FrameDiPT: SE(3) Diffusion Model for Protein Structure Inpainting

Cheng Zhang InstaDeep
Adam Leach InstaDeep
Thomas Makkink InstaDeep
Miguel Arbesú InstaDeep
Ibtissem Kadri InstaDeep

Daniel Luo InstaDeep
Liron Mizrahi InstaDeep
Sabrine Krichen InstaDeep
Maren Lang BioNTech
Andrey Tovchigrechko BioNTech

Nicolas Lopez Carranza InstaDeep
Uğur Şahin BioNTech
Karim Beguir InstaDeep
Michael Rooney BioNTech
Yunguan Fu InstaDeep

Abstract

Protein structure prediction field has been revolutionised by deep learning with protein folding models such as AlphaFold 2 and ESMFold. These models enable rapid *in silico* prediction and have been integrated into *de novo* protein design and protein-protein interaction (PPI) prediction. However, biologically relevant features dependent on conformational distributions cannot be estimated with these models. Diffusion models, a novel class of generative models, have been developed to learn conformational distributions and applied to *de novo* protein design. Limited work has been done on protein structure inpainting, where a masked section is recovered by simultaneously conditioning on its sequence and the rest of the structure. In this work, we propose **FrameD**iff **inP**ain**T**ing (FrameDiPT), a generalised model for protein inpainting. This is important for T-cells given the hyper-variability of the complementarity determining region (CDR) loops. We evaluated the model on CDR loop design for T-cell receptors and achieved comparable prediction accuracy to ProteinGenerator and RFdiffusion with limited training data and learnable parameters. Different from deterministic structure prediction models, FrameDiPT captures the conformational distribution at different regions and binding states, highlighting a key advantage of generative models.

1 Introduction

Proteins play an essential part in almost all cellular processes. Accurate modelling of protein structure is important to assess the behaviour of existing and *de novo* proteins. While models such as AlphaFold 2 [Jumper et al., 2021, Evans et al., 2021], RoseTTAFold [Baek et al., 2021] and ESMFold [Lin et al., 2022] have revolutionised computational protein modelling with high-quality predictions, their deterministic nature does not reflect the dynamic nature of proteins. Diffusion denoising models [Sohl-Dickstein et al., 2015, Ho et al., 2020], a novel class of generative models, have achieved superior performance in image synthesis. RFdiffusion [Watson et al., 2023] integrated the diffusion model with a pre-trained RoseTTAFold to perform *de novo* protein design, motif scaffolding and binder design. While recent works have leveraged diffusion models for conformational distribution tasks by training on molecular force fields [Abdin and Kim, 2023, Zheng et al., 2023], limited work has been done on protein structure inpainting tasks where only a subset of the residues are of interest. By fixing the majority of residue positions, we sample the conformational distribution in the area of interest while avoiding incorrect global structure predictions.

Machine Learning for Structural Biology Workshop, NeurIPS 2023.

In this work, we focus on T-cell receptors (TCRs) and peptide-major histocompatibility complexes (pMHCs), which are crucial for cell-mediated immunity. The complementarity determining region (CDR) loops of TCRs, especially the CDR3 loops, are highly variable and thereby able to bind with different pMHCs [Minguet et al., 2007, Xu et al., 2020]. Understanding the conformational distribution is therefore beneficial to downstream tasks such as TCR maturation and binding prediction. We choose to model the distribution of CDR3 loops and keep the rest of the structure fixed as contextual information. This differs from DiffAB [Luo et al., 2022] which designs antibody CDR sequences and structures given antibody-antigen frameworks, since, in our inpainting task, the amino acid sequences are known. We extended FrameDiff [Yim et al., 2023], an SE(3) diffusion model for *de novo* protein backbone generation, to generic protein structure inpainting, a model we term **FrameD**iff in**P**ain**T**ing (FrameDiPT). After training for 2 GPU-weeks on 32K monomer structures, an 18M parameter FrameDiPT model reached satisfying prediction accuracy compared to deterministic models including AlphaFold 2, ESMFold as well as diffusion-based models such as RFdiffusion and ProteinGenerator [Lisanza et al., 2023]. Importantly, the compared methods were trained on larger datasets including TCR-like structures, while FrameDiPT training intentionally excluded all structures similar to TCRs and antibodies. FrameDiPT also captured conformational distribution differences at different regions and binding states.

2 Related work

Diffusion models [Sohl-Dickstein et al., 2015, Ho et al., 2020] were first applied to image generation, where images are synthesised via progressive denoising. Different from images which have no limitation in sampling space, molecules and proteins have intrinsic geometric constraints on bond angles and lengths. AlphaFold 2 [Jumper et al., 2021] represents each residue as a rigid frame thus preventing structural violations. Similarly, heavy atom side chain positions can be described by dihedral angles in lieu of bond lengths [Baek et al., 2021]. With these rigid representations, diffusion models on the Riemannian manifolds SO(2) and SO(3) [Huang et al., 2022, Leach et al., 2022] are applied to proteins to ensure plausible structure sampling.

While small molecule designs [Shi et al., 2021] model all atoms, most protein diffusion models focus on *de novo* protein design where only backbone atoms [Watson et al., 2023, Yim et al., 2023, Wu et al., 2022] are generated, reducing the model complexity and computational costs. Full-atom structures can be obtained by sampling a sequence via inverse folding models such as ProteinMPNN [Dauparas et al., 2022], and then predicting structure using protein folding models. Alternatively, full-atom generation models such as cg2all [Heo and Feig, 2023] can be used to generate full structures based on coarse-grain representations. Other works such as Chroma [Ingraham et al., 2022], which attempts to preserve the average bond angle and radius of gyration through an anisotropic diffusion process, and EigenFold [Jing et al., 2023] which performs diffusion on a harmonic decomposition of the protein chain instead of directly in the coordinate space of residues, attempt to integrate physical priors into the diffusion processes. Some work also attempts to learn sequence-structure joint distributions and thus performs protein sequence-structure co-design [Zhang et al., 2023, Anand and Achim, 2022, Luo et al., 2022, Chu et al., 2023]. However, there is limited research focusing on the task of protein structure inpainting. Wang et al. [2022] proposed a deep learning model for scaffolding protein functional sites, however, with a deterministic model. RFdiffusion targets various tasks including motif scaffolding but relies on ProteinMPNN to generate sequence. Similarly, Gao et al. [2023] proposed a language diffusion model DiffSDS for unknown-sequence protein backbone inpainting.

3 Method

3.1 FrameDiff

Yim et al. [2023] introduced FrameDiff, a graph-based SE(3)-equivariant neural network, which achieves comparable performance to RFdiffusion for *de novo* protein backbone design with much less training data and without a pre-trained structure prediction network. FrameDiff uses the backbone rigid-frame representation $\mathbf{T}$ with an additional torsion angle ψ to determine the position of the backbone carbonyl oxygen atom. The forward diffusion process is performed on the rigid-frame representation, i.e. the translation $\mathbf{X}$ given by the coordinates of the C_α atom and the rotation $\mathbf{R}$ defined by the frame formed by the $N - C_\alpha - C$ atoms for each residue. The torsion angle ψ is not involved in the diffusion process but is predicted by the model. The backward diffusion process is

defined by denoising score matching [Vincent, 2011]. Training losses consist of both the rigid-frame losses and the atom-level losses. We refer the readers to Yim et al. [2023] for further details.

3.2 FrameDiPT

We extend FrameDiff to **FrameD**iff in**P**ain**T**ing (FrameDiPT) for protein structure inpainting with the following modifications.

Randomly mask a contiguous region for diffusion For all training monomers, a contiguous region of 8-50 amino acids is randomly selected and the diffusion process is applied only in this region. For the input node features, the diffusion timestep `t` is set to 10^{-5} in the fixed region to indicate the residues are not being diffused. For convenience, we call the randomly selected contiguous region the "diffused region" and the remaining part of the structure the "context region".

Add amino acid types as node features In contrast to *de novo* protein design, the amino acid sequence is given in the inpainting task. An extra node feature `aatype` is concatenated to the original node features of FrameDiff. For a protein with N_{res} residues, the `aatype` node feature is of shape (N_{res}, 21) containing 20 standard amino acid types and 1 for unknown amino acid type. Similarly, edge features are modified accordingly.

Training loop During the training loop, the diffusion noising process is only performed in the diffused region. The resulting noised structure is then given as input to the model to predict the original structure. Loss is only computed over the diffused region. Different from FrameDiff, we train the model on clustered data with a sequence similarity threshold of 90%. In each epoch, only one structure is randomly sampled from each cluster. This approach outperforms FrameDiPT trained with non-clustered data (Appendix B).

Inference loop In the inference loop, the context region along with the sampled random noise in the diffused region are given as the initial state. In each inference step, the model predicts the backbone rigid frames $\hat{\mathbf{T}}^{(0)}$ from $\mathbf{T}^{(t)}$, where $\mathbf{T}^{(t)}$ represents the rigid frames at time t and $t = 0$ corresponds to the ground truth. A reverse diffusion step is performed to get $\mathbf{T}^{(t-dt)}$ from $\mathbf{T}^{(t)}$ and $\hat{\mathbf{T}}^{(0)}$ where dt is the step size. Then the context region in $\mathbf{T}^{(t-dt)}$ is replaced by the original structure, to ensure the context region stays fixed. Finally, we extended our model to run inference on multimers by adding a residue gap of 200 between different chains, following the trick used in Motmaen et al. [2023].

4 Experiment setting

Training We train FrameDiPT using data from RCSB Protein Data Bank (RCSB PDB)[1] (Appendix A) with the same hyperparameters as FrameDiff. It was trained for 2 weeks with 1 NVIDIA A100 GPU for 95 epochs with a length batching strategy. Each batch contains a collection of different diffused instances of the same backbone structure, and the number of samples per batch adapts to the sequence length of the structure with a maximum batch size of 128.

Evaluation A TCR and TCR:pMHC dataset has been curated for evaluation (Appendix C). It contains three splits: 21 unbound TCR structures; 62 TCR:pMHC class I complexes; and 18 TCR:pMHC class II complexes. For each TCR or TCR:pMHC sample, the CDR3 loop in both TCR alpha and beta chains are masked as the diffused region and 100 inference steps are performed to get the final prediction. `cg2all`[2] is used to convert the predicted structure to an all-atom structure. We report the root-mean-square-deviation (RMSD) on the backbone and full-atom structure for evaluation. For diffusion models, we developed a sample selection strategy using kernel density estimation. A Gaussian kernel with standard deviation of 30Å is fitted over the carbon alpha coordinates of the inpainted region to estimate density. The sample corresponding to the highest density is used as a proxy for the "most-likely" sample.

Baseline diffusion models ProteinGenerator performs diffusion in sequence space and structure is predicted via RoseTTAFold given the generated sequence. We adapted ProteinGenerator to the inpainting task by supplying the input sequence. RFdiffusion is used by masking out both the structure and sequence of TCR CDR3 loops, thus full-atom RMSD is not applicable. Empirically, we found

[1] `https://www.rcsb.org/docs/programmatic-access/file-download-services`

[2] `https://github.com/huhlim/cg2all`

providing sequences to RFdiffusion deteriorated the performance since the model was not trained in this mode.

5 Results and discussions

Table 1: Backbone and full-atom RMSD comparison. A signed Wilcoxon paired two-sided rank statistical test between FrameDiPT and ProteinGenerator is performed at significance level p-value < 0.05. Underline means significantly better than ProteinGenerator which outperforms RFdiffusion.

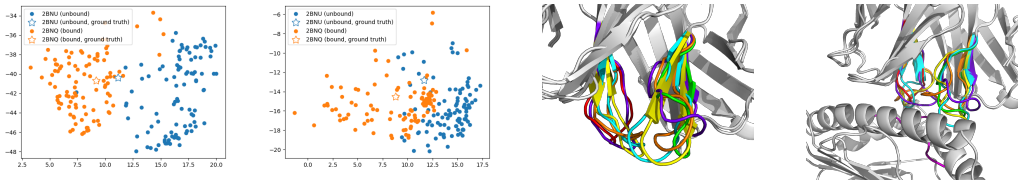
Method	RMSD	TCR	TCR:pMHC-I	TCR:pMHC-II
ProteinGenerator	Backbone	2.87 ± 0.48	2.34 ± 0.73	3.01 ± 0.55
	Full-atom	3.58 ± 0.62	3.24 ± 0.79	$\mathbf{3.59 \pm 0.39}$
RFdiffusion	Backbone	3.23 ± 0.93	3.32 ± 0.72	3.81 ± 0.91
FrameDiPT	Backbone	$\mathbf{2.70 \pm 0.43}$	$\mathbf{2.18 \pm 0.45}$	$\mathbf{2.91 \pm 0.54}$
	Full-atom	$\mathbf{3.48 \pm 0.52}$	$\mathbf{2.93 \pm 0.59}$	3.90 ± 0.46

Comparison to protein diffusion models We compared FrameDiPT with existing diffusion models, ProteinGenerator and RFdiffusion, for inpainting tasks (Table 1). The original FrameDiff is not applicable as it does not take any sequential or structural information as input. Across five samples, FrameDiPT achieved median RMSD of 2.70Å, 2.18Å, and 2.91Å on unbound TCR, bound TCR with pMHC-I, and pMHC-II, respectively. This performance is comparable with ProteinGenerator and RFdiffusion which have larger networks and have been trained on larger datasets (Appendix D). A signed Wilcoxon paired two-sided rank test was performed which indicated no statistically significant difference between results. Importantly, the training of FrameDiPT explicitly excluded TCR:pMHC and antibody structures (Appendix A), demonstrating the strong generalization capacity of the proposed method. The RMSD is further reduced across all datasets when generating 25 samples (Table 5). This indicates that FrameDiPT is capable of sampling structures that are specified in crystal structures. It is important to highlight that, although the structures were evaluated using RMSD, RMSD to X-ray structure is not the perfect metric as the crystalised conformation represents only a snapshot of the dynamic.

Capturing conformational distributions While CDR3 loops are highly flexible, the N- and C-terminal flank regions should be more constrained, as they are mainly beta strands. We thus performed inpainting on these flanks of the same length as the CDR3 loop to compare the variance of generated samples, which is defined as the average inter-sample backbone RMSD (Table 2). Lower backbone RMSD and sample variance were observed in the flanks. Notably, the C-terminal flank has a smaller variance than N-terminal, correctly reflecting how the C-terminus fully encompasses a beta strand while the N-terminal flank starts on the small loop leading the beta strand before the CDR3 loop. The correlation between normalised carbon-alpha B-factors and sampling variances (Figure 9b) is 0.366. The moderate correlation is not surprising, as B-factors aggregate different effects from experimental uncertainty to structural flexibility in a variable manner between structures.

Table 2: Backbone RMSD and sample variance of different diffused regions. N- and C-terminal flanks have lower RMSD and variance, consistent with structural properties of the diffused regions.

Diffused region	Metric	TCR	TCR:pMHC-I	TCR:pMHC-II
CDR3	Backbone RMSD	2.70 ± 0.43	2.18 ± 0.45	2.91 ± 0.54
	Variance	1.60 ± 0.20	1.58 ± 0.34	1.85 ± 0.34
N-terminal flank	Backbone RMSD	0.89 ± 0.23	1.59 ± 0.48	1.71 ± 0.39
	Variance	0.92 ± 0.37	0.96 ± 0.23	0.93 ± 0.30
C-terminal flank	Backbone RMSD	0.71 ± 0.12	0.69 ± 0.08	0.76 ± 0.14
	Variance	0.26 ± 0.10	0.32 ± 0.13	0.38 ± 0.10



(a) CDR3α PCA (b) CDR3β PCA (c) PDB 2BNU (d) PDB 2BNQ

Figure 1: Plots of the first two principal components of PCA describing a) CDR3α and b) CDR3β C_α position loop conformations with ground truth marked as a star. A clear distinction between samples drawn from the unbound (2BNU) and bound (2BNQ) conformations can be seen on the alpha chain. A clear difference in modes can be seen on the beta chain. c) 2BNU and d) 2BNQ visualizations with context structure, ground truth CDR3 alpha, ground truth CDR3 beta and peptide; FrameDiPT predictions in cyan, yellow and orange alongside ESMFold prediction.

Conformation changes upon binding The CDR conformations differ in the bound and unbound states [Armstrong et al., 2008]. This is due to intermolecular forces and steric constraints induced by the pMHC complex influencing the position of loop residues. Pairs of structures (6 for pMHC-I and 2 for pMHC-II) are found that represent unbound and bound states of the same TCRs. For each structure, 100 samples were generated, with inpainting applied to the CDR3 loops on the alpha and beta chains. Figure 1 shows a clear separation between unbound and bound samples in terms of CDR3 loop conformations for 2BNU and 2BNQ, with significant differences between distribution centroids. More examples can be found in Appendix E.3. This underlines the strength of diffusion models in capturing different conformations, which could be useful for downstream binding classification tasks. For example, sampling loop conformations of different or mutated bound TCR:pMHC complexes and evaluating them through energy-scoring methods could help to discriminate weak and strong binders.

Quantifying uncertainty Figure 9a shows the correlation between the median backbone RMSD and the variance of generated samples. Interestingly, when the sampled structures are different between them, the samples themselves may be inaccurate. The variance can thus be used as a metric to analyse the overall quality of generated samples for a test instance.

Comparison to deterministic protein folding models FrameDiPT has been compared with pre-trained deterministic protein folding models such as AlphaFold 2 (Table 7). AlphaFold 2 with custom templates where CDR3 loops were masked had lower RMSD to the ground truth crystal structures. The difference to FrameDiPT is significant, indicating further room for improvement.

6 Conclusion

In this work, we proposed a novel inpainting task for protein structure generation. We trained the proposed FrameDiff inpainting (FrameDiPT) model on 32K monomer structures and evaluated it on TCR CDR3 loop design. With no TCR and antibody structures present in the training data and only 18M parameters, FrameDiPT achieved similar RMSD to other TCR-aware pre-trained large diffusion models such as ProteinGenerator and RFdiffusion. Despite only being trained on monomers, FrameDiPT could capture the conformational distribution of the diffused region and the TCR:pMHC binding interaction. While FrameDiPT is able to sample structures close to the crystal structures, the RMSD remains significantly higher than AlphaFold 2 which has a larger network and larger training set. In the future, FrameDiPT could be improved with more training data and scaling up the network to close the gap. Moreover, downstream applications such as TCR:pMHC binding classification can be considered.

References

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, 2021. doi:10.1038/s41586-021-03819-2.

Richard Evans, Michael O'Neill, Alexander Pritzel, Natasha Antropova, Andrew Senior, Tim Green, Augustin Žídek, Russ Bates, Sam Blackwell, Jason Yim, Olaf Ronneberger, Sebastian Bodenstein, Michal Zielinski, Alex Bridgland, Anna Potapenko, Andrew Cowie, Kathryn Tunyasuvunakool, Rishub Jain, Ellen Clancy, Pushmeet Kohli, John Jumper, and Demis Hassabis. Protein complex prediction with alphafold-multimer. bioRxiv, 2021. doi:10.1101/2021.10.04.463034. URL https://www.biorxiv.org/content/early/2021/10/04/2021.10.04.463034.

Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N. Kinch, R. Dustin Schaeffer, Claudia Millán, Hahnbeom Park, Carson Adams, Caleb R. Glassman, Andy DeGiovanni, Jose H. Pereira, Andria V. Rodrigues, Alberdina A. van Dijk, Ana C. Ebrecht, Diederik J. Opperman, Theo Sagmeister, Christoph Buhlheller, Tea Pavkov-Keller, Manoj K. Rathinaswamy, Udit Dalwadi, Calvin K. Yip, John E. Burke, K. Christopher Garcia, Nick V. Grishin, Paul D. Adams, Randy J. Read, and David Baker. 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.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. bioRxiv, 2022.

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. arXiv:1503.03585v8, 2015.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. arXiv:2006.11239v2, 2020.

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, 620(7976):1089–1100, Aug 2023. ISSN 1476-4687. doi:10.1038/s41586-023-06415-8. URL https://doi.org/10.1038/s41586-023-06415-8.

Osama Abdin and Philip M. Kim. Pepflow: direct conformational sampling from peptide energy landscapes through hypernetwork-conditioned diffusion. bioRxiv, 2023. doi:10.1101/2023.06.25.546443. URL https://www.biorxiv.org/content/early/2023/06/26/2023.06.25.546443.

Shuxin Zheng, Jiyan He, Chang Liu, Yu Shi, Ziheng Lu, Weitao Feng, Fusong Ju, Jiaxi Wang, Jianwei Zhu, Yaosen Min, He Zhang, Shidi Tang, Hongxia Hao, Peiran Jin, Chi Chen, Frank Noé, Haiguang Liu, and Tie-Yan Liu. Towards predicting equilibrium distributions for molecular systems with deep learning. arXiv:2306.05445v1, 2023.

Susana Minguet, Mahima Swamy, Balbino Alarcón, Immanuel F. Luescher, and Wolfgang W.A. Schamel. Full activation of the t cell receptor requires both clustering and conformational changes at CD3. Immunity, 26(1):43–54, January 2007. doi:10.1016/j.immuni.2006.10.019. URL https://doi.org/10.1016/j.immuni.2006.10.019.

Xinyi Xu, Hua Li, and Chenqi Xu. Structural understanding of t cell receptor triggering. Cellular & Molecular Immunology, 17(3):193–202, February 2020. doi:10.1038/s41423-020-0367-1. URL https://doi.org/10.1038/s41423-020-0367-1.

Shitong Luo, Yufeng Su, Xingang Peng, Sheng Wang, Jian Peng, and Jianzhu Ma. 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/2022/07/11/2022.07.10.499510.

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. arXiv preprint arXiv:2302.02277v3, 2023.

Sidney Lyayuga Lisanza, Jake Merle Gershon, Sam Tipps, Lucas Arnoldt, Samuel Hendel, Jeremiah Nelson Sims, Xinting Li, and David Baker. Joint generation of protein sequence and structure with rosettafold sequence space diffusion. bioRxiv, 2023. doi:10.1101/2023.05.08.539766. URL https://www.biorxiv.org/content/early/2023/05/10/2023.05.08.539766.

Chin-Wei Huang, Milad Aghajohari, Avishek Joey Bose, Prakash Panangaden, and Aaron Courville. Riemannian diffusion models. arXiv:2208.07949v1, 2022.

Adam Leach, Sebastian M Schmon, Matteo T. Degiacomi, and Chris G. Willcocks. Denoising diffusion probabilistic models on SO(3) for rotational alignment. In ICLR 2022 Workshop on Geometrical and Topological Representation Learning, 2022. URL https://openreview.net/forum?id=BY88eBbkpe5.

Chence Shi, Shitong Luo, Minkai Xu, and Jian Tang. Learning gradient fields for molecular conformation generation. arXiv:2105.03902v3, 2021.

Kevin E. Wu, Kevin K. Yang, Rianne van den Berg, James Y. Zou, Alex X. Lu, and Ava P. Amini. Protein structure generation via folding diffusion. arXiv:2209.15611v2, 2022.

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. bioRxiv, 2022. doi:10.1101/2022.06.03.494563. URL https://www.biorxiv.org/content/early/2022/06/04/2022.06.03.494563.

Lim Heo and Michael Feig. One particle per residue is sufficient to describe all-atom protein structures. bioRxiv, 2023. doi:10.1101/2023.05.22.541652. URL https://www.biorxiv.org/content/early/2023/05/23/2023.05.22.541652.

John Ingraham, Max Baranov, Zak Costello, Vincent Frappier, Ahmed Ismail, Shan Tie, Wujie Wang, Vincent Xue, Fritz Obermeyer, Andrew Beam, and Gevorg Grigoryan. Illuminating protein space with a programmable generative model. bioRxiv, 2022. doi:10.1101/2022.12.01.518682. URL https://www.biorxiv.org/content/early/2022/12/02/2022.12.01.518682.

Bowen Jing, Ezra Erives, Peter Pao-Huang, Gabriele Corso, Bonnie Berger, and Tommi Jaakkola. Eigenfold: Generative protein structure prediction with diffusion models. arXiv:2304.02198v1, 2023.

Zuobai Zhang, Minghao Xu, Aurélie Lozano, Vijil Chenthamarakshan, Payel Das, and Jian Tang. Pre-training protein encoder via siamese sequence-structure diffusion trajectory prediction. arXiv:2301.12068v2, 2023.

Namrata Anand and Tudor Achim. Protein structure and sequence generation with equivariant denoising diffusion probabilistic models. arXiv:2205.15019v1, 2022.

Alexander E. Chu, Lucy Cheng, Gina El Nesr, Minkai Xu, and Po-Ssu Huang. 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.

Jue Wang, Sidney Lisanza, David Juergens, Doug Tischer, Joseph L. Watson, Karla M. Castro, Robert Ragotte, Amijai Saragovi, Lukas F. Milles, Minkyung Baek, Ivan Anishchenko, Wei Yang, Derrick R. Hicks, Marc Expòsit, Thomas Schlichthaerle, Jung-Ho Chun, Justas Dauparas, Nathaniel Bennett, Basile I. M. Wicky, Andrew Muenks, Frank DiMaio, Bruno Correia, Sergey Ovchinnikov, and David Baker. 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.

Zhangyang Gao, Cheng Tan, and Stan Z. Li. Diffsds: A language diffusion model for protein backbone inpainting under geometric conditions and constraints. arXiv:2301.09642v1, 2023.

Pascal Vincent. A connection between score matching and denoising autoencoders. Neural Computation, 23(7):1661–1674, 2011. doi:10.1162/NECO_a_00142.

Amir Motmaen, Justas Dauparas, Minkyung Baek, Mohamad H. Abedi, David Baker, and Philip Bradley. Peptide-binding specificity prediction using fine-tuned protein structure prediction networks. Proceedings of the National Academy of Sciences, 120(9):e2216697120, 2023. doi:10.1073/pnas.2216697120. URL https://www.pnas.org/doi/abs/10.1073/pnas.2216697120.

Kathryn M. Armstrong, Kurt H. Piepenbrink, and Brian M. Baker. Conformational changes and flexibility in T-cell receptor recognition of peptide–MHC complexes. Biochemical Journal, 415 (2):183–196, 09 2008. ISSN 0264-6021. doi:10.1042/BJ20080850. URL https://doi.org/10.1042/BJ20080850.

Jinwoo Leem, Saulo HP de Oliveira, Konrad Krawczyk, and Charlotte M Deane. STCRDab: the structural T-cell receptor database. Nucleic Acids Research, 46(D1):D406–D412, 10 2017. ISSN 0305-1048. doi:10.1093/nar/gkx971. URL https://doi.org/10.1093/nar/gkx971.

A Training data

Data used to train FrameDiPT model was downloaded from RCSB PDB with the following data cleaning procedure. Only X-ray structures were kept, i.e. the structures from non-X-ray assays, or ModelArchive[3] and the AlphaFold 2 predicted structures were removed. The training data cleaning process also included the removal of structures belonging to any of the following categories: 1) have only non-standard residues; 2) have a resolution larger than 9Å; 3) have a single amino acid accounting for more than 80% of the structure; 4) have more than 4950 residues. In the end, any structures that could not be parsed by `biopython`[4] were removed.

The data cleaning process involved also removing samples that are similar to the TCR test data to avoid data leakage. Similar samples to TCR test data are identified using the 70% sequence similarity clusters. A cluster is considered as "leaking" if it contains any chain from the TCR test data. Afterwards, all samples from "leaking" clusters are removed from the training set. Such removal ensures that structures in the training set have a maximum sequence similarity of 70% compared to any structure in the TCR test data.

We followed the same data processing procedure as in the original FrameDiff, which leads to 32K monomers for training and the training strategy on clustered data leads to 9K clusters for the 32K monomers.

B Training strategy

The training on clustered monomers with 9K clusters is evaluated against baseline training on all 32K monomers for both *de novo* protein design model FrameDiff and inpainting model FrameDiPT. Figure 2 shows self-consistency RMSD, which is the RMSD between the generated backbone and ESMFold predictions of the ProteinMPNN generated sequences, for different designed lengths. The model trained on clustered data shows consistently better results than the baseline model. Table 3 shows median backbone RMSD on CDR3 loop design where better performance is observed with training on clustered data.

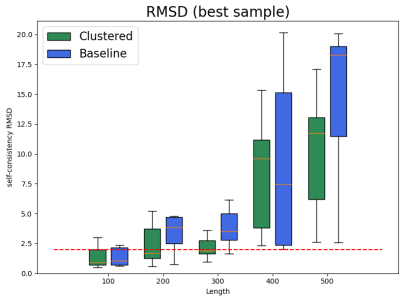


Figure 2: Best sample self-consistency RMSD of *de novo* protein design with the baseline model and the model trained on clustered data. The latter shows consistently better performance for all designed lengths.

Table 3: Backbone RMSD comparison between baseline training and training on clustered data. A signed Wilcoxon paired two-sided rank statistical test between baseline and clustered training is performed at significance level p-value < 0.05. Underline means significantly better.

	Baseline	Clustered
TCR	2.77 ± 0.47	$\mathbf{2.70 \pm 0.43}$
TCR:pMHC-I	2.86 ± 0.74	$\underline{\mathbf{2.18 \pm 0.45}}$
TCR:pMHC-II	3.15 ± 0.46	$\underline{\mathbf{2.91 \pm 0.54}}$

C Evaluation data

Curated sets of high-resolution, annotated structures of TCRs and TCR:pMHC complexes without any other companion proteins were assembled and fetched from the RCSB. First, the lists of PDB IDs for the different dataset types were fetched from the Structural T-Cell Receptor Database (STCRDab) [Leem et al., 2017]. The corresponding structures were downloaded from the RCSB and only X-ray structures with resolution < 3.5Å were kept. Then, the PDBe REST API was used to map

[3] `https://www.modelarchive.org/`

[4] `https://biopython.org/`

the structures' chains to UniProt IDs, when available. The UniProt metadata was used to label the chains in each structure (e.g. TCR alpha or beta chain, peptide, MHC alpha or beta chain) based on keyword and gene name matching. Other proteins or not annotated ones were flagged as such. Only structures with the expected TCR, peptide, or MHC chains[5] were kept; structures containing other proteins, or unlabelled ones were filtered out.

D Experiment settings

Experiment settings including number of model parameters, training time and training data of different models are summarised in Table 4. As ProteinGenerator and RFdiffusion are fine-tuned RoseTTAFold model, the experiment setting for RoseTTAFold is reported. FrameDiPT model has significantly fewer parameters and is trained with much less data and training time.

Table 4: Experiment setting comparison

Method	Params	Training Time		Training Data	
		Device	Time	Size	Description
AlphaFold$^{\alpha}$	93M	128 TPU	2 weeks	10M	PDB+self-distillation
AlphaFold$^{\beta}$	93M	128 TPU	11 days	40K+350K	PDB+self-distillation
ESMFold	3B	/	/	>12M	PDB+UniRef50
RoseTTAFold	130M	64 GPU	4 weeks	>208K	PDB+AlphaFold prediction
ProteinGenerator	130M	/	/	>208K	/
RFdiffusion	130M	8 GPU	3 days	>208K	monomers from PDB
FrameDiPT	17.5M	1 GPU	12 days	32K	monomers from PDB

AlphaFold$^{\alpha}$: Multimer and AlphaFold$^{\beta}$: Monomer

E Further results and discussions

E.1 Generating specific conformation

Table 5: Backbone RMSD w.r.t. number of generated samples

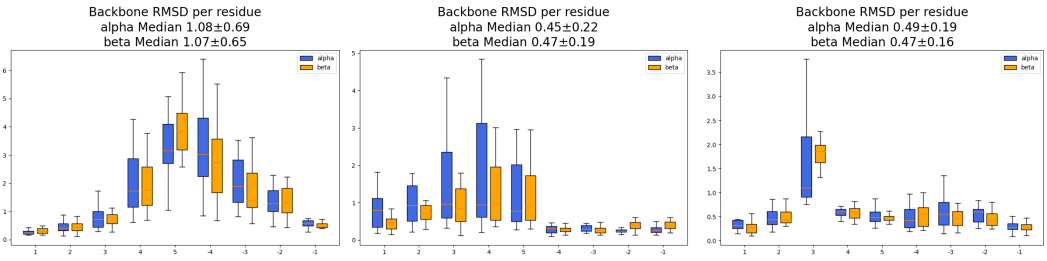
Number of samples	TCR	TCR:pMHC-I	TCR:pMHC-II
5	2.70 ± 0.43	2.18 ± 0.45	2.91 ± 0.54
25	$\mathbf{2.49 \pm 0.32}$	$\mathbf{2.05 \pm 0.49}$	$\mathbf{2.60 \pm 0.49}$

E.2 Capturing conformational distributions

Backbone RMSD per residue is also computed to evaluate how our model performs at different residue positions and we observe different patterns for different diffused regions. Figure 4 visualises an example (PDB 1KGC) of generated samples for CDR3 N-terminal and C-terminal flanks, which shows consistent structural properties. Figure 3 shows backbone RMSD per residue of CDR3, N-terminal and C-terminal flanks of CDR3. The CDR3 loop shows larger RMSDs in the middle of the loop. For N-terminal flank, the positions close to CDR3 loop usually consist of beta strands for which small RMSDs are obtained while for the positions further away from CDR3 loop, more potential conformations are predicted therefore leading to higher RMSD. For the C-terminal flank, the RMSD is consistently low which is coincident with end of the loop becoming a beta strand. Position 3 is an

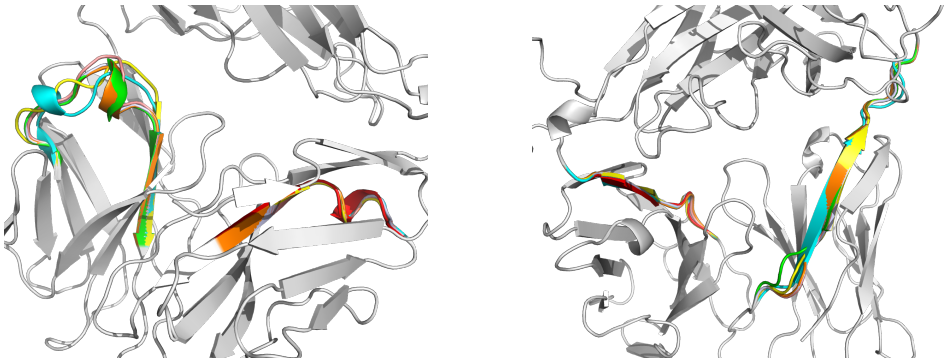
[5]TCR:pMHC class I structures missing MHC beta chains were kept, since the domain is not involved in the TCR:pMHC interface.

exception showing a local increase of RMSD, consistent with a kink following the loop at the start of the beta strand.



(a) CDR3 (b) N-terminal flank (c) C-terminal flank

Figure 3: Backbone RMSD per residue on TCR dataset of (a) CDR3, (b) N-terminal flank and (c) C-terminal flank. CDR3 loop shows greater RMSD in the middle of the loop while N-terminal flank shows smaller RMSD at positions close to CDR3 and C-terminal flank shows small RMSD in general except the third position.



(a) Diffused region on N-terminal flank (b) Diffused region on C-terminal flank

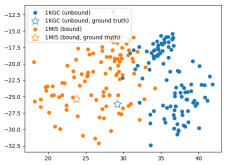
Figure 4: PDB 1KGC with context structure, ground truth alpha, ground truth beta and FrameDiPT predictions in other colors for (a) diffused region on N-terminal flank and (b) diffused region on C-terminal flank. Structural properties correspond to backbone RMSD per residue shown in Figure 3 where positions with smaller RMSD are usually beta strands and those with bigger RMSD are usually loops, especially the 3rd position of C-terminal flank corresponds to a kink in the loop structure.

Table 6: Backbone RMSD and generated sample variance of different CDR loops

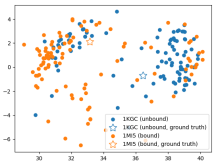
CDR loop	Metric	TCR	TCR:pMHC-I	TCR:pMHC-II
CDR1	Backbone RMSD	1.22 ± 0.23	1.28 ± 0.26	1.60 ± 0.50
	Variance	0.96 ± 0.40	1.11 ± 0.21	1.40 ± 0.35
CDR2	Backbone RMSD	1.07 ± 0.17	1.19 ± 0.33	1.51 ± 0.26
	Variance	0.75 ± 0.22	0.85 ± 0.26	0.77 ± 0.30
CDR3	Backbone RMSD	2.53 ± 0.56	2.44 ± 0.45	3.00 ± 0.57
	Variance	1.87 ± 0.22	2.06 ± 0.44	2.78 ± 0.68

All CDR loops diffusion Loops are usually flexible structures while different loops could have different structure flexibility, for example CDR3 loops are the most variable w.r.t. CDR1 and CDR2 loops in TCR chains. We also performed diffusion on all the three CDR loops and compared the backbone RMSD and sample variance in Table 6. Smaller backbone RMSD and sample variance are obtained for CDR1 and CDR2 loops.

E.3 Conformation change upon binding

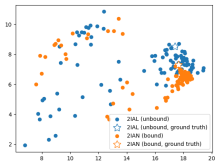


(a) CDR3α PCA

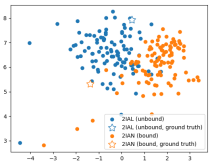


(b) CDR3β PCA

Figure 5: Conformational distributions of 1KGC (unbound) and 1MI5 (bound)

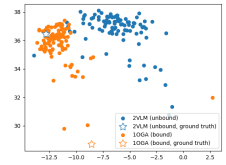


(a) CDR3α PCA

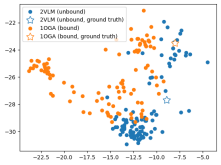


(b) CDR3β PCA

Figure 6: Conformational distributions of 2IAL (unbound) and 2IAN (bound)

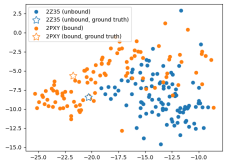


(a) CDR3α PCA

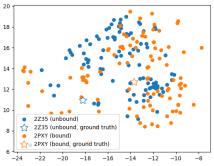


(b) CDR3β PCA

Figure 7: Conformational distributions of 2VLM (unbound) and 1OGA (bound)



(a) CDR3α PCA

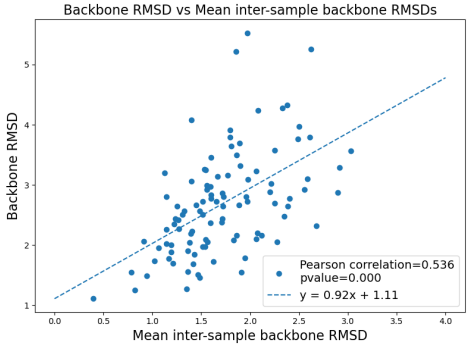


(b) CDR3β PCA

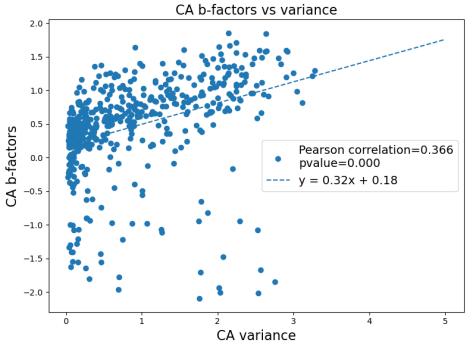
Figure 8: Conformational distributions of 2Z35 (unbound) and 2PXY (bound)

E.4 Quantifying uncertainty

We analysed the correlation between backbone RMSD and sampling variance (Figure 9a) and between normalised carbon-alpha B-factors and sampling variance (Figure 9b). Though we performed a standard normalisation of B-factors over the whole protein structure to remove intrinsic factors, no evident correlation between B-factors and sampling variance was observed.



(a)



(b)

Figure 9: Correlation between a) median backbone RMSD and the variance of generated sample which is computed as mean inter-sample backbone RMSD; b) normalised B-factors and the sampling variance on carbon-alpha of each residue. A Pearson correlation > 0.5 was observed between backbone RMSD and sampling variance while no strong correlation between normalised B-factors and variance.

E.5 Comparison to deterministic protein folding models

Table 7: Backbone and full-atom RMSD comparison of TCR CDR3 loop design. A signed Wilcoxon paired two-sided rank statistical test between FrameDiPT and the best AlphaFold model is performed at significance level p-value < 0.05. Underline means significantly different from the best AlphaFold model.

Method	RMSD	TCR	TCR:pMHC-I	TCR:pMHC-II
AlphaFold$^{\alpha}$	Backbone	2.60 ± 0.68	2.46 ± 0.68	2.33 ± 0.65
	Full-atom	3.15 ± 0.54	2.97 ± 0.76	2.91 ± 0.65
AlphaFold$^{\beta}$	Backbone	$\mathbf{1.75 \pm 0.68}$	1.58 ± 0.51	1.75 ± 1.08
	Full-atom	$\mathbf{2.20 \pm 0.72}$	$\mathbf{2.27 \pm 0.75}$	$\mathbf{2.04 \pm 1.03}$
AlphaFold$^{\gamma}$	Backbone	2.03 ± 0.68	$\mathbf{1.34 \pm 0.48}$	$\mathbf{1.37 \pm 0.43}$
	Full-atom	2.49 ± 0.66	$\mathbf{2.27 \pm 0.87}$	2.10 ± 0.65
ESMFold	Backbone	2.58 ± 0.86	2.13 ± 0.42	2.27 ± 0.85
	Full-atom	3.24 ± 0.80	2.83 ± 0.66	2.41 ± 0.70
FrameDiPT (25 samples)	Backbone	$\underline{2.49 \pm 0.32}$	$\underline{2.05 \pm 0.49}$	$\underline{2.60 \pm 0.49}$
	Full-atom	$\underline{3.45 \pm 0.47}$	$\underline{2.83 \pm 0.65}$	$\underline{3.26 \pm 0.56}$

AlphaFold$^{\alpha}$: AlphaFold Multimer with searched templates from PDB70.

AlphaFold$^{\beta}$: AlphaFold Multimer with custom templates by masking CDR3 loops.

AlphaFold$^{\gamma}$: AlphaFold Monomer with custom templates by masking CDR3 loops.