Gainza et al., 2023). The A1-A20 helical portion of DBL1_03 forms the interacting interface with PD-L1. We developed a motif scaffolding pipeline (in Appendix A.1.4). The best-designed structure, predicted using Alphafold multimer, is shown in the right panel of Figure B.1 in

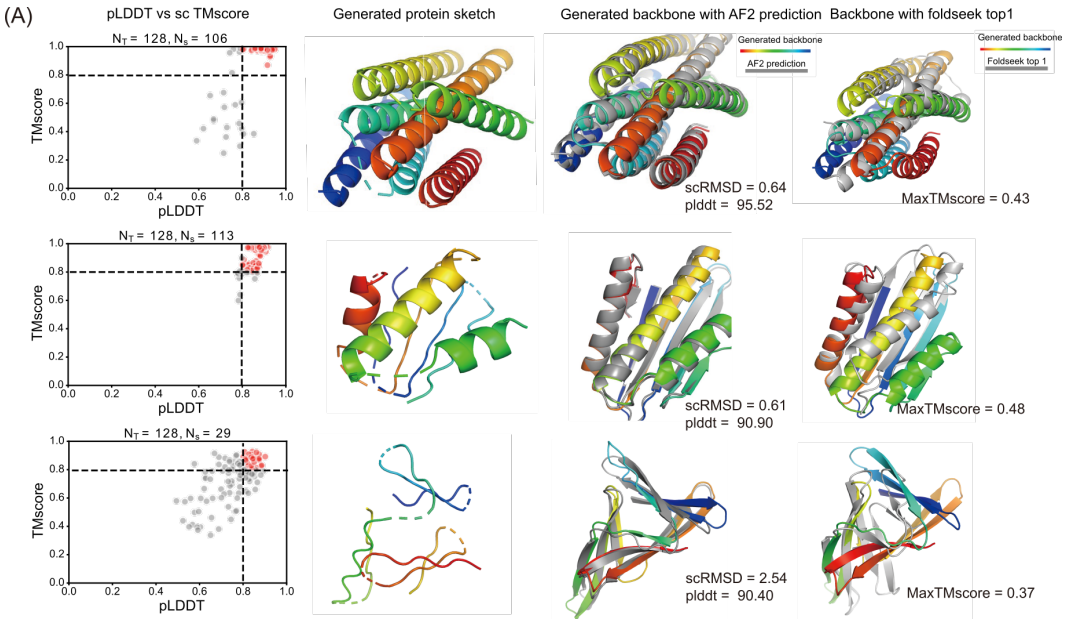

Figure 4: Explorations of novel folds using DiffTopo. Results for novel fold exploration in all-α(top), mixed α-β(middle) and all-β (bottom) folds. The plots show TM-scores between generated backbones and Alphafold2 predictions against predicted plDDT. Red points indicate sequences with high designability (TM-score $>$0.8, plDDT $>$80). The second column displays DiffTopo-generated protein sketches. The third column presents backbone alignment with Alphafold2 predictions, and the last column compare the closest structure found by Foldseek.

the Appendix. The predicted binding positions by Alphafold align well with the ground truth. This demonstrates the potential of our method in constructing scaffolds for tasks involving the creation of motifs with existing secondary structures.

4 Discussion

We show that a new representation termed CG topology is able to capture the three-dimensional features of protein folds correctly using very few data points, reducing the sampling space for protein design. With our generative diffusion model, DiffTopo, we not only learn the distribution of relative positions of secondary structures but also sample diverse topologies. By passing protein sketches to RFdiffusion, we are able to generate designable protein backbones in 3D space and explore the protein fold space with fewer degrees of freedom. Comparing with RFdiffusion only as benchmark, and using ProteinMPNN and AlphaFold2 as an orthogonal test, our framework improves the diversity of protein space exploration while still maintaining good designability. Our model also shows the capability in motif scaffolding problem when motifs with secondary structures are available.

Our results demonstrate that the DiffTopo+RFdiffusion framework is capable of designing new proteins and explore novel folds non-existent in nature. Compared to all other backbone atom level models like RFdiffusion and Framediff, our framework can massively sample topological space with different secondary structure arrangements in a short time. On the other hand, compared to Form and GENESIS(Harteveld et al., 2022), we can make protein topology sampling automatically instead of manually construct it. Our work provides a new perspective on protein structure representation and we believe there is much to explore in the space of this representation to improve performance in protein universe searching and protein design. For example, our method can be leveraged to generate customized protein backbones exhibiting highly regular patterns, similar to nanomaterials. Moreover, our approach holds the potential to create novel scaffolds for existing functional motifs, offering applications in de novo enzyme and antibody design.

REFERENCES

Namrata Anand and Tudor Achim. Protein Structure and Sequence Generation with Equivariant Denoising Diffusion Probabilistic Models, May 2022. URL http://arxiv.org/abs/2205.15019. arXiv:2205.15019 [cs, q-bio].

Namrata Anand, Raphael Eguchi, and Po-Ssu Huang. Fully differentiable full-atom protein backbone generation. April 2019. URL https://openreview.net/forum?id=SJxnVL8YOV.

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured Denoising Diffusion Models in Discrete State-Spaces, February 2023. URL http://arxiv.org/abs/2107.03006. arXiv:2107.03006 [cs].

J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, A. Courbet, R. J. de Haas, N. Bethel, P. J. Y. Leung, T. F. Huddy, S. Pellock, D. Tischer, F. Chan, B. Koepnick, H. Nguyen, A. Kang, B. Sankaran, A. K. Bera, N. P. King, and D. Baker. Robust deep learning-based protein sequence design using ProteinMPNN. *Science (New York, N.Y.)*, 378 (6615):49–56, October 2022. ISSN 1095-9203. doi: 10.1126/science.add2187.

Raphael R. Eguchi, Christian A. Choe, and Po-Ssu Huang. Ig-VAE: Generative modeling of protein structure by direct 3D coordinate generation. *PLOS Computational Biology*, 18(6):e1010271, June 2022. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1010271. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1010271. Publisher: Public Library of Science.

Fabian B. Fuchs, Daniel E. Worrall, Volker Fischer, and Max Welling. SE(3)-Transformers: 3D Roto-Translation Equivariant Attention Networks, November 2020. URL http://arxiv.org/abs/2006.10503. arXiv:2006.10503 [cs, stat].

Pablo Gainza, Sarah Wehrle, Alexandra Van Hall-Beauvais, Anthony Marchand, Andreas Scheck, Zander Harteveld, Stephen Buckley, Dongchun Ni, Shuguang Tan, Freyr Sverrisson, Casper Goverde, Priscilla Turelli, Charlène Raclot, Alexandra Teslenko, Martin Pacesa, Stéphane Rosset, Sandrine Georgeon, Jane Marsden, Aaron Petruzzella, Kefang Liu, Zepeng Xu, Yan Chai, Pu Han, George F. Gao, Elisa Oricchio, Beat Fierz, Didier Trono, Henning Stahlberg, Michael Bronstein, and Bruno E. Correia. De novo design of protein interactions with learned surface fingerprints. *Nature*, 617(7959):176–184, May 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-05993-x.

Zander Harteveld, Joshua Southern, Michaël Defferrard, Andreas Loukas, Pierre Vandergheynst, Micheal Bronstein, and Bruno Correia. Deep sharpening of topological features for de novo protein design. April 2022. URL https://openreview.net/forum?id=DwN81YIXGQP.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models, December 2020. URL http://arxiv.org/abs/2006.11239. arXiv:2006.11239 [cs, stat].

TM Jacobs, B Williams, T Williams, X Xu, A Eletsky, JF Federizon, T Szyperski, and B Kuhlman. Design of structurally distinct proteins using strategies inspired by evolution. *Science (New York, N.Y.)*, 352(6286):687–690, May 2016. ISSN 0036-8075. doi: 10.1126/science.aad8036. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4934125/.

Bowen Jing, Stephan Eismann, Patricia Suriana, Raphael J. L. Townshend, and Ron Dror. Learning from Protein Structure with Geometric Vector Perceptrons, May 2021. URL http://arxiv.org/abs/2009.01411. arXiv:2009.01411 [cs, q-bio, stat].

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, August 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2. URL https://www.nature.com/articles/s41586-021-03819-2. Number: 7873 Publisher: Nature Publishing Group.

Michael Knudsen and Carsten Wiuf. The CATH database. *Human Genomics*, 4(3):207–212, February 2010. ISSN 1473-9542. doi: 10.1186/1479-7364-4-3-207. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3525972/.

Brian Kuhlman and Philip Bradley. Advances in protein structure prediction and design. *Nature Reviews Molecular Cell Biology*, 20(11):681–697, November 2019. ISSN 1471-0080. doi: 10.1038/s41580-019-0163-x. URL https://www.nature.com/articles/s41580-019-0163-x. Number: 11 Publisher: Nature Publishing Group.

Yeqing Lin and Mohammed AlQuraishi. Generating Novel, Designable, and Diverse Protein Structures by Equivariantly Diffusing Oriented Residue Clouds, June 2023. URL http://arxiv.org/abs/2301.12485. arXiv:2301.12485 [cs, q-bio].

Shitong Luo, Yufeng Su, Xingang Peng, Sheng Wang, Jian Peng, and Jianzhu Ma. Antigen-Specific Antibody Design and Optimization with Diffusion-Based Generative Models for Protein Structures, October 2022. URL https://www.biorxiv.org/content/10.1101/2022.07.10.499510v5. Pages: 2022.07.10.499510 Section: New Results.

Craig O. Mackenzie, Jianfu Zhou, and Gevorg Grigoryan. Tertiary alphabet for the observable protein structural universe. *Proceedings of the National Academy of Sciences of the United States of America*, 113(47):E7438–E7447, November 2016. ISSN 1091-6490. doi: 10.1073/pnas.1607178113.

Shintaro Minami, Naohiro Kobayashi, Toshihiko Sugiki, Toshio Nagashima, Toshimichi Fujiwara, Rie Tatsumi-Koga, George Chikenji, and Nobuyasu Koga. Exploration of novel $\alpha\beta$-protein folds through de novo design. *Nature Structural & Molecular Biology*, pp. 1–9, July 2023. ISSN 1545-9985. doi: 10.1038/s41594-023-01029-0. URL https://www.nature.com/articles/s41594-023-01029-0. Publisher: Nature Publishing Group.

Xingjie Pan and Tanja Kortemme. Recent advances in de novo protein design: Principles, methods, and applications. *Journal of Biological Chemistry*, 296:100558, January 2021. ISSN 0021-9258. doi: 10.1016/j.jbc.2021.100558. URL https://www.sciencedirect.com/science/article/pii/S0021925821003367.

Victor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) Equivariant Graph Neural Networks, February 2022. URL http://arxiv.org/abs/2102.09844. arXiv:2102.09844 [cs, stat].

K. T. Simons, C. Kooperberg, E. Huang, and D. Baker. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *Journal of Molecular Biology*, 268(1):209–225, April 1997. ISSN 0022-2836. doi: 10.1006/jmbi.1997.0959.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-Based Generative Modeling through Stochastic Differential Equations, February 2021. URL http://arxiv.org/abs/2011.13456. arXiv:2011.13456 [cs, stat].

William R. Taylor, Gail J. Bartlett, Vijayalakshmi Chelliah, Daniel Klose, Kuang Lin, Tom Sheldon, and Inge Jonassen. Prediction of protein structure from ideal forms. *Proteins*, 70(4):1610–1619, March 2008. ISSN 1097-0134. doi: 10.1002/prot.21913.

Brian L. Trippe, Jason Yim, Doug Tischer, David Baker, Tamara Broderick, Regina Barzilay, and Tommi Jaakkola. Diffusion probabilistic modeling of protein backbones in 3D for the motif-scaffolding problem, March 2023. URL http://arxiv.org/abs/2206.04119. arXiv:2206.04119 [cs, q-bio, stat].

Michel van Kempen, Stephanie S. Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron L. M. Gilchrist, Johannes Söding, and Martin Steinegger. Fast and accurate protein structure search with Foldseek. *Nature Biotechnology*, pp. 1–4, May 2023. ISSN 1546-1696. doi: 10.1038/s41587-023-01773-0. URL https://www.nature.com/articles/s41587-023-01773-0. Publisher: Nature Publishing Group.

Joseph L. Watson, David Juergens, Nathaniel R. Bennett, Brian L. Trippe, Jason Yim, Helen E. Eisenach, Woody Ahern, Andrew J. Borst, Robert J. Ragotte, Lukas F. Milles, Basile I. M. Wicky, Nikita Hanikel, Samuel J. Pellock, Alexis Courbet, William Sheffler, Jue Wang, Preetham Venkatesh, Isaac Sappington, Susana Vázquez Torres, Anna Lauko, Valentin De Bortoli, Emile Mathieu, Sergey Ovchinnikov, Regina Barzilay, Tommi S. Jaakkola, Frank DiMaio, Minkyung Baek, and David Baker. De novo design of protein structure and function with RFdiffusion. *Nature*, pp. 1–3, July 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06415-8. URL https://www.nature.com/articles/s41586-023-06415-8. Publisher: Nature Publishing Group.

Jinrui Xu and Yang Zhang. How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics*, 26(7):889–895, April 2010. ISSN 1367-4803. doi: 10.1093/bioinformatics/btq066. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2913670/.

Jianyi Yang, Ivan Anishchenko, Hahnbeom Park, Zhenling Peng, Sergey Ovchinnikov, and David Baker. Improved protein structure prediction using predicted interresidue orientations. *Proceedings of the National Academy of Sciences*, 117(3):1496–1503, January 2020. doi: 10.1073/pnas.1914677117. URL https://www.pnas.org/doi/10.1073/pnas.1914677117. Publisher: Proceedings of the National Academy of Sciences.

Jason Yim, Andrew Campbell, Andrew Y. K. Foong, Michael Gastegger, José Jiménez-Luna, Sarah Lewis, Victor Garcia Satorras, Bastiaan S. Veeling, Regina Barzilay, Tommi Jaakkola, and Frank Noé. Fast protein backbone generation with SE(3) flow matching, October 2023a. URL http://arxiv.org/abs/2310.05297. arXiv:2310.05297 [q-bio].

Jason Yim, Brian L. Trippe, Valentin De Bortoli, Emile Mathieu, Arnaud Doucet, Regina Barzilay, and Tommi Jaakkola. SE(3) diffusion model with application to protein backbone generation, May 2023b. URL http://arxiv.org/abs/2302.02277. arXiv:2302.02277 [cs, q-bio, stat].

Yang Zhang and Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins*, 57(4):702–710, December 2004. ISSN 1097-0134. doi: 10.1002/prot.20264.

A Appendix

A.1 Method

A.1.1 Coarse-grained topology representation

Inspired by the Form representation (Taylor et al., 2008) and protein sketches (Harteveld et al., 2022), where protein structures are built by the assembly of their secondary structures, we adopt a simplified representation referred to as a "Coarse-grained topology (CG topology)". The topology reduces the intricate protein structure into a stacked arrangement of Secondary Structure Elements (SSEs), achieved by representing secondary structures with three carbon alpha centroids, as illustrated in Figure 1A. CG topology involves capturing the geometric position of helices and strands through centroids. For helices, centroids include the first and last four Cα atoms, as well as the total Cα atom centroid. Strands are represented by centroids of the first and last two Cα atoms, along with the overall Cα atom centroid within the strand. CG topology can also be easily represented as a protein sketch, which is a rough 3D approximation of a native protein structure with standard SSEs, lacking loops, and AA side chains. Compared with Form and protein sketch, this representation method of CGtopo gets rid of the concept of layers and has a higher freedom in structural representation to represent structures like beta barrels, and it is still simple with a small number of degrees of freedom.

A.1.2 Dataset

Using the CG topology approach we can easily convert a protein structure database, such as PDB, from standard protein models to CG topology structures. In this paper, all PDB structural data are sourced from CATH, a database organized as a classification of protein structures (Knudsen &

Wiuf, 2010). To mitigate the influence of structural redundancy on the data distribution, we employ the CATH-dataset-nonredundant S40 for training data, yielding 31,886 non-redundant protein structures. 90% of structures are used for training and 10% rest are used for independent validation.

A.1.3 DIFFTOPO+RFDIFFUSION FRAMEWORK

First, we introduce DiffTopo, a Equivariant diffusion model for generating CG topology conditioned on SSEs strings. Based on previous works on denoising diffusion (Song et al., 2021; Ho et al., 2020; Austin et al., 2023), given a data point sampled from a real data distribution $z_0 \sim q(\mathbf{z})$, the forward diffusion process is defined by adding Gaussian noise gradually to the data point z_0 form corrupted samples z_t. For one CG topology data z_0, there are three points x and corresponding secondary structure type u. At time step $t = 0, \ldots, \mathrm{T}$, the conditional distribution of the intermediate data state z_t given the previous state is defined by the multivariate normal distribution,

$$q\left(z_t \mid z_{t-1}\right) = \mathcal{N}\left(z_t; \sqrt{1-\beta_t} z_{t-1}, \beta_t I\right)$$

The step sizes are controlled by a variance schedule $\{\beta_t \in (0,1)\}_{t=1}^{T}$. The process is constructed to be Markovian and if we let $\alpha_t = 1 - \beta_t$ and $\overline{\alpha_t} = \prod_{i=1}^{t} \alpha_i$, we could obtain the distribution of z_t given x

$$q\left(z_t \mid z_0\right) = \mathcal{N}\left(z_t; \sqrt{\overline{\alpha_t}} z_0, \left(1-\overline{\alpha_t}\right) I\right)$$

For the backward process, we need to learn a model p to approximate these conditional probabilities.

$$p\left(z_{t-1} \mid z_t, u\right) = \mathcal{N}\left(z_{t-1}; \mu\left(z_t, t, u\right), \Sigma\left(z_t, t\right)\right)$$

$\mu_\theta\left(x_t, t, u\right)$ is predicted by the neural network. In this paper function μ that predicts noise ϵ_θ is implemented as geometric vector perceptrons (Jing et al., 2021). The input to GVPs is the noised version of the point coordinates x_t and point feature h_t at time t and context u. Note that the predicted noise ϵ includes coordinate and feature components, $\epsilon = [\epsilon_{\mathbf{x}},\ \epsilon_{\mathbf{h}}]$. The predicted noise could be calculated by

$$\epsilon_\theta\left(x_t, t\right) = GVP\left(x_t, h_t, u, t\right) - \left[x_t, h_t\right]$$

Following the generation of the CG topology via the diffusion process, utilizing CG topology allows us to straightforwardly obtain the positions and lengths of standard Secondary Structure Elements (SSEs). Since CG topology retains a certain level of fuzziness, we need a backbone level model to systematically search for the designable backbone corresponding to CG topology. Here, we leverage RFdiffusion, a highly effective backbone generation model to generate designable backbones from CG topology. Initially, we construct a protein sketch from standard SSEs based on CG topology and utilize the motif modeling function inherent in RFdiffusion to directly connect the SSEs. Subsequently, we employ the partial diffusion approach to identify a reasonable backbone structure within this interconnected structure. Then we employ the Protein Message Passing Neural Network (ProteinMPNN) (Dauparas et al., 2022) for fixed backbone sequence design algorithm to obtain amino acid sequences. These sequences are then input into the structure prediction algorithm Alphafold2 (Jumper et al., 2021) to predict the structure, which is use to validate the designability of generated backbones.

A.1.4 MOTIF SCCAFOLD MODELING

First, we extract the initial helix combined with PD-L1 in DBL_03 and convert it to CGtopo. Subsequently, we employ Difftopo's conditional generation protocol to sample the entire scaffold's CG-topo, using the condition 'HEHEHE' while keeping the first helix unchanged. The sampled CGtopo is then constructed into a protein sketch, with the first helix replaced by a functional motif.

Next, RFdiffusion's fixcontig protocol is utilized to connect the loops, followed by the partial diffusion protocol to optimize the entire backbone while preserving the first 20 amino acids. Lastly, ProteinMPNN is employed to design sequences for positions other than the first 20 amino acids. Alphafold is then used to predict the monomer structure, and Alphafold multimer is applied to predict the complex structure.

B SUPPLEMENTARY FIGURES

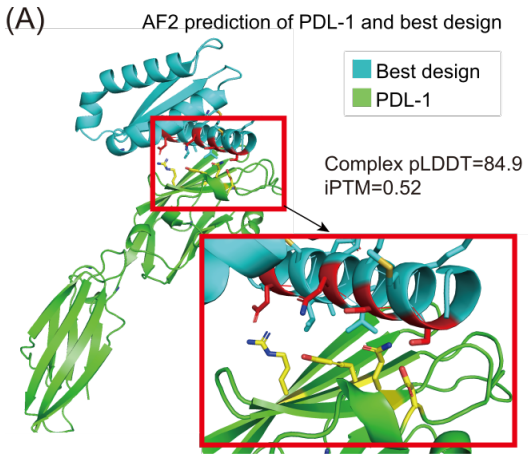

Figure B.1: Motif scaffolding application. (A) Prediction of multimer structure of best design and PD-L1. Zoomed-in part shows the AlphaFold2-predicted interface, with red residues representing the functional motif binding to PD-L1.