

Evaluating Protein Binding Interfaces With PUMBA

Azam Shirali¹, Vitalii Stebliankin², and Giri Narasimhan^{1,3,2*}

¹ Bioinformatics Research Group (BioRG), Knight Foundation School of Computing and Information Sciences, Florida International University, Miami, FL, USA

ashir018@fiu.edu

² Euleris LLC, FL, USA

vitalii@euleris.ai

³ Biomolecular Sciences Institute, Florida International University, Miami, FL, USA

giri@fiu.edu

Abstract

Protein–protein docking tools help in studying interactions between proteins, and are essential for drug, vaccine, and therapeutic development. However, the accuracy of a docking tool depends on a robust scoring function that can reliably differentiate between native and non-native complexes. PIsToN is a state-of-the-art deep learning–based scoring function that uses Vision Transformers in its architecture. Recently, the Mamba architecture has demonstrated exceptional performance in both natural language processing and computer vision, often outperforming Transformer-based models in their domains. In this study, we introduce PUMBA (**P**rotein–protein interface **e**valuation with **V**ision **M**amba), which improves PIsToN by replacing its Vision Transformer backbone with Vision Mamba. This change allows us to leverage Mamba’s efficient long-range sequence modeling for sequences of image patches. As a result, the model’s ability to capture both global and local patterns in protein–protein interface features is significantly improved. Evaluation on several widely-used, large-scale public datasets demonstrates that PUMBA consistently outperforms its original Transformer-based predecessor, PIsToN.

1 Introduction

Protein-protein docking is crucial for understanding protein function, elucidating disease mechanisms, and guiding the development of therapies. A computational method for protein docking generally involves two main steps. The first step is the sampling stage, where a large set of candidate conformations is generated. The second step is the scoring stage, during which a scoring function evaluates and ranks these conformations to identify the structure that most closely resembles the native protein complex. However, designing an efficient and reliable scoring function remains a major challenge [21]. Significant advancements have been achieved on addressing this problem, with deep learning (DL) methods frequently surpassing traditional physics-based and knowledge-based techniques. Among deep learning-based methods, PIsToN [23] is a recent tool that demonstrates state-of-the-art performance in comparison to its competitors [21].

*Corresponding author.

PIsToN (evaluating **P**rotein binding **I**nterfaces with **T**ransformer **N**etworks) utilizes hand-crafted chemical, physical, and geometric features extracted from the surfaces of the 3D structures of protein docking models. For each model, it computes the features representing them as a pair of surface patches—one from each interacting protein. Each patch is then transformed into a multi-channel image, with each channel representing a specific feature. The resulting pairs of images is then processed by a Vision Transformer (ViT) [9] to identify patterns that help distinguish native protein pairs from non-native decoys. The ViT embeddings are combined with "hybrid" empirical energy terms, allowing the model to capture the energetic characteristics of protein binding. A multi-attention mechanism then organizes the energy terms and interface features into five distinct groups, each processed by its own transformer network. Finally, the outputs of these networks are integrated through a transformer encoder, which generates the final classification score indicating whether the protein is native or non-native.

Our main contributions: In this work, we introduce PUMBA (**P**rotein-protein interface **e**valuation with **M**amba), an enhanced version of PIsToN that replaces the vision transformer backbone with a Vision Mamba (ViM) architecture [29]. ViM is based on the Mamba architecture [11], which enhances efficient modeling of long-range dependencies and bidirectional context propagation for sequences of image patches. By utilizing a selective state space mechanism, ViM can capture local and global patterns in protein interface representations more efficiently than the attention-based model. Additional innovations in PUMBA include a reformulation that exposes implicit attention matrices, thereby introducing explainability features. These enhancements enable PUMBA to learn discriminative features from multi-channel interface images more effectively, leading to consistent improvements over PIsToN across classification metrics and docking evaluation measures. Our results highlight the potential of state space-based architectures as a powerful alternative to transformers for modeling protein interactions while preserving the strong explainability characteristics of PIsToN through different mechanisms.

2 Related work

2.1 Deep Learning in protein scoring functions

Just as deep learning (DL) has achieved remarkable success in various applications, including text-related tasks (machine translation and sentiment analysis), image processing tasks (object detection and image segmentation), and graph analysis (social network analysis and molecular property prediction), it has also been applied to diverse problems in structural biology, including the design of powerful DL-based scoring functions for protein docking.

Wang et al. [26] developed GNN-DOVE by improving DOVE [27] and utilizing two subgraphs to represent protein complex interfaces. Node features are encoded with one-hot vectors to capture properties like atom type, connection counts, and hydrogen atom numbers, while edges represent covalent and non-covalent bonds. A gate-augmented attention mechanism is used to identify atomic interaction patterns and assess their significance. [18] introduced DeepRank-GNN, an improvement on [19]. This model represents protein-protein interfaces as residue-level graphs, incorporating features like residue type, charge, and buried surface area. It constructs separate intra- and inter-protein subgraphs, processed with a graph attention network (GAT) [15] and an edge-aggregated graph attention network (EGRAT) [15], respectively. The model calculates weighted contributions from neighboring residues to distinguish between native and non-native complexes. [24] developed dMaSIF, enhancing its predecessor, MaSIF [10], by representing protein surfaces as point clouds. Unlike MaSIF, dMaSIF assesses interactions between all atoms without precomputed features. It calculates geometric and chemical features for each point on

the interface, which are processed through convolutional layers to learn interaction patterns.

AlphaFold3 (AF3) [2] significantly advanced the prediction of biomolecular structure by modeling proteins, nucleic acids, small molecules, and ions, and integrating a scoring function directly into its architecture for protein–protein complexes. Utilizing the Pairformer module, AF3 captured residue-level relationships and employed a diffusion-based [12] refinement network for high-accuracy 3D structures. Key metrics include the predicted Template Modeling score (ipTM) [28] for interchain accuracy, predicted Template Modeling score (pTM) [28] for global structural accuracy, predicted Local Distance Difference Test (pLDDT) [16] for per-atom confidence, and Predicted Aligned Error (PAE) for positional uncertainty. These metrics enable effective ranking and evaluation of predicted models. For a broader comparison of scoring functions, see [21].

2.2 Deep Learning in machine vision

Before the emergence of transformers in vision tasks, convolutional neural networks (CNNs) were the dominant architecture, achieving significant success in areas such as image classification, detection, and segmentation. CNNs effectively capture local spatial patterns through hierarchical feature extraction and learn translation-invariant features, making them the standard architecture for computer vision. However, the impressive success of transformers in natural language processing (NLP), driven by their self-attention mechanism and capability to model long-range dependencies, sparked interest in applying them to vision tasks. This interest led to the development of ViT [9], which treats an image as a sequence of patches and utilizes transformer-based attention to capture global context. ViT quickly demonstrated competitiveness with, and often superiority over, CNNs, especially with large-scale training data, due to its flexibility in modeling complex relationships across image regions. Despite these advantages, transformers, including ViT, have inherent limitations, such as quadratic complexity with respect to sequence length, high memory consumption, and reduced efficiency on very large or high-resolution inputs.

To address these challenges, Mamba [11] introduced a selective state space model (SSM) that processes sequences with linear complexity, enabling greater scalability while preserving strong modeling capacity. The success of Mamba in NLP tasks has inspired ViM [29] to adapt this approach to images by replacing the self-attention mechanism in ViT-like architectures with efficient SSM-based blocks. This modification enables faster training and improves handling of long or high-resolution sequences, without sacrificing accuracy.

2.3 Explainability and hidden attention in ViM

“Explainability” refers to the ability of a deep learning model to provide insight into the reasoning behind its decisions. In transformer-based models, this is often analyzed through the self-attention mechanism, which models dependencies between tokens by assigning weights that reflect their relative importance within the input. By examining the attention patterns, one can identify which parts of the input most influence the model’s output, as explored in [6] and [1].

Unlike transformers, which produce explicit attention matrices that can be directly analyzed for interpretability, Mamba models do not generate analogous attention structures, making their internal reasoning less transparent. The absence of interpretability tools further limits their reliable deployment in sensitive domains where explainability is essential. However, [3] demonstrates that Mamba models can be reformulated to expose implicit attention matrices. This study demonstrates that Mamba implements a variant of causal self-attention. In contrast to transformers, where each attention head produces a single attention matrix, a single Mamba layer generates substantially more attention matrices due to its multiple channels, enabling finer-grained tracking of token-to-token influence.

3 Model Design and Architecture

Building on PIsToN [23], PUMBA replaces the ViT backbone with ViM, an SSM-based architecture introduced for visual sequence modeling. Similar to PIsToN, PUMBA is trained to distinguish native-like from non-native docking conformations by learning from the surface characteristics of binding interfaces.

Given the PDB format [5] of a protein-protein complex, the structure is first refined using FireDock [4], which also computes binding free energy, including Van der Waals, desolvation, insiderness, hydrogen bonds, disulfide bonds, electrostatics, π -stacking, cation- π and aliphatic interactions [23]. The refined complex is then cropped within a distance r_{surf} from the interaction center, defined as the geometric center of all atoms located within 5\AA of any atom in the binding partner. The solvent-excluded surface is subsequently triangulated and rescaled to 1\AA granularity. Surface features are then computed over circular patches on the interacting regions, where a patch consists of surface vertices within geodesic distance r_{surf} from the interaction center. For every surface point, six features are extracted which are shape index, curvature, hydrogen bonding potential, charge, hydrophathy, and Relative Accessible Surface Area (RASA). Additionally, a ‘‘patch dist’’ feature is derived from the Euclidean distances between corresponding points on paired patches. RASA values are obtained using DSSP v2.3 [25]. All features are converted into multi-channel 2D images by projecting patch surface points onto a plane using multidimensional scaling (MDS) [17], producing a $2r_{\text{surf}} \times 2r_{\text{surf}}$ grid for each feature, with pixel intensities smoothed over nearest neighbors (see [22, 23] for details). These multi-channel images are then processed by a ViM backbone, which treats sequences of image patches as ordered tokens, enabling efficient modeling of both local and long-range dependencies across surface features.

The multi-channel images obtained from the interface are provided as inputs to a ViM network. The ViM architecture follows the introduced in [29], with modifications to accommodate N input channels corresponding to the number of computed surface features, rather than the conventional RGB channels. Each image is subdivided into square patches of size l , producing $P = (a/l)^2$ mini-patches, where a denotes the image size (i.e. the length of each side in pixels). These patches are then linearly projected into M -dimensional token embeddings [23]. A class token is inserted at the center of the embedding sequence, as this position has been shown to yield optimal performance [29]. The output provides a ranking score for each protein docking model, distinguishing between native-like and non-native conformations. As in PIsToN (Fig.1a), related surface features and energy terms are organized into five functional groups: shape, RASA, charge, hydrogen bonds, and hydrophathy. In PUMBA (Fig.1b), each group is processed by an independent hybrid module in which the ViT encoder from PIsToN is replaced with a ViM encoder, followed by a fully connected layer. The resulting representations are then aggregated through a final ViM encoder to generate a binding score for each input docking model.

The challenge with image-like protein surface patches is that they are inherently non-sequential. In attention-based models like ViT, the attention mechanism can directly access all tokens simultaneously, making spatial order less critical as long as positional encodings are included. In contrast, ViM updates its hidden state recurrently based on both the previous state and the current input. This sequential nature makes the input order important; early states carry less contextual information than later ones when processed in a single forward pass. To address this challenge, ViM utilizes Bidirectional idea, which consists of two separate SSMs within each encoder block: the Forward and Backward SSM to process the patch embeddings in the forward and reverse orders. This approach effectively allows future patch information to be passed backward through the model. The outputs of these two SSMs are fused, enabling each token to access both past and future contexts. As suggested in the ViM architecture, in

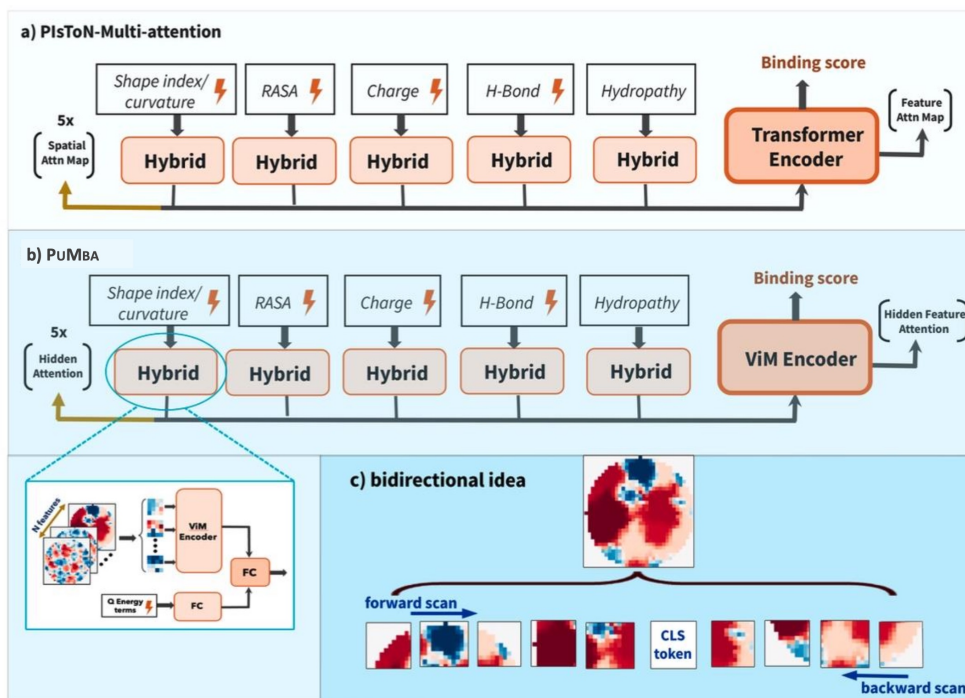


Figure 1: Overview of PIsToN and PUMBA architectures: (a, b) Multi-attention architecture with hybrid feature processing in PIsToN and PUMBA, (c) bidirectional scans of patches of each feature in forward and backward SSM blocks in PUMBA. The CLS token is inserted into the middle of the patch sequence to maximize bidirectional context utilization.

PUMBA we added a class token as an additional token positioned in the middle to enhance classification performance. Fig.1c illustrates an example of patch embedding, where the image of the hydropathy feature is split into small patches with a class token placed in the middle. In PUMBA, the internal operations of each encoder layer—including SSM computation, bidirectional fusion, and positional encoding—are consistent with those described in the ViM paper [29].

To incorporate the explainability features of PIsToN, we follow the framework established in [3]. This ensures that the important feature of explainability is not sacrificed by replacing the transformer-based model with a Mamba-based model. As in PIsToN, our model offers a two-stage attention interpretation: at the token level within each branch, it captures the token-to-token dependency dynamics across the intermediate Mamba layers, identifying which pixels the model pays more attention to for each feature. At the feature level across branches (in the final ViM encoder), it determines which feature(s) contribute most significantly to the final decision. An example of this interpretation is provided in Section 4. We follow the training procedure described by [23], adopting the same training dataset. The loss formulation remains unchanged, combining binary cross-entropy, supervised contrastive loss [13], and margin ranking loss [7]. The training strategy is also replicated, including input preprocessing, optimizer choice, and learning rate schedule, to ensure consistency with their reported performance.

4 Experimental Results

To evaluate the effectiveness of the proposed PUMBA architecture, we conducted a series of experiments comparing its performance to the original PIsToN using popular and well-known datasets. We employed eight datasets reviewed in [21] to ensure a broad and diverse evaluation of PIsToN and PUMBA. These datasets span widely used benchmark collections, covering a variety of docking scenarios and difficulty levels. More details about each dataset can be found in [21]. Training and testing were performed on a system equipped with an NVIDIA GeForce GTX 1080 Ti (8 GB GPU memory), 256 GB RAM, and a 28-core Intel Xeon E5-2650 CPU.

We evaluated PUMBA against its predecessor PIsToN across the same eight widely used protein-protein docking benchmark datasets: CAPRI Score v2022 (Difficult/Easy), CAPRI Score Refined, BM4, BM5, Dockground, PDB 2023, and MaSIF test [21]. The evaluation considered several classification metrics, including the area under the ROC curve (AUC ROC), average precision (AP), balanced accuracy (BA), F1 score, precision, and recall.

Table 1 demonstrates that PUMBA consistently matched or outperformed PIsToN across various datasets, with particularly notable gains in the challenging datasets. For instance, in the CAPRI v2022 Difficult dataset, AUC ROC increased by 4.06%, AP by 7.47%, precision by 5.60%, and recall by 6.03%. In Dockground dataset, PUMBA achieved a more than two-fold improvement in AP (from 10.71 to 25.33) along with a higher F1 score (2.58 increase). Even in datasets where PIsToN already performed at a high level, such as BM5 and PDB 2023, PUMBA maintained or slightly surpassed the performance.

The success rate evaluation in Table 2 further supports these improvements. Success rate indicates how frequently a docking method yields at least one model of acceptable quality among its top K predicted docking models. A higher success rate directly improves drug discovery by enhancing early enrichment, reducing the number of compounds requiring experimental testing, and accelerating the design-make-test cycle. Table 2 shows that in CAPRI v2022 Difficult, the top 1 success rate increased from 17% to 22%, and in Dockground, it rose from 15% to 32%. Gains were also evident in BM4, where the top 1 rate improved from 54% to 57%, and in CAPRI v2022 Easy, where it increased from 41% to 48%. In the top 100 and 200, PUMBA achieved perfect or near-perfect success rates across several datasets, further demonstrating its ability to effectively prioritize native-like docking models.

To assess the quality of protein docking models, predictions are categorized as incorrect, acceptable, medium, or high according to the CAPRI (Critical Assessment of Predicted Interactions) criteria [14] using the CAPRI-Q tool [8]. Table 3 shows the ability of PIsToN and PUMBA in ranking docking models of varying quality among different top-k predictions. Across all eight benchmark datasets, PUMBA consistently identifies a greater number of acceptable-to-high-quality models than PIsToN, particularly within the top 10 and top 100 predictions. For example, in BM5, PUMBA identified 44 acceptable-quality docking models in the top 10 and 78 in the top 100, compared to 40 and 70 for PIsToN. In CAPRI Score v2022 (Easy), PUMBA found 92 acceptable models in the top 100, exceeding PIsToN’s 66. Overall, the models ranked by PUMBA are of higher quality than those by PIsToN.

To assess the ability of PUMBA to be interpretable and a fair comparison, we followed the explainability protocol of PIsToN while adapting it to PUMBA’s hidden attention values. For each protein complex, two sets of attention maps were extracted (Fig.1b): hidden attention maps and hidden feature attention maps from the final binding score module. These maps were averaged across all heads and layers, and the z-score was calculated to highlight the most influential pixels and regions ($z\text{-score} \geq -1.96$, $p \leq 0.05$). The resulting high-importance regions were overlaid onto the corresponding feature images to visualize feature(s) contributed most

Dataset	Method	AUC ROC	AP	BA	F1	Precision	Recall
CAPRI v2022 (Difficult)	PIsToN	64.77	3.02	63.05	66.16	3.47	70.49
	PUMBA	68.83	10.49	62.71	67.23	9.07	76.52
CAPRI v2022 (Easy)	PIsToN	80.15	46.24	73.08	60.51	55.89	65.95
	PUMBA	79.91	48.15	72.12	63.46	58.19	69.80
CAPRI Refined	PIsToN	81.25	24.86	73.24	42.21	30.37	69.16
	PUMBA	79.95	25.19	74.21	39.38	29.15	76.56
BM4	PIsToN	70.84	36.46	62.71	43.67	53.75	36.77
	PUMBA	72.50	37.27	64.64	43.04	52.27	45.13
BM5	PIsToN	91.84	23.64	74.90	45.14	38.16	55.24
	PUMBA	92.02	22.31	72.74	46.86	41.10	56.49
Dockground	PIsToN	72.04	10.71	60.40	22.13	15.62	37.92
	PUMBA	72.94	25.33	64.12	24.71	18.15	39.46
PDB 2023	PIsToN	93.86	45.85	78.21	66.05	78.89	56.80
	PUMBA	93.91	46.11	80.23	66.72	75.13	58.66
MaSIF test	PIsToN	93.55	81.08	85.43	85.97	87.72	82.40
	PUMBA	93.42	81.45	85.97	86.34	87.69	84.12

Table 1: Evaluation of PIsToN and PUMBA across eight benchmark datasets.

Dataset	Method	top1	top10	top25	top100	top200
CAPRI v2022 (Difficult)	PIsToN	17	28	46	71	78
	PUMBA	22	31	47	69	85
CAPRI v2022 (Easy)	PIsToN	41	82	87	94	97
	PUMBA	48	83	87	96	100
CAPRI Refined	PIsToN	38	69	76	76	100
	PUMBA	40	69	75	81	100
BM4	PIsToN	54	89	89	94	100
	PUMBA	57	90	90	92	100
BM5	PIsToN	66	93	100	100	100
	PUMBA	64	95	100	100	100
Dockground	PIsToN	15	55	81	100	100
	PUMBA	32	59	83	100	100
PDB 2023	PIsToN	88	96	98	100	100
	PUMBA	89	99	99	100	100

Table 2: Success rates (%) of PIsToN and PUMBA on eight different datasets.

to the predicted score. Fig.2 illustrates an example of the explainability of PUMBA, which is consistent with PIsToN used. Notably, PUMBA focused on the same features as PIsToN when deciding about classifying complexes, demonstrating that the model maintains strong explainability while delivering improved predictive accuracy. For this example, both models placed greater attention on the hydrophathy feature when predicting a native protein complex. Furthermore, by mapping the high-importance interface pixels back to the original structure in PyMOL [20], we observed that the residues highlighted for the hydrophathy feature were predominantly hydrophobic, confirming the correctness of our model’s attention and showing

Dataset	Method	acceptable			medium			high		
		top1	top10	top100	top1	top10	top100	top1	top10	top100
CAPRI v2022 (Difficult)	PIsToN	17	28	67	0	3	35	0	0	0
	PUMBA	19	32	71	2	6	33	0	0	3
CAPRI v2022 (Easy)	PIsToN	17	51	84	20	66	87	2	20	30
	PUMBA	16	53	92	24	65	89	5	23	37
CAPRI Refined	PIsToN	7	46	61	30	46	53	0	7	23
	PUMBA	12	45	63	32	46	57	0	7	26
BM4	PIsToN	5	26	42	5	26	42	5	21	42
	PUMBA	15	33	41	11	28	40	6	23	43
BM5	PIsToN	6	40	80	40	60	66	20	73	100
	PUMBA	5	44	78	38	64	72	18	77	100
Dockground	PIsToN	0	1	28	24	71	98	1	8	17
	PUMBA	1	4	22	31	75	98	6	15	31
PDB 2023	PIsToN	1	44	48	9	30	30	76	92	96
	PUMBA	2	36	48	13	33	41	76	95	98

Table 3: Quality of docking models scored by PIsToN and PUMBA across eight benchmark datasets. Values represent the number of docking models classified into each CAPRI quality category among the top 1, 10, and 100 predictions.

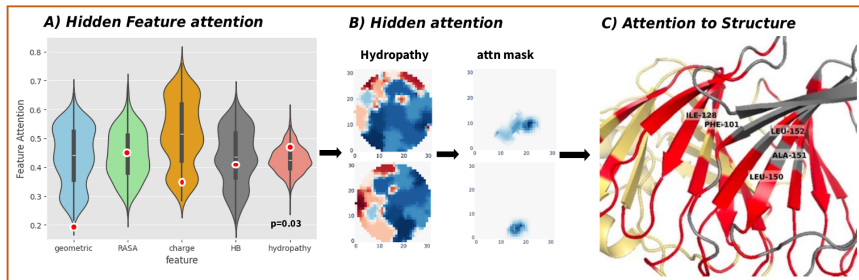


Figure 2: Explainability in PUMBA with homodimer (PDB ID 1J3R, chains A and B).

that its focus aligns precisely with the biologically relevant interaction regions.

5 Conclusions

We introduced PUMBA, a scoring tool designed to rank protein docking models. The ViT backbone of PIsToN was replaced with the bidirectional ViM architecture in PUMBA. By utilizing selective state space modeling, PUMBA effectively captures both global and local information. Hidden attention analysis further supports interpretability, even in the absence of explicit attention maps. Experimental results show that PUMBA consistently outperforms PIsToN across standard metrics and the CAPRI docking quality criteria. These findings underscore the potential of Mamba-based architectures to advance the assessment of biomolecular docking.

Acknowledgments: Sincere thanks to the PIsToN designers for sharing their insights.

Data and code availability: Datasets and code are available at: <https://zenodo.org/records/12681335> and <https://github.com/Azam-Shi/PuMba>, respectively.

References

- [1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*, 2020.
- [2] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Babrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*.
- [3] Ameen Ali, Itamar Zimerman, and Lior Wolf. The hidden attention of mamba models. *CoRR*, abs/2403.01590, 2024.
- [4] Nelly Andrusier, Ruth Nussinov, and Haim J Wolfson. FireDock: fast interaction refinement in molecular docking. *Proteins: Structure, Function, and Bioinformatics*, 69(1):139–159, 2007.
- [5] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The Protein Data Bank. *Nucleic acids research*, 28(1):235–242, 2000.
- [6] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 782–791, 2021.
- [7] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32, 2019.
- [8] Keeley W Collins, Matthew M Copeland, Guillaume Brysbaert, Shoshana J Wodak, Alexandre MJJ Bonvin, Petras J Kundrotas, Ilya A Vakser, and Marc F Lensink. CAPRI-Q: The CAPRI resource evaluating the quality of predicted structures of protein complexes. *Journal of molecular biology*, 436(17):168540, 2024.
- [9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [10] Pablo Gainza, Freyr Sverrisson, Frederico Monti, Emanuele Rodola, Davide Boscaini, Michael M Bronstein, and Bruno E Correia. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nature methods*, 17(2):184–192, 2020.
- [11] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.
- [12] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35:26565–26577, 2022.
- [13] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschiot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.
- [14] Marc F Lensink, Raúl Méndez, and Shoshana J Wodak. Docking and scoring protein complexes: Capri 3rd edition. *Proteins: Structure, Function, and Bioinformatics*, 69(4):704–718, 2007.
- [15] Sazan Mahbub and Md Shamsuzzoha Bayzid. EGRET: edge aggregated graph attention networks and transfer learning improve protein–protein interaction site prediction. *Briefings in Bioinformatics*, 23(2):bbab578, 2022.
- [16] Valerio Mariani, Marco Biasini, Alessandro Barbato, and Torsten Schwede. IDDT: a local superposition-free score for comparing protein structures and models using distance difference tests.

- Bioinformatics*, 29(21):2722–2728, 2013.
- [17] Al Mead. Review of the development of multidimensional scaling methods. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 41(1):27–39, 1992.
 - [18] Manon Réau, Nicolas Renaud, Li C Xue, and Alexandre MJJ Bonvin. DeepRank-GNN: a graph neural network framework to learn patterns in protein–protein interfaces. *Bioinformatics*, 39(1):btac759, 2023.
 - [19] Nicolas Renaud, Cunliang Geng, Sonja Georgievska, Francesco Ambrosetti, Lars Ridder, Dario F Marzella, Manon F Réau, Alexandre MJJ Bonvin, and Li C Xue. DeepRank: a deep learning framework for data mining 3D protein-protein interfaces. *Nature communications*, 12(1):7068, 2021.
 - [20] L Schrödinger. The PyMOL molecular graphics system, version 1.8. (*No Title*), 2015.
 - [21] Azam Shirali, Vitalii Stebliankin, Ukesh Karki, Jimeng Shi, Prem Chapagain, and Giri Narasimhan. A comprehensive survey of scoring functions for protein docking models. *BMC bioinformatics*, 26(1):25, 2025.
 - [22] Azam Shirali, Vitalii Stebliankin, Jimeng Shi, Prem Chapagain, and Giri Narasimhan. Evaluating and scoring ebolavirus protein-protein docking models using piston. *arXiv preprint arXiv:2511.13583*, 2025.
 - [23] Vitalii Stebliankin, Azam Shirali, Prabin Baral, Jimeng Shi, Prem Chapagain, Kalai Mathee, and Giri Narasimhan. Evaluating protein binding interfaces with transformer networks. *Nature Machine Intelligence*, 5(9):1042–1053, 2023.
 - [24] Freyr Sverrisson, Jean Feydy, Bruno E Correia, and Michael M Bronstein. Fast end-to-end learning on protein surfaces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15272–15281, 2021.
 - [25] Wouter G Touw, Coos Baakman, Jon Black, Tim AH Te Beek, Elmar Krieger, Robbie P Joosten, and Gert Vriend. A series of PDB-related databanks for everyday needs. *Nucleic acids research*, 43(D1):D364–D368, 2015.
 - [26] Xiao Wang, Sean T Flannery, and Daisuke Kihara. Protein docking model evaluation by graph neural networks. *Frontiers in Molecular Biosciences*, 8:647915, 2021.
 - [27] Xiao Wang, Genki Terashi, Charles W Christoffer, Mengmeng Zhu, and Daisuke Kihara. Protein docking model evaluation by 3D deep convolutional neural networks. *Bioinformatics*, 36(7):2113–2118, 2020.
 - [28] Jinrui Xu and Yang Zhang. How significant is a protein structure similarity with TM-score= 0.5? *Bioinformatics*, 26(7):889–895, 2010.
 - [29] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.