Explaining Protein Folding Networks Using Integrated Gradients and Attention Mechanisms

Rukmangadh Sai Myana[1] and Sumit Kumar Jha[1]

Department of Computer Science, Florida International University, Miami, FL, USA
rmyan001@fiu.edu,jha@cs.fiu.edu

Abstract. Protein folding prediction models like AlphaFold and ColabFold have revolutionized structural biology by providing accurate protein structures. However, these models present challenges when it comes to understanding how they arrive at their decisions. In this paper, we propose the application of Explainable AI (XAI) techniques, specifically Integrated Gradients and Attention Mechanisms, to elucidate the decision-making process of these complex networks. We conduct computational experiments to evaluate the effectiveness of these methods and discuss potential implications for the field.

Keywords: Protein Folding · Explainable AI · Integrated Gradients · Attention Mechanisms · AlphaFold · ColabFold

1 Introduction

Protein folding represents a central and enduring challenge within computational biology, as the elucidation of these intricate processes is paramount to deciphering the molecular underpinnings of biological function and disease etiology. The functionality of a protein is inextricably linked to its complex three-dimensional conformation, a structure uniquely determined by the physicochemical attributes inherent in its amino acid sequence. The precise prediction of protein structures is therefore of considerable importance, finding broad utility in diverse applications such as drug discovery [12], enzyme engineering [71], the study of genetic regulation [50], and the detailed analysis of molecular interactions [40].

The advent of deep learning methodologies has instigated a paradigm shift within the realm of protein structure prediction. Groundbreaking models, notably AlphaFold2 [46] and ColabFold [58], have demonstrated an unprecedented capacity to predict protein structures with accuracy approaching experimental levels [1]. These sophisticated innovations have effectively revolutionized the field of structural biology, empowering researchers to address and resolve intricate structural complexities previously deemed intractable [95].

Despite the remarkable predictive capabilities of these deep learning models, their operation as complex, 'black-box' systems presents significant challenges. The inherent opacity of their decision-making processes raises critical concerns regarding interpretability. Organizations encounter difficulties in discerning the rationale behind model predictions, evaluating performance consistency across diverse applications, and identifying potential systematic errors or biases embedded within the model's learning and predictive mechanisms [57, 73]. This lack of transparency not only undermines trust in model outputs but also impedes the seamless integration of these powerful tools into critical domains demanding high confidence and accountability, such as clinical diagnostics and therapeutic development, necessitating the urgent address of interpretability to fully realize their transformative potential in biological research [88].

Prior to the deep learning era, protein structure prediction relied primarily on computational modeling techniques rooted in established biophysical principles. Homology modeling, a prominent approach, constructs atomic-resolution models of target proteins by exploiting sequence similarity with homologous templates [25], with widely used tools including IntFOLD [56], RaptorX [64], and Modeller [32]. Protein threading, conversely, tackles structure prediction for proteins lacking clear homologous templates by leveraging statistical relationships between known protein folds and target sequences, exemplified by tools such as I-TASSER [72], HHpred [81], and Phyre [47].

Both traditional methodologies and contemporary machine learning models are fundamentally underpinned by the physics governing protein folding [31]. The incorporation of these foundational principles into machine

learning interpretability frameworks offers a promising avenue for achieving a more profound understanding of model predictions. Experimental observations have consistently demonstrated that proteins typically adopt well-defined three-dimensional structures, denoted as the 'native state' [48, 91]. Furthermore, the transition of many proteins from an unfolded state to their native state is characterized by an 'all-or-none' mechanism [5,67]. This characteristic transition imparts robustness to protein structures, thereby ensuring their requisite biological functionality, and understanding these internal mechanisms is critical for deciphering complex cellular functions, molecular interactions, and disease mechanisms [4, 14, 40, 90].

Explainable AI (XAI) techniques offer a potent framework for interpreting the intricacies of complex models [28]. These methodologies are designed to bridge the gap between model complexity and user comprehension by providing insights into the determinants of model predictions [60]. For instance, feature attribution methods [35, 36, 61] can identify critical sequence motifs, physicochemical properties, or interaction patterns that exert significant influence on predicted structures. Similarly, attention mechanisms [10, 55, 87] within deep learning architectures can elucidate relationships between input sequence regions within a structural context. Visualization tools, including saliency maps [80] and gradient-based overlays [76], further facilitate the correlation of model outputs with underlying biological, physical, and chemical phenomena, thereby aiding hypothesis generation, experimental validation, and broader acceptance of computational predictions within high-stakes applications such as precision medicine and drug design [88]. Integrating domain-specific knowledge, particularly principles of protein physics and thermodynamics, into XAI methodologies further enhances interpretability by grounding model predictions within established biological frameworks, fostering cross-disciplinary collaboration and making computational outputs more accessible and actionable for experimental biologists and clinicians [17, 19, 68, 74].

2 Background

Machine learning (ML) models have garnered significant attention and efficacy, broadening their impact beyond computer science into fields such as medicine, environmental science, and physics. Key innovations like the Multi-headed Attention mechanism in Transformer models [87] have transformed medical applications, enhancing diagnostics, preventive care, and medical workflow management. Specifically, Transformer-based models like ScoreNet [82], T2T-ViT [98], IL-MCAM [22], and SEViT [3] have been instrumental in histopathology image classification, while Uni4Eye [16] and LAT [85] are used for fundus image classification. Models like Chest L-Transformer [38] and FeSTA [63] demonstrate robust performance in X-ray classification.

Additionally, Transformer models have advanced *medical image segmentation* with examples like RANT [62], PCAT-UNet [20], and AMGB-Transformer [52]; *medical image detection* through models such as UTRAD [24] and STCovidNet [89]; and *medical image reconstruction* via models like SSTrans-3D [92] ReconFormer [39], and DSFormer [99]. In clinical information extraction, BioBERT [51], ClinicalBERT [42], and BlueBERT [65] have made notable contributions. Finally, applications in critical care employ models such as SimTA [94], and BENDR [49], while public health monitoring has benefited from models like BERTweet [2].

Moreover, breakthroughs in protein structure prediction include trRosetta [93], ESMFold [53], RoseTTAFold2 [9], and AlphaFold2 [46]. In pharmacology, models like TranSynergy [54], TP-DDI [97], and MolTrans [43] address drug interactions, while drug synthesis is facilitated by Molecular Transformer [75]. These advancements demonstrate the broad and transformative potential of attention based models across diverse sectors.

2.1 AlphaFold, AlphaFold2 and ColabFold

AlphaFold, a deep learning model developed by DeepMind, accurately predicts protein structures. Introduced in 2018 [77], AlphaFold generates a mean force potential for a given protein, allowing the 3D structure to be obtained via gradient descent in this potential space. This prediction is enabled by a central convolutional residual network that computes a distogram — a matrix representing distances between C_β atoms of residue pairs in a protein sequence — based on multiple sequence alignment (MSA) features. For each residue pair (i, j), AlphaFold predicts a discrete probability distribution over distances from 2 Å to 22 Å, split into 64 bins. Cubic spline interpolation is then used to construct the protein's distance potential from these discrete distributions. The torsion angles of side chains are modeled as Von Mises distributions, whose parameters

are also predicted by AlphaFold. The final mean force potential is computed by summing the log-likelihoods of the distance distributions across all residue pairs (i, j), along with the log-likelihood of the torsion angle distributions across all residues.

In 2020, DeepMind significantly advanced their previous work with the introduction of AlphaFold2 [46], a model that set a new standard in protein structure prediction by achieving unprecedented performance in the CASP14 competition. This breakthrough was largely attributed to AlphaFold2's innovative Transformer-based architecture and the integration of novel attention mechanisms, including Axial Attention [41] and Invariant Point Attention [46]. Axial Attention generalizes the Self-Attention operation to multiple dimensions while maintaining computational efficiency, while Invariant Point Attention offers self-attention that is invariant to global euclidean transformations. Together, these mechanisms allowed AlphaFold2 to accurately model complex spatial relationships within protein structures, setting it apart from competing approaches.

To construct diverse multiple sequence alignment (MSA) features, extensive protein sequence datasets from both public and environmental databases [26] [59] are queried by AlphaFold2 using advanced homology detection tools, specifically HMMer [29] and HHblits [83], which leverage hidden Markov models (HMMs). Given the considerable volume of these databases, search processes for a single protein can require several hours and demand multiple terabytes of storage. Additionally, deep neural network computations necessitate graphics processing units (GPUs) with substantial GPU RAM, particularly for proteins with larger sequence lengths. ColabFold [58] was thus introduced to accelerate single-protein predictions while maintaining a performance comparable to AlphaFold2. ColabFold achieves this by substituting AlphaFold2's homology search with the significantly faster MMseqs2 (Many-against-Many sequence searching) [84], achieving a 40 to 60-fold speed increase. For batch predictions, ColabFold reduces runtime by approximately 90-fold by eliminating recompilation and incorporating an early stopping criterion.

2.2 Explainable AI Techniques

Current interpretability methods can be broadly categorized into five approaches [7]: feature attribution, inherently interpretable models, hierarchical explanations, contrastive explanations, and counterfactual explanations.

Feature attribution methods aim to assign scores or ranks to parts of the input, quantifying their influence on a model's prediction. Techniques such as LIME [69] approximate neural network decision-making by fitting an interpretable model, like a decision tree, to the local neighborhood of the input. Anchors [70] provides explanations for individual predictions of any black-box classification model by identifying IF-THEN decision rules that "anchor" the prediction. A rule is considered an anchor if rule-abiding changes to feature values do not affect the prediction, ensuring robustness of the explanation. Other feature attribution methods like relevance propagation [8] [78], Integrated Gradients [86], and Grad-CAM [76] generate feature importance maps by back-propagating gradients through the network.

Inherently interpretable models produce a 'rationale' [11] [18] [96] [6], a concise, human-understandable explanation of their predictions. These methods typically consist of two components: a generator that extracts the rationale and an encoder that uses it to make a prediction. Such models are often termed 'self-explainable' due to their intrinsic transparency. Simpler models like linear regression [33], decision trees [15], and RuleFit [34] also fall under this category, as they naturally lend themselves to straightforward interpretation.

Hierarchical explanations aim to elucidate model outputs at multiple levels. For example, [21] employs Shapley values to provide both feature attributions and feature interactions, while [79] extends Integrated Gradients to also capture these interactions. This multi-layered approach offers a more comprehensive understanding of the prediction process. On the other hand, contrastive explanations focus on building inherently interpretable models capable of addressing questions of the form *Why 'p'*? and *Why 'p' instead of 'q'?*. While most existing post-hoc techniques, such as gradient-based methods, are limited to addressing *Why 'p'?*, contrastive methods fill this gap by explicitly comparing alternatives. Finally, Counterfactual explanations identify minimal changes to the input data that would effectively "flip" the model's prediction.

3 Methodology

3.1 Integrated Gradients

Integrated Gradients (IG) is a popular technique used in XAI to attribute the contribution of each input feature to a model's prediction. Integrated Gradients quantify how much each feature contributes to the model's output relative to a baseline. The baseline is typically chosen to represent the "absence" of information (e.g. a zero matrix of numerical inputs, a completely black image). IG then computes the cumulative gradients of the model's output with respect to the inputs, moving along a straight line from the baseline to the actual input. The integrated gradient for the i-th feature is calculated as:

$$IG_i(x) = (x_i - x_i') \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \cdot (x - x'))}{\partial x_i} d\alpha$$

where x is the actual input, x' is the baseline input, α is a scaling factor that interpolates between the baseline and the actual input, and F is the model's output function. Integrated Gradients (IG) offers several desirable properties for feature attribution. It ensures that features with no impact on the output receive zero attribution, maintains consistency across mathematically equivalent model implementations (model-agnostic), and satisfies the completeness property, where the sum of all feature attributions equals the difference between the model's prediction for the actual input and the baseline. In the context of protein structure prediction with ColabFold, the output function corresponds to the calculation of the model confidence metric while the baseline corresponds to absence of any input sequence (i.e. zero vectors) to the model.

Integrated Gradients (IG) has been effectively employed as a versatile interpretability technique across a range of domains, underscoring its utility in elucidating the decision-making processes of deep learning models. Within drug discovery, IG has facilitated the analysis of retrosynthetic reaction predictions [44] and the prediction of crucial drug-CYP3A4 enzyme interactions for understanding drug metabolism [45]. Furthermore, IG has been instrumental in enhancing the transparency of AI-driven diagnostics by interpreting disease detection models [27, 66] and in environmental science by identifying key factors in precipitation predictions [30]. These diverse applications collectively illustrate the robustness of IG as an interpretability tool, offering valuable insights for a spectrum of scientific and practical problems.

3.2 Attention Mechanisms

Attention mechanisms, introduced in the domain of machine translation, demonstrated significant improvements in sequence-to-sequence models by dynamically focusing on relevant parts of the input during prediction [10]. With the advent and subsequent success of Transformer models, attention mechanisms have gained widespread adoption across a diverse array of tasks spanning domains such as speech recognition, natural language processing, object detection, and time series prediction illustrating their versatility and efficacy.

A notable milestone in the evolution of attention-based models is AlphaFold2, which leveraged attention mechanisms for predicting protein structures with remarkable accuracy. Specifically, AlphaFold2 employs Axial Attention, a generalized and computationally efficient variant of self-attention, designed to handle high-dimensional inputs such as protein sequences and their spatial features. The formulation of Self-Attention (SA), a foundational mechanism in modern deep learning models, is expressed mathematically as follows:

$$\mathrm{SA}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V$$

In self-attention mechanisms, query(Q), key(K), andvalue (V) matrices, derived from input data, are utilized, with d_k representing the key vector dimensionality. Self-attention computes weighted sums of value vectors, where weights are determined by compatibility scores between query and key vectors, thereby measuring the attention one residue pays to another. Self-attention over these triplets enables the integration of contextual information across the sequence, capturing long-range dependencies crucial for accurate structure prediction.

Axial Attention (AA) [41] extends the standard self-attention mechanism by restricting the computation of attention to specific axes of a high-dimensional input tensor. This approach enhances computational efficiency by avoiding the full-dimensional attention computation, which can be prohibitively expensive for large-scale inputs. When Axial Attention layers are stacked such that each layer corresponds to a different axis of the input, the combined operation is equivalent to applying standard self-attention across all axes.

4 Computational Experiments

To interpret the inter-residue relationships learned by the model, we applied Integrated Gradients (IG) to the attention, or compatibility, matrices. This approach was undertaken to extract the specific inter-residue interactions that are most salient to the model's predictions. The aim was to ascertain whether these identified interactions align with established physicochemical principles known to govern protein folding and stability, thereby providing a biologically meaningful interpretation of the model's internal representations.

4.1 Dataset Selection

Protein sequences have been obtained from protein Data Bank (PDB) [13] and Uniprot [26]. The choice of protein sequence has been determined by the availability of experimentally determined 3D structure for the protein. In this context, the B1 domain of streptococcal protein G, and Sso10a from Sulfolobus solfataricus have been explored.

B1 Domain - Steptococcal Protein G [37] Protein G helps *Streptococcus* evade host defenses by binding to key host proteins. A repeating 55-residue domain in Protein G binds to the F_c region of immunoglobulin G (IgG) and α2-macroglobulin, a major protease inhibitor in human plasma. Protein G from strain GX7809 contains two such repeats, while strain GX7805 has three, with over 90% sequence identity between repeats. This domain of Protein G named as B1 domain is a significant analytical tool in immunology due to its extreme physicochemical properties and has been extensively studied. Its experimentally determined structure thus provides a benchmark for residue importance predictions by AlphaFold2.

The B1 domain (Figure 1c) exhibits a novel topology with a four-stranded β-sheet containing a central parallel pair and a +3x crossover connecting outer strands via an α-helix, a configuration unprecedented in structural databases. Its exceptional thermal stability arises from several features: (1) 95% of its residues participate in secondary structures, contributing 45 hydrogen bonds (41 backbone-backbone, 3 side chain-backbone, and 1 side chain-side chain), which stabilize turns and helices; (2) a hydrophobic core formed by tightly packed aromatic residues, such as Trp43 and Tyr45, and hydrophobic interactions involving residues like Leu5, Leu7, and Val39; and (3) a solvent-exposed surface enriched in polar and charged residues, including Thr, Lys, Glu, and Asp. These structural features together enhance stability and functionality, making the B1 domain a remarkable model for protein studies.

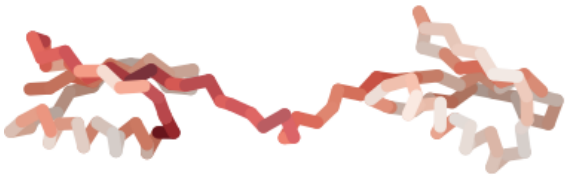
Sso10a - Sulfolobus Solfatarius [23] The crystal structure of Sso10a reveals an elongated dimer formed through crystallographic 2-fold rotation, with dimensions of 27 Å × 80 Å × 27 Å. Each monomer comprises four α-helices and three β-strands arranged in an H1–B1–H2–H3–B2–B3–H4 topology, with α-helices making up 76% of the structure and β-sheets contributing 13%. The protein is organized into two distinct domains: an N-terminal Winged Helix Domain (residues 1-59) that contains a DNA-binding helix-turn-helix motif and a wing structure formed by β-strands, and a C-terminal H4 Helix Domain that forms an essential antiparallel coiled-coil for dimerization.

The C-terminal domain's H4 helix (residues 60-92) forms a nine-turn structure that pairs with its counterpart to create a coiled-coil interface spanning 3.7 heptad repeats. This interface is stabilized by hydrophobic residues at the heptad's a and d positions, with an unusual feature being Asp69 occupying a typically hydrophobic d position to form a solvent-exposed ion pair with Lys86. The dimerization interface is characterized by a left-handed coiled-coil with a crossing angle of approximately 25°, stabilized by hydrophobic interactions, specific ion pairs (including Glu65-Lys86), and hydrogen bonds.

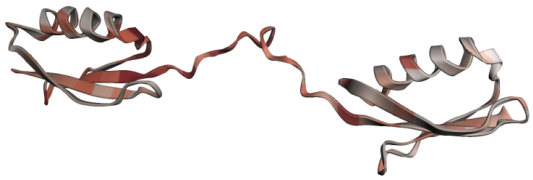
5 Results

5.1 B1 Domain - Steptococcal Protein G

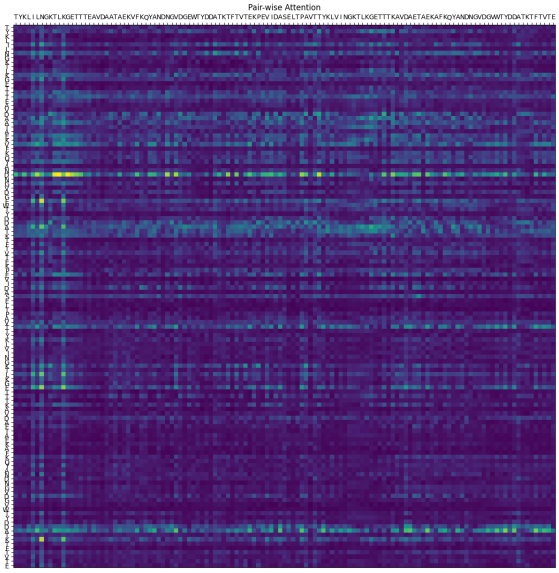
(a) Relative importances of the residues



(b) Relative importances in pseudo 3D



(c) Relative importances in 3D



(d) Attention Map

Fig. 1: *B1 Domain* of protein G

Despite the remarkable thermal stability of protein G, arising from its intricate structural features, including 45 hydrogen bonds (41 backbone-backbone, 3 side chain-backbone, and 1 side chain-side chain) that stabilize essential structural motifs such as turns and helices; robust hydrophobic interactions involving residues like Leu5, Leu7, and Val39; and a solvent-exposed surface enriched with polar and charged residues, our analysis suggests that current interpretation techniques, such as Integrated Gradients and Attention, do not fully capture the underlying physicochemical principles in a meaningful way.

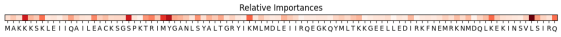
From a physiochemical perspective, the pairwise attention that each residue allocates to other residues should logically align with physiochemical interactions, such as disulfide (S-S) bonds, hydrogen bonds and other molecular interactions, which play critical roles in stabilizing protein structure. However, such relationships are conspicuously absent in the attention maps shown in Fig-1d. For example in an α helix structure, there exists a hydrogen bond between the i^{th} and $(i-4)^{th}$ peptides but such a pattern of attention is not visible in the figure. The misalignment between the attention matrix and fundamental biophysical interactions indicates that this matrix fails to capture the core physical principles that determine how protein residues interact with one another.

Similarly, while integrated gradients highlight certain residues in the protein structure, as depicted in Fig-1a, Fig-1b, and Fig-1c, these highlighted residues do not correspond meaningfully to their importance from the perspective of physics, chemistry, or biology. For instance, residues critical for forming stabilizing interactions or participating in catalytic activity remain underemphasized or misrepresented.

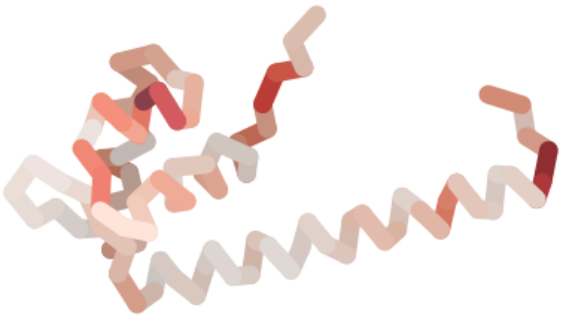
5.2 Sso10a - Sulfolobus Solfatarius

Correspondingly, while the Sso10a protein exhibits a well-defined structure with four α-helices and three β-strands (comprising 76% and 13% of the structure respectively), current attempts to interpret AlphaFold2's predictions using methods like Integrated Gradients and Attention have not aligned with our understanding of these structural elements.

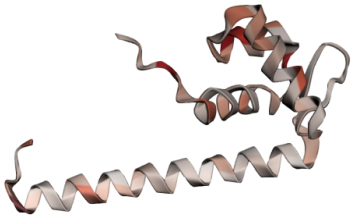
From a biophysical perspective, one would expect the attention matrix to reflect fundamental physicochemical interactions that stabilize protein structure, such as disulfide bonds and hydrogen bonding networks. However, the attention maps presented in Fig-2d do not appear to clearly reflect these essential molecular relationships. The attention patterns appear disconnected from the underlying physical principles that govern residue-residue interactions in protein structures.



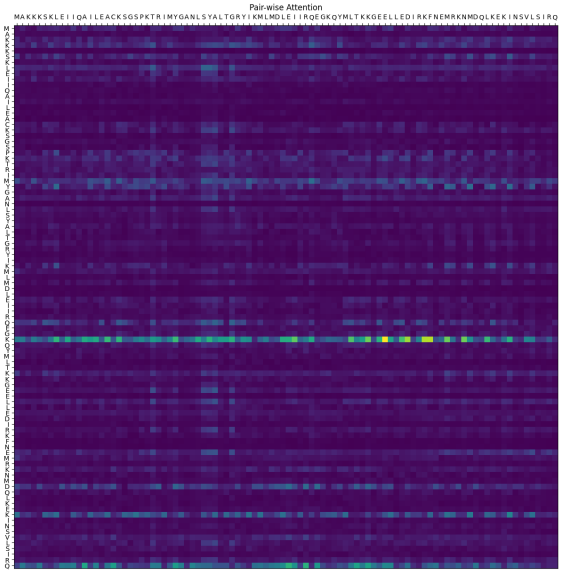
(a) Relative importances of the residues



(b) Relative importances in pseudo 3D



(c) Relative importances in 3D



(d) Attention Map

Fig. 2: *Sso10a* protein

The Integrated Gradients analysis, visualized across multiple representations in Fig-2a, Fig-2b, and Fig-2c, similarly falls short of providing biologically meaningful insights. The residues highlighted by this method do not correlate well with their known functional or structural importance. Key residues involved in structural stability or enzymatic function are often overlooked or incorrectly weighted.

6 Discussion

6.1 Implications for Structural Biology

Our analysis reveals significant challenges in applying current explainable AI methodologies to protein structure prediction networks, particularly AlphaFold2. While techniques such as Integrated Gradients and Attention mechanisms have shown promise in other domains of deep learning, their application to protein folding networks presents challenges in deriving biochemically meaningful interpretations. The disconnect between these computational outputs and established physicochemical principles underscores a fundamental limitation in our ability to decode the decision-making processes within these networks.

This interpretability gap has profound implications for structural biology. As protein structure prediction models become increasingly integrated into biological research workflows, the inability to validate their predictions against known biophysical principles poses potential risks. While these models achieve high accuracy, their black-box nature poses challenges in leveraging insights for advancing our understanding of protein folding mechanisms.

Following the paradigm established by Rudin [73] in other machine learning domains, we propose that the field of protein structure prediction would benefit from a fundamental shift toward inherently interpretable architectures. Rather than attempting to retrofit explanation methods onto existing models, future development should prioritize architectures that incorporate explicit representations of physical and chemical principles. Such models would ideally maintain or exceed the accuracy of current approaches while providing transparent reasoning paths that align with established biological knowledge.

6.2 Limitations and Future Directions

While our study highlights the inadequacy of current interpretability methods for AlphaFold2, several important caveats merit consideration. First, the limitations we observe may be specific to the particular combination of model architecture and explanation techniques employed.

The complexity and scale of protein folding networks, may demand novel interpretability approaches that are specifically designed for this domain. We propose that future work should focus on developing custom interpretability frameworks that explicitly incorporate biochemical and physical principles. Such frameworks might include: Physics-aware attention mechanisms that track and visualize specific types of molecular interactions, Gradient-based methods that are calibrated to known structure-stability relationships, Interpretability layers that map network activations to recognized biological motifs and interaction patterns, Hybrid approaches that combine machine learning interpretability with molecular dynamics insights. These developments would not only enhance our confidence in model predictions but also potentially provide new insights into the physical principles governing protein folding and structure.

7 Conclusion

Our investigation into explainable AI methods for protein folding networks, focusing on Integrated Gradients and Attention Mechanisms, has revealed both opportunities and significant challenges. While these methods offer initial insights into model behavior, their current implementations fall short of providing interpretations that align meaningfully with established biochemical principles and protein folding mechanisms. We plan to conduct extensive computational experiments to evaluate the robustness of proposed interpretation methods across diverse protein structures. Future work will also focus on incorporating physical principles into interpretability methods by developing attribution methods that reflect molecular interactions and creating attention mechanisms aligned with physical force fields.

While our current work has established a foundation for explaining protein folding networks, there are still notable challenges in developing interpretability methods that align closely with biochemical principles. The path forward requires close collaboration between computational scientists and structural biologists to ensure that future methods not only provide mathematical explanations but also offer genuine insights into the biological principles governing protein structure. Success in this endeavor would not only advance

our understanding of protein folding networks but also contribute to the broader goal of creating more interpretable and trustworthy AI systems for scientific applications.

References

1. Ahdritz, G., Bouatta, N., Floristean, C., Kadyan, S., Xia, Q., Gerecke, W., O'Donnell, T.J., Berenberg, D., Fisk, I., Zanichelli, N., et al.: Openfold: Retraining alphafold2 yields new insights into its learning mechanisms and capacity for generalization. Nature Methods pp. 1–11 (2024)
2. Ahne, A., Khetan, V., Tannier, X., Rizvi, M.I.H., Czernichow, T., Orchard, F., Bour, C., Fano, A., Fagherazzi, G., et al.: Extraction of explicit and implicit cause-effect relationships in patient-reported diabetes-related tweets from 2017 to 2021: deep learning approach. JMIR medical informatics **10**(7), e37201 (2022)
3. Almalik, F., Yaqub, M., Nandakumar, K.: Self-ensembling vision transformer (sevit) for robust medical image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 376–386. Springer (2022)
4. Ananthakrishnan, R., Ehrlicher, A.: The forces behind cell movement. International journal of biological sciences **3**(5), 303 (2007)
5. Anfinsen, C.B., Haber, E., Sela, M., White Jr, F.: The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. Proceedings of the National Academy of Sciences **47**(9), 1309–1314 (1961)
6. Antognini, D., Faltings, B.: Rationalization through concepts. arXiv preprint arXiv:2105.04837 (2021)
7. Babiker, H.K.B.: Deep interpretable modelling. Education and Research Archive (2023)
8. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one **10**(7), e0130140 (2015)
9. Baek, M., Anishchenko, I., Humphreys, I.R., Cong, Q., Baker, D., DiMaio, F.: Efficient and accurate prediction of protein structure using rosettafold2. BioRxiv pp. 2023–05 (2023)
10. Bahdanau, D.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
11. Bastings, J., Aziz, W., Titov, I.: Interpretable neural predictions with differentiable binary variables. arXiv preprint arXiv:1905.08160 (2019)
12. Batool, M., Ahmad, B., Choi, S.: A structure-based drug discovery paradigm. International journal of molecular sciences **20**(11), 2783 (2019)
13. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E.: The protein data bank. Nucleic acids research **28**(1), 235–242 (2000)
14. Blair, D.F., Dutcher, S.K.: Flagella in prokaryotes and lower eukaryotes. Current Opinion in Genetics & Development **2**(5), 756–767 (1992)
15. Breiman, L.: Classification and regression trees. Routledge (2017)
16. Cai, Z., Lin, L., He, H., Tang, X.: Uni4eye: unified 2d and 3d self-supervised pre-training via masked image modeling transformer for ophthalmic image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 88–98. Springer (2022)
17. Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., Elhadad, N.: Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1721–1730 (2015)
18. Chang, S., Zhang, Y., Yu, M., Jaakkola, T.: Invariant rationalization. In: International Conference on Machine Learning. pp. 1448–1458. PMLR (2020)
19. Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., Su, J.K.: This looks like that: deep learning for interpretable image recognition. Advances in neural information processing systems **32** (2019)
20. Chen, D., Yang, W., Wang, L., Tan, S., Lin, J., Bu, W.: Pcat-unet: Unet-like network fused convolution and transformer for retinal vessel segmentation. PloS one **17**(1), e0262689 (2022)
21. Chen, H., Zheng, G., Ji, Y.: Generating hierarchical explanations on text classification via feature interaction detection. arXiv preprint arXiv:2004.02015 (2020)
22. Chen, H., Li, C., Li, X., Rahaman, M.M., Hu, W., Li, Y., Liu, W., Sun, C., Sun, H., Huang, X., et al.: Il-mcam: An interactive learning and multi-channel attention mechanism-based weakly supervised colorectal histopathology image classification approach. Computers in Biology and Medicine **143**, 105265 (2022)
23. Chen, L., Chen, L.R., Zhou, X.E., Wang, Y., Kahsai, M.A., Clark, A.T., Edmondson, S.P., Liu, Z.J., Rose, J.P., Wang, B.C., et al.: The hyperthermophile protein sso10a is a dimer of winged helix dna-binding domains linked by an antiparallel coiled coil rod. Journal of molecular biology **341**(1), 73–91 (2004)
24. Chen, L., You, Z., Zhang, N., Xi, J., Le, X.: Utrad: Anomaly detection and localization with u-transformer. Neural Networks **147**, 53–62 (2022)
25. Chothia, C., Lesk, A.M.: The relation between the divergence of sequence and structure in proteins. The EMBO journal **5**(4), 823–826 (1986)

26. Consortium, U.: Uniprot: a worldwide hub of protein knowledge. Nucleic acids research **47**(D1), D506–D515 (2019)
27. DeGrave, A.J., Janizek, J.D., Lee, S.I.: Ai for radiographic covid-19 detection selects shortcuts over signal. Nature Machine Intelligence **3**(7), 610–619 (2021)
28. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 (2017)
29. Eddy, S.R.: Accelerated profile hmm searches. PLoS computational biology **7**(10), e1002195 (2011)
30. Espeholt, L., Agrawal, S., Sønderby, C., Kumar, M., Heek, J., Bromberg, C., Gazen, C., Carver, R., Andrychowicz, M., Hickey, J., et al.: Deep learning for twelve hour precipitation forecasts. Nature communications **13**(1), 1–10 (2022)
31. Finkelstein, A., Galzitskaya, O.: Physics of protein folding. Physics of Life reviews **1**(1), 23–56 (2004)
32. Fiser, A., Šali, A.: Modeller: generation and refinement of homology-based protein structure models. In: Methods in enzymology, vol. 374, pp. 461–491. Elsevier (2003)
33. Freedman, D.: Statistical Models : Theory and Practice. Cambridge University Press (2005)
34. Friedman, J.H., Popescu, B.E.: Predictive learning via rule ensembles. The Annals of Applied Statistics (2008)
35. Garson, G.D.: Interpreting neural-network connection weights. AI Expert **6**(4), 46–51 (Apr 1991)
36. Greenwell, B.M., Boehmke, B.C., McCarthy, A.J.: A simple and effective model-based variable importance measure. arXiv preprint arXiv:1805.04755 (2018)
37. Gronenborn, A.M., Filpula, D.R., Essig, N.Z., Achari, A., Whitlow, M., Wingfield, P.T., Clore, G.M.: A novel, highly stable fold of the immunoglobulin binding domain of streptococcal protein g. Science **253**(5020), 657–661 (1991)
38. Gu, H., Wang, H., Qin, P., Wang, J.: Chest l-transformer: local features with position attention for weakly supervised chest radiograph segmentation and classification. Frontiers in Medicine **9**, 923456 (2022)
39. Guo, P., Mei, Y., Zhou, J., Jiang, S., Patel, V.M.: Reconformer: Accelerated mri reconstruction using recurrent transformer. IEEE transactions on medical imaging (2023)
40. Hegyi, H., Gerstein, M.: The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. Journal of molecular biology **288**(1), 147–164 (1999)
41. Ho, J., Kalchbrenner, N., Weissenborn, D., Salimans, T.: Axial attention in multidimensional transformers. arXiv preprint arXiv:1912.12180 (2019)
42. Huang, K., Altosaar, J., Ranganath, R.: Clinicalbert: Modeling clinical notes and predicting hospital readmission. arXiv preprint arXiv:1904.05342 (2019)
43. Huang, K., Xiao, C., Glass, L.M., Sun, J.: Moltrans: molecular interaction transformer for drug–target interaction prediction. Bioinformatics **37**(6), 830–836 (2021)
44. Ishida, S., Terayama, K., Kojima, R., Takasu, K., Okuno, Y.: Prediction and interpretable visualization of retrosynthetic reactions using graph convolutional networks. Journal of chemical information and modeling **59**(12), 5026–5033 (2019)
45. Jiménez-Luna, J., Grisoni, F., Schneider, G.: Drug discovery with explainable artificial intelligence. Nature Machine Intelligence **2**(10), 573–584 (2020)
46. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al.: Highly accurate protein structure prediction with alphafold. nature **596**(7873), 583–589 (2021)
47. Kelley, L.A., Sternberg, M.J.: Protein structure prediction on the web: a case study using the phyre server. Nature protocols **4**(3), 363–371 (2009)
48. Kendrew, J.C., Bodo, G., Dintzis, H.M., Parrish, R., Wyckoff, H., Phillips, D.C.: A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. Nature **181**(4610), 662–666 (1958)
49. Kostas, D., Aroca-Ouellette, S., Rudzicz, F.: Bendr: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of eeg data. Frontiers in Human Neuroscience **15**, 653659 (2021)
50. Latchman, D.S.: Transcription factors: an overview. International journal of experimental pathology **74**(5), 417 (1993)
51. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**(4), 1234–1240 (2020)
52. Li, Y., Zeng, G., Zhang, Y., Wang, J., Jin, Q., Sun, L., Zhang, Q., Lian, Q., Qian, G., Xia, N., et al.: Agmb-transformer: Anatomy-guided multi-branch transformer network for automated evaluation of root canal therapy. IEEE Journal of Biomedical and Health Informatics **26**(4), 1684–1695 (2021)
53. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al.: Evolutionary-scale prediction of atomic-level protein structure with a language model. Science **379**(6637), 1123–1130 (2023)
54. Liu, Q., Xie, L.: Transynergy: Mechanism-driven interpretable deep neural network for the synergistic prediction and pathway deconvolution of drug combinations. PLoS computational biology **17**(2), e1008653 (2021)

55. Luong, M.T.: Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025 (2015)
56. McGuffin, L.J., Adiyaman, R., Maghrabi, A.H., Shuid, A.N., Brackenridge, D.A., Nealon, J.O., Philomina, L.S.: Intfold: an integrated web resource for high performance protein structure and function prediction. Nucleic acids research **47**(W1), W408–W413 (2019)
57. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial intelligence **267**, 1–38 (2019)
58. Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., Steinegger, M.: Colabfold: making protein folding accessible to all. Nature methods **19**(6), 679–682 (2022)
59. Mitchell, A.L., Almeida, A., Beracochea, M., Boland, M., Burgin, J., Cochrane, G., Crusoe, M.R., Kale, V., Potter, S.C., Richardson, L.J., et al.: Mgnify: the microbiome analysis resource in 2020. Nucleic acids research **48**(D1), D570–D578 (2020)
60. Murdoch, W.J., Singh, C., Kumbier, K., Abbasi-Asl, R., Yu, B.: Definitions, methods, and applications in interpretable machine learning. Proceedings of the National Academy of Sciences **116**(44), 22071–22080 (2019)
61. Olden, J.D., Joy, M.K., Death, R.G.: An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. Ecological Modelling **178**(3), 389–397 (2004). https://doi.org/https://doi.org/10.1016/j.ecolmodel.2004.03.013, `https://www.sciencedirect.com/science/article/pii/S0304380004001565`
62. Pan, X., Bai, W., Ma, M., Zhang, S.: Rant: A cascade reverse attention segmentation framework with hybrid transformer for laryngeal endoscope images. Biomedical Signal Processing and Control **78**, 103890 (2022)
63. Park, S., Kim, G., Kim, J., Kim, B., Ye, J.C.: Federated split vision transformer for covid-19 cxr diagnosis using task-agnostic training. arXiv preprint arXiv:2111.01338 (2021)
64. Peng, J., Xu, J.: Raptorx: exploiting structure information for protein alignment by statistical inference. Proteins: Structure, Function, and Bioinformatics **79**(S10), 161–171 (2011)
65. Peng, Y., Yan, S., Lu, Z.: Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In: Demner-Fushman, D., Cohen, K.B., Ananiadou, S., Tsujii, J. (eds.) Proceedings of the 18th BioNLP Workshop and Shared Task. pp. 58–65. Association for Computational Linguistics, Florence, Italy (Aug 2019). https://doi.org/10.18653/v1/W19-5006, `https://aclanthology.org/W19-5006`
66. Placido, D., Yuan, B., Hjaltelin, J.X., Zheng, C., Haue, A.D., Chmura, P.J., Yuan, C., Kim, J., Umeton, R., Antell, G., et al.: A deep learning algorithm to predict risk of pancreatic cancer from disease trajectories. Nature medicine **29**(5), 1113–1122 (2023)
67. Privalov, P., Khechinashvili, N.: A thermodynamic approach to the problem of stabilization of globular protein structure: a calorimetric study. Journal of molecular biology **86**(3), 665–684 (1974)
68. Razavian, N., Blecker, S., Schmidt, A.M., Smith-McLallen, A., Nigam, S., Sontag, D.: Population-level prediction of type 2 diabetes from claims data and analysis of risk factors. Big Data **3**(4), 277–287 (2015)
69. Ribeiro, M.T., Singh, S., Guestrin, C.: "why should I trust you?": Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016. pp. 1135–1144 (2016)
70. Ribeiro, M.T., Singh, S., Guestrin, C.: Anchors: High-precision model-agnostic explanations. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
71. Ringe, D., Petsko, G.A.: How enzymes work. Science **320**(5882), 1428–1429 (2008)
72. Roy, A., Kucukural, A., Zhang, Y.: I-tasser: a unified platform for automated protein structure and function prediction. Nature protocols **5**(4), 725–738 (2010)
73. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature machine intelligence **1**(5), 206–215 (2019)
74. Rudin, C., Ustun, B.: Optimized scoring systems: Toward trust in machine learning for healthcare and criminal justice. Interfaces **48**(5), 449–466 (2018)
75. Schwaller, P., Petraglia, R., Zullo, V., Nair, V.H., Haeuselmann, R.A., Pisoni, R., Bekas, C., Iuliano, A., Laino, T.: Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. Chemical science **11**(12), 3316–3325 (2020)
76. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: visual explanations from deep networks via gradient-based localization. International journal of computer vision **128**, 336–359 (2020)
77. Senior, A.W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A.W., Bridgland, A., et al.: Improved protein structure prediction using potentials from deep learning. Nature **577**(7792), 706–710 (2020)
78. Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. In: International conference on machine learning. pp. 3145–3153. PMlR (2017)

79. Sikdar, S., Bhattacharya, P., Heese, K.: Integrated directional gradients: Feature interaction attribution for neural nlp models. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 865–878 (2021)
80. Simonyan, K.: Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013)
81. Söding, J., Biegert, A., Lupas, A.N.: The hhpred interactive server for protein homology detection and structure prediction. Nucleic acids research **33**(suppl_2), W244–W248 (2005)
82. Stegmüller, T., Bozorgtabar, B., Spahr, A., Thiran, J.P.: Scorenet: Learning non-uniform attention and augmentation for transformer-based histopathological image classification. In: Proceedings of the IEEE/CVF winter Conference on applications of computer vision. pp. 6170–6179 (2023)
83. Steinegger, M., Meier, M., Mirdita, M., Vöhringer, H., Haunsberger, S.J., Söding, J.: Hh-suite3 for fast remote homology detection and deep protein annotation. BMC bioinformatics **20**, 1–15 (2019)
84. Steinegger, M., Söding, J.: Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nature biotechnology **35**(11), 1026–1028 (2017)
85. Sun, R., Li, Y., Zhang, T., Mao, Z., Wu, F., Zhang, Y.: Lesion-aware transformers for diabetic retinopathy grading. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10938–10947 (2021)
86. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: International conference on machine learning. pp. 3319–3328. PMLR (2017)
87. Vaswani, A.: Attention is all you need. Advances in Neural Information Processing Systems (2017)
88. van der Velden, B.H.: Explainable ai: current status and future potential. European Radiology **34**(2), 1187–1189 (2024)
89. Wang, B., Zhang, D., Tian, Z.: Stcovidnet: Automatic detection model of novel coronavirus pneumonia based on swin transformer. Research Square (2022). https://doi.org/10.21203/rs.3.rs-1401026/v1, `https://europepmc.org/article/PPR/PPR465713`
90. Wang, J., Luttrell, J., Zhang, N., Khan, S., Shi, N., Wang, M.X., Kang, J.Q., Wang, Z., Xu, D.: Exploring human diseases and biological mechanisms by protein structure prediction and modeling. Translational Biomedical Informatics: A Precision Medicine Perspective pp. 39–61 (2016)
91. Wüthrich, K.: Nmr of proteins and nucleic acids (1986)
92. Xie, H., Thorn, S., Liu, Y.H., Lee, S., Liu, Z., Wang, G., Sinusas, A.J., Liu, C.: Deep-learning-based few-angle cardiac spect reconstruction using transformer. IEEE transactions on radiation and plasma medical sciences **7**(1), 33–40 (2022)
93. Yang, J., Anishchenko, I., Park, H., Peng, Z., Ovchinnikov, S., Baker, D.: Improved protein structure prediction using predicted interresidue orientations. Proceedings of the National Academy of Sciences **117**(3), 1496–1503 (2020)
94. Yang, Y., Yang, J., Shen, L., Chen, J., Xia, L., Ni, B., Ge, L., Wang, Y., Lu, S.: A multi-omics-based serial deep learning approach to predict clinical outcomes of single-agent anti-pd-1/pd-l1 immunotherapy in advanced stage non-small-cell lung cancer. American journal of translational research **13**(2), 743 (2021)
95. Yang, Z., Zeng, X., Zhao, Y., Chen, R.: Alphafold2 and its applications in the fields of biology and medicine. Signal Transduction and Targeted Therapy **8**(1), 115 (2023)
96. Yu, M., Chang, S., Zhang, Y., Jaakkola, T.S.: Rethinking cooperative rationalization: Introspective extraction and complement control. arXiv preprint arXiv:1910.13294 (2019)
97. Zaikis, D., Vlahavas, I.: Tp-ddi: Transformer-based pipeline for the extraction of drug-drug interactions. Artificial Intelligence in Medicine **119**, 102153 (2021)
98. Zhao, C., Shuai, R., Ma, L., Liu, W., Wu, M.: Improving cervical cancer classification with imbalanced datasets combining taming transformers with t2t-vit. Multimedia tools and applications **81**(17), 24265–24300 (2022)
99. Zhou, B., Dey, N., Schlemper, J., Salehi, S.S.M., Liu, C., Duncan, J.S., Sofka, M.: Dsformer: A dual-domain self-supervised transformer for accelerated multi-contrast mri reconstruction. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 4966–4975 (2023)