TriFold: A New Architecture for Predicting Protein Sequences from Structural Data

Harish Srinivasan
Lyda Hill Department of Bioinformatics
University of Texas Southwestern Medical Center
harish.srinivasan@utsouthwestern.edu

Jian Zhou
Lyda Hill Department of Bioinformatics
University of Texas Southwestern Medical Center
jian.zhou@utsouthwestern.edu

Abstract

The inverse protein folding challenge aims to identify specific amino acid sequences that fold into a predetermined protein structure. Despite advancements like AlphaFold2, it remains a complex issue in protein engineering. This paper introduces a novel architecture inspired by the self-attention mechanisms in AlphaFold2 and RoseTTAFold2, adapted for solving the inverse folding problem. Our approach, contrasted with previous graph-based models, leverages attention-based transformer architecture to efficiently integrate information across the entire protein. We combine attention mechanisms, such as invariant point attention, with those designed for sequence and pair representations, resulting in enhanced performance in the inverse protein folding task. Furthermore, we introduce a novel feature representation of protein structure used as an inductive bias in pair representation. The proposed model is trained and tested using the OpenFold codebase on the Protein Data Bank and the AlphaFold distillation dataset, achieving performance improvements over ProteinMPNN regarding sequence recovery. The model's validation on the CAMEO dataset, which comprises proteins released from October 16th, 2021 – January 16th, 2022, further substantiates its efficacy in enhanced sequence recovery across short, single, and multiple chains.

1 Introduction

In the field of protein engineering, addressing the inverse protein folding problem stands as a notable challenge. This problem concerns identifying a specific amino acid sequence that will fold to a predetermined protein backbone structure. Recent groundbreaking solutions for the inverse problem have emerged, with graph-based neural networks at the forefront, like ProteinMPNN and PiFold, GVP-GNN, and alphadesign [5, 6, 9, 7]. These models have shown to be accurate while also being small in model size. Nonetheless, there remains scope for enhancement in the performance of current approaches.

The advances in computational structural biology have been greatly propelled by the development of AlphaFold2[10], which made substantial progress in the protein structure prediction problem. The exploration into the crucial features that underlie the success of AlphaFold2 and RosettaFold2[3] has paved the way for numerous new applications, and we expect them to provide new insight into potential solutions for the inverse protein folding problem.

Machine Learning for Structural Biology Workshop, NeurIPS 2023.

In this manuscript, we propose a new and flexible architecture inspired by the self-attention mechanisms in AlphaFold2 and RoseTTAFold2 but adapted to solving the inverse folding problem. Attention-based architecture can efficiently integrate information from the entire protein, in contrast to graph-based models such as ProteinMPNN, which are limited to integrating information from a local neighborhood. However, despite success in structure prediction tasks, such architectures are not designed and readily applicable to the inverse-folding problem. While GVP-transformer[8] incorporated transformer in its architecture, it employed a generic transformer architecture that doesn't fully leverage structure-based features. We achieve this by integrating invariant point attention for structure representation with attention mechanisms designed for sequence and pair representations. Additionally, we introduce a novel feature representation of a protein structure called pairwise relative distance representation as a substitute for the commonly used Euclidean distance representation. Our novel approach introduces all-to-all interactions across different representation types, aiming to improve performance in the inverse protein folding problem.

2 Methods

We propose a new architecture, drawing inspiration from the attention mechanisms in both AlphaFold and RoseTTAFold2. Our implementation is based on the OpenFold codebase[1], a faithful reimplementation of AlphaFold2. This architecture outperforms ProteinMPNN in sequence recovery on our holdout dataset. At the core of our architecture lies a versatile module that takes in and outputs single, pair, and rigid representations simultaneously. Below, we describe these representations specifically for the inverse protein folding problem:

- Single Representation: We initialize this as a trainable parameter and subsequently use it to predict the amino acid sequence. Specifically, a multilayer perceptron is used to predict the sequence identity from the single representation. Its programmable nature allows our model to adapt or extend to various generative models, such as diffusion.
- Pair Representation: The pair representation is initialized as the positional embedding, based on the residue index, of the protein. The process involves computing the relative distances by determining the difference in residue indices between pairs of residues. This calculated distance is then transformed into a one-hot vector clipped within a range of [-48 to 48]. A linear projection is applied to this one hot vector to initialize the pair representation.
- Rigid Representation: Following the methodology employed by AlphaFold2, the Gram-Schmidt process is applied to the backbone atom coordinates. In this context, each residue is represented using N, C_a, C backbone atoms in the Gram-Schmidt process to construct a rigid frame transformation. The transformation, denoted as T , is expressed as $T = (R, \vec{t})$, where R is the rotation matrix and $\vec{t}$ is the translation vector which is also the C_α position of the residue.

Since each module has the same inputs and outputs, we can stack these modules, allowing subsequent module outputs to evolve into latent representations. Within each module, we subject all representations to attention-based updates and interchange between the three representations. The single representation is updated by self-attention biased by the pair representation. This pair representation plays a role in updating the single representation by biasing the dot product affinities matrix during attention. The single representation reshaped to the form of pair representation $(N_{res}, N_{res}, hidden)$ through an outer sum influences the pair representation updates. The updated single and pair representation influences the rigid representation through Invariant Point Attention(IPA), and the single representation is also updated. Unlike AlphaFold2, we initialize the rigids to the protein backbone coordinates instead of the block hole initialization for IPA. In addition, the rigids are used to update pair representation instead of keeping pair representation unchanged.

RoseTTAFold2 showed that Triangular multiplicative update and biased axial attention are required for the best performance to update the pair representation. The biased row and column-wise axial tied attention has the added benefit of updating the pair representation while allowing information to flow from the rigid representation to the pair as a bias. We utilize these attention mechanisms in our pair representation update, and here, we present a novel method to create the pair bias that is invariant to rotations and translations.

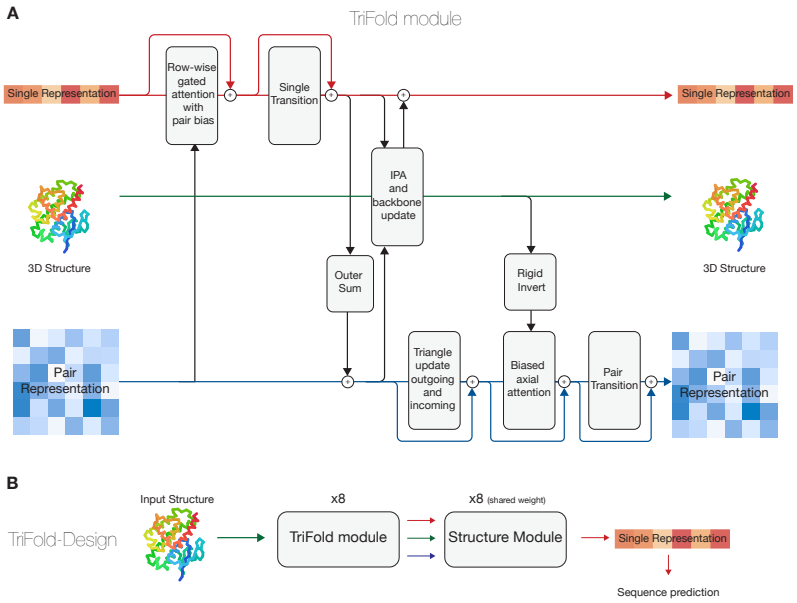

Figure 1: Model Architecture A)The architecture of one unit of a block that takes single, rigid, and pair representations as input and output. Arrows indicate the direction of flow for the representation, and after an attention-based update, the previous representation is added, which is indicated by a '+' B) Architecture of the entire model for sequence prediction

2.1 Pairwise Relative Distance Representation

To introduce information flow from the structure to pair presentation, we devised a new type of rotational and translational invariant pairwise representation of the rigids, which we refer to as Pairwise Relative Distance Representation. This representation achieved better performance compared to the commonly used squared Euclidean distance representation in our experiments. This representation introduces structure information from the updated rigid representation from IPA and backbone update to pair representation updates.

Specifically, for a protein chain with N residues, each rigid frame T_i represents a transformation from the local to a global reference. We calculate the relative distance coordinate for each pair of residues i, j where $i, j \in [0, N-1]$. We do this by taking the inverse transformation and applying it to all C_α positions x_j, and the following formulas explain this:

The inverse transformation is defined as

$$T^{-1} = (R, \vec{t})^{-1} = (R^{-1}, -R^{-1}\vec{t})$$

We apply the inverse transformation across all residues to create a $(N_{res}, N_{res}, 3)$ matrix as:

$$d_{ij} = T_i^{-1} \circ x_j$$
$$d_{ij} = R_i^{-1} x_j + -R_i^{-1}\vec{t}_i$$

d_{ij} is interpreted as a relative distance matrix invariant to rotations and translations. This representation for individual residue pairs preserves more information than Euclidean distance, as the relative orientation information is also preserved. We apply a linear transformation on this pairwise relative distance matrix to bias the pair representation through biased axial attention.

2.2 Dataset and Training

We train our model on Protein Data Bank[4] with a cutoff date of October 10th, 2021, and the alphafold distillation dataset that is generously open-sourced by the openfold team[2]. The dataset

was filtered to exclude low-resolution (greater than 9 Angstrom) protein chains and protein chains where a single amino acid dominated the entire sequence (80% of the sequence). We also utilized the stochastic filters and sampled chains inversely proportional to the cluster size that was used in openfold. The protein chains were also sampled with 75% probability from the self-distillation dataset and 25% from the PDB dataset. Each training sample consisted of one protein chain cropped to a residue size of 256. The residue crop start position is sampled from start $\sim$ Uniform$[1, 256 - x + 1]$, where $x \sim$ Uniform$[0, 256]$. If the protein chain is smaller than 256, we keep it as it is and pad the sequence until 256 residues. Our model comprises of eight trifold modules and eight shared blocks of structure module. The final model has 25 million parameters with a weight file size of approximately 99.8MB. We train with a learning rate warmup with a maximum lr = 1e-3. The batch size during training is set to 8, and we trained on four GPU A100 that took 36GB of memory per GPU for 125 epochs. The model was trained with the cross-entropy loss as the objective function and averaged across all TriFold modules and structure module blocks. The loss function is defined below:

$$L(y, \hat{y}) = -\sum_{i=1}^{C} y_i \log(\hat{y}_i)$$

where y is the true label vector, $\hat{y}$ is the predicted probability vector, C is the number of classes (C = 21 for 20 AA and a missing token), and i is the index of the class.

3 Result

In this paper, the validation of our model is conducted on the CAMEO dataset[12], similarly utilized for the validation dataset in openfold. Our CAMEO dataset comprises of proteins with a maximum sequence length of 700, released from October 16, 2021 – January 16, 2022, culminating in 183 samples. Within this dataset, 55 are single-chain proteins. Among these, 7 have fewer than 100 amino acids and are classified as short. Additionally, 121 samples possess multiple chains. In calculating statistics such as sequence recovery, residues without positional value in the file are omitted, and sequence recovery across all chains is calculated. The performance of our model is validated by computing the sequence recovery and contrasting it with ProteinMPNN, a model trained on PDB until August 02, 2021. To our understanding, ProteinMPNN is the nearest model trained on a dataset with a comparable date cutoff, thus facilitating a more straightforward comparison of our model's efficacy. The default setting in ProteinMPNN is employed, wherein all available chains are designed, the temperature parameter is set to 0.1, and two sequences are predicted. The sequence with the higher sequence recovery is selected for our validation result.

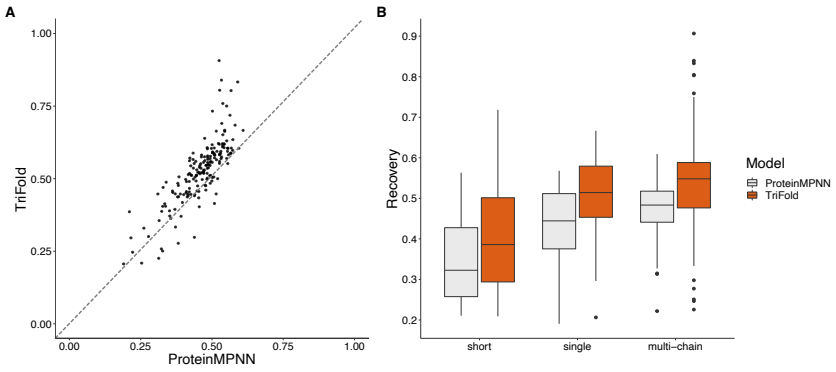

Figure 2: Sequence Recovery in ProteinMPNN and our Model using CAMEO dataset A) Sequence recovery comparison between our model and ProteinMPNN B) Comparison across small, single and multi-chain proteins

Compared to ProteinMPNN, our model demonstrates an enhanced capability to recover sequence identity for most proteins in our validation. Figure 2A illustrates that our model attains higher

sequence recovery for most samples in the CAMEO dataset. In several instances, our model achieves high accuracy (greater than 80%), a feat ProteinMPNN does not accomplish in any of the samples. Furthermore, Figure 2B underscores that our model consistently outperforms across short, single, and multiple chains, with a higher median sequence recovery.

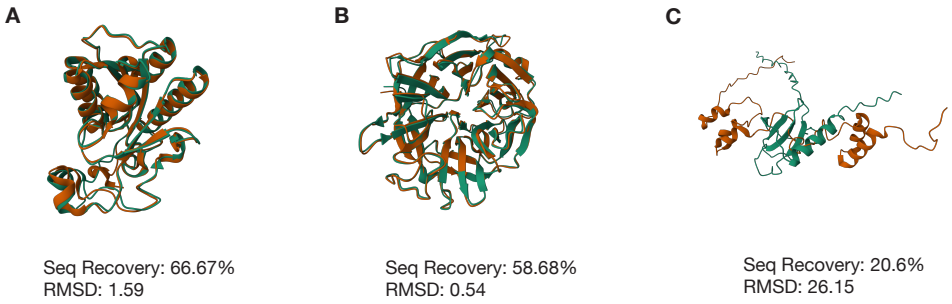

Figure 3: Example structure predictions using the designed sequence in AlphaFold2 for a) 7l6j, b)7lbu, c)7f7n. Green is the predicted structure, and burnt orange is the ground truth

In addressing the inverse folding problem, it is crucial to demonstrate that the designed sequence can fold to the desired structure. AlphaFold2[10, 13]is utilized to predict the structure of the designed sequences, as depicted in Figure 3. Figures 3A and 3B display a good alignment of the designed sequence to the ground truth. However, as Figure 3C reveals, the model encounters difficulties for proteins with disordered regions or regions lacking secondary structures, indicated by the suboptimal sequence recovery and alignment.

4 Discussion

The presented results in this paper underline the efficacy of the TriFold architecture in addressing the inverse protein folding problem. The architecture leverages attention-based capabilities to extract information from the entire protein chain. This allows TriFold to outperform existing models like ProteinMPNN regarding sequence recovery. An essential component of the architecture's success is the integration of different attention mechanisms, such as invariant point attention and those designed for single and pair representations, and allowing the exchange of representations across the module. To our knowledge, we are the first to integrate IPA with single and pair attention-based updates in a module for the inverse folding problem. Moreover, we introduced a novel approach to bias the pair representation using a pairwise relative distance representation invariant to rotations and translations. We found this to be a better feature compared to squared Euclidean distance to bias the pair representation and allow information to flow from structure to pair representation. We also hope to benchmark our model in other validation datasets, such as CATH[14], to grant insights into our model's strengths and capabilities. We also believe that our model architecture can aid in designing generative models. For example, current advancements in diffusion models like RF-diffusion[15], Frame-diff[17] for structure design, GRADE-IF[16] for sequence design, and ProteinGenerator[11] for sequence-structure design are setting the standard for contemporary generative models. We are optimistic that our architecture will contribute significantly to these emerging domains.

References

[1] Gustaf Ahdritz, Nazim Bouatta, Christina Floristean, Sachin Kadyan, Qinghui Xia, William Gerecke, Timothy J O'Donnell, Daniel Berenberg, Ian Fisk, Niccolò Zanichelli, Bo Zhang, Arkadiusz Nowaczynski, Bei Wang, Marta M Stepniewska-Dziubinska, Shang Zhang, Adegoke Ojewole, Murat Efe Guney, Stella Biderman, Andrew M Watkins, Stephen Ra, Pablo Ribalta Lorenzo, Lucas Nivon, Brian Weitzner, Yih-En Andrew Ban, Peter K Sorger, Emad Mostaque, Zhao Zhang, Richard Bonneau, and Mohammed AlQuraishi. OpenFold: Retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization. August 2023.

[2] Gustaf Ahdritz, Nazim Bouatta, Sachin Kadyan, Lukas Jarosch, Daniel Berenberg, Ian Fisk, Andrew M Watkins, Stephen Ra, Richard Bonneau, and Mohammed AlQuraishi. OpenProteinSet: Training data for structural biology at scale. *ArXiv*, August 2023.

[3] Minkyung Baek, Ivan Anishchenko, Ian R Humphreys, Qian Cong, David Baker, and Frank DiMaio. Efficient and accurate prediction of protein structure using RoseTTAFold2. May 2023.

[4] Helen Berman, Kim Henrick, Haruki Nakamura, and John L Markley. The worldwide protein data bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.*, 35 (Database issue):D301–3, January 2007.

[5] 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*, 378(6615):49–56, 2022.

[6] Zhangyang Gao, Cheng Tan, Pablo Chacón, and Stan Z Li. PiFold: Toward effective and efficient protein inverse folding. September 2022.

[7] Zhangyang Gao, Cheng Tan, and Stan Z Li. AlphaDesign: A graph protein design method and benchmark on AlphaFoldDB. February 2022.

[8] Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, and Alexander Rives. Learning inverse folding from millions of predicted structures. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 8946–8970. PMLR, 2022.

[9] Bowen Jing, Stephan Eismann, Patricia Suriana, Raphael J L Townshend, and Ron Dror. Learning from protein structure with geometric vector perceptrons. September 2020.

[10] 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.

[11] 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. May 2023.

[12] L McGuffin and Y Zhou. CAMEO - continuous automated model EvaluatiOn - welcome. `https://www.cameo3d.org/`. Accessed: 2023-10-1.

[13] Milot Mirdita, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger. ColabFold: making protein folding accessible to all. *Nat. Methods*, 19(6): 679–682, June 2022.

[14] Ian Sillitoe, Natalie Dawson, and Christine Orengo. CATH protein domain classification (version 4.2), April 2019.

[15] 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, August 2023.

[16] Kai Yi, Bingxin Zhou, Yiqing Shen, Pietro Liò, and Yu Guang Wang. Graph denoising diffusion for inverse protein folding. June 2023.

[17] 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. February 2023.