DIFFMASIF: Surface-based Protein-Protein Docking with Diffusion Models

Freyr Sverrisson*[1] Mehmet Akdel*[2] Dylan Abramson[2] Jean Feydy[3]
Alexander Goncearenco[2] Yusuf Adeshina[2] Daniel Kovtun[2] Céline Marquet[2]
Xuejin Zhang[2] David Baugher[2] Zachary Carpenter[2] Luca Naef[2]
Michael M. Bronstein[2] Bruno Correia[1]

[1]École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland
firstname.surname@epfl.ch
[2]VantAI, New York, NY 10003, United States
firstname@vant.ai
[3]Équipe Inria HeKA, PariSanté Campus, 2 - 10 Rue d'Oradour-sur-Glane, 75015 Paris, France
jean.feydy@inria.fr
* These authors contribted equally to this work.

Abstract

Predicting protein-protein complexes is a central challenge of computational structural biology. Existing state-of-the-art methods rely on co-evolution learned on large amino acid sequence datasets and thus often fall short on both transient and engineered interfaces (which are of particular interest in therapeutic applications) where co-evolutionary signals are absent or minimal. To address this, we introduce DIFFMASIF, a novel score-based diffusion model for rigid protein-protein docking. Instead of sequence-based features, DIFFMASIF uses a protein molecular surface-based encoder-decoder architecture to effectively learn physical complementarity. The encoder uses learned geometric features extracted from protein surface point clouds. It directly learns binding site complementary through prediction of contact sites as an auxiliary loss, and also allows for specification of known binding sites during inference. It is followed by a decoder predicting rotation and translation via $\mathrm{SO}(3)$ diffusion. We show that DIFFMASIF achieves state-of-the-art among deep Llearning methods for rigid body docking, in particular on structurally novel interfaces and low sequence conservation. This provides a significant advance towards accurate modelling of low co-evolution protein interactions and their many practical applications.

1 Introduction

Proteins orchestrate most cellular functions, many of which are derived from the way in which they mutually interact. A protein's three-dimensional structure directly defines its function and interactions with other molecules. Recent groundbreaking work (Jumper et al., 2021) showed that deep learning methods could be used to predict a significant fraction of protein structures to near-experimental accuracy using the protein sequence and information about its evolutionary history. The accurate prediction of protein-protein interactions, however, still remains an open challenge (Ozden et al., 2023).

Traditionally, protein-protein complexes are structurally modelled through *docking*, where one attempts to predict the conformations of proteins in the complex from the individual unbound structures of the interacting proteins. Protein-protein docking methods typically involve constructing

Machine Learning for Structural Biology Workshop, NeurIPS 2023.

a pseudo-energy function derived from physical principles fitted on known protein-protein complexes, potentially combined with known templates and heuristics (Vajda & Kozakov, 2009). Black-box stochastic optimization techniques are then used to search for minima within the energy functions. However, the search space of all possible conformations including backbone and side chain torsions is infeasible to explore exhaustively (Harmalkar & Gray, 2021), therefore sampling techniques such as Monte Carlo simulations are applied (Marze et al., 2018). As an initial approximation, *rigid-body docking* (where the relative pose of one protein with respect to the other is determined) is often performed, sometimes followed by an iterative refinement allowing backbones and side chains to relax in presence of its interacting partner (Desta et al., 2020).

Current deep learning methods for protein-protein docking typically build on the same principles as structure prediction, leveraging sequence representations trained via masked-language modelling on large evolutionary sequence databases(Ketata et al., 2023; Jumper et al., 2021). While these tend to perform well in case of co-evolved stable interfaces, they fail to capture the many structurally diverse, transient and flexible interactions many proteins participate in. In addition, de-novo designed interfaces as well as heavily recombined sequences such as antibody hypervariable regions, which are very commonly used for therapeutic applications, lack co-evolution data. This leads to subpar performance of existing deep learning approaches (Ozden et al., 2023).

It is known, however, that all protein interactions are mediated and understandable through steric and electrostatic complementarity of the interface (Lawrence & Colman, 1993; Jones & Thornton, 1996). Early rigid-body docking approaches (Katchalski-Katzir et al., 1992) in fact relied on implicit representations of protein surfaces and by using fast Fourier transform of a correlation function to assess the degree of shape complementarity. Later, in a deep learning context, learned protein surface representations (*molecular surface interaction fingerprinting*, or *MaSIF*), which can capture this steric and electrostatic complementarity have proven to be powerful in predicting protein interactions Gainza et al. (2020); Sverrisson et al. (2021); Gainza et al. (2023). In this paper, as a way to address the limitations of co-evolution based approaches, we propose DIFFMASIF, the first score-based diffusion model for rigid-body docking using a versatile surface representation of proteins.

Main contributions. DIFFMASIF is the first protein surface-based diffusion model, addressing the limitations of current co-evolution reliant models. Second, we propose a novel joint-training strategy for simultaneous binding site and protein-protein pose prediction. This enables site prediction at inference time as well, as an easy way to add conditioning for site-specific docking. Third, we use a novel encoder-decoder architecture that combines a surface-based vector-neuron (DGCNN) encoder with $\mathrm{E}(3)$-equivariant graph convolution decoder, trained to learn binding site structural complementarity and rigid body docking via $\mathrm{SO}(3)$ diffusion, respectively. Finally, we show state-of-the-art rigid body docking results, surpassing current machine learning methods on structurally novel interfaces, as well as on docking of predicted AlphaFold monomers.

2 Methods

2.1 Data

To address the limitations of the typically used sequence-centric and small benchmark set, Dataset of Interacting ProteinS (DIPS) (Townshend et al., 2019), with a test split comprised of Docking Benchark 5 (DB5) (Vreven et al., 2015), we make use of a recently introduced dataset, PINDER with a novel splitting strategy based on structural interfaces specifically designed to assess the protein-protein docking task (Akdel et al., 2023). We detail the methods and splits produced by PINDER in Appendix section **3**. The large test set PINDER-xl consists of 1,756 dimers representing novel structural interfaces, and the PINDER-af2 set of 72 dimers representing novel structural interfaces released after the AlphaFold-Multimer (AF2MM) training date. To further test our method on the more realistic use-case of using predicted structures for docking, for 90% of the complexes in the PINDER-xl test set we use pairs of corresponding AlphaFold monomer predicted structures as input for docking.

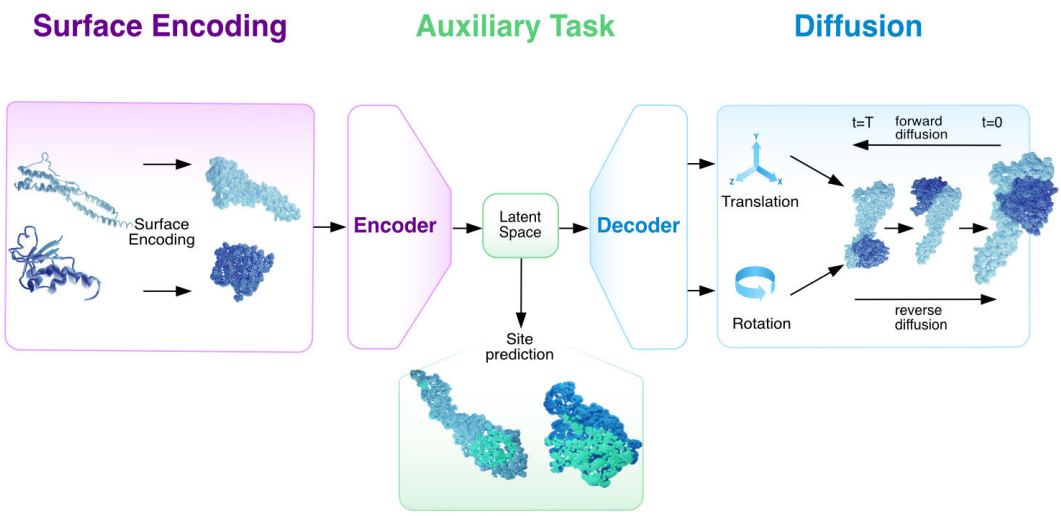

Figure 1: Overview of the DIFFMASIF method. Protein surface point clouds are generated and fed into a encoder-decoder network. The model learns both binding site prediction via an encoder, and denoising a reverse diffusion process over rotations and translations via a decoder.

2.2 Model Architecture

The architecture of DIFFMASIF has two main components: an encoder and a decoder. Our study adopts a diffusion process approach akin to the methodologies presented in Corso et al. (2022); Ketata et al. (2023), described in Appendix section **1.2**.

Encoder The encoder takes atom level features as input for both proteins, containing one hot encoding of atom types and coordinates, which are passed to a dMaSIF layer to generate (1) surface normals, (2) surface point coordinates, and (3) scalar embeddings from dMaSIF's geodesic convolution layer, which are further scaled using MLP layers. k nearest neighbor graphs ($k = 12$) connecting these surface nodes are constructed for the receptor (stationery chain) and ligand (movable chain) separately and no cross information is communicated at this stage.

Binding-Site Auxiliary Task Building off the insights from DockGPT (McPartlon & Xu), where including contact points improved complex prediction, we sought to construct a loss that differentiates interaction site prediction from pose prediction. This auxiliary loss passes the result of the dMaSIF MLP from both the ligand and the receptor through a cross-attention mechanism to predict whether a surface node is part of the binding site or not. True binding site nodes are defined as those < 3Å from the other surface. For both the ligand and receptor, only the top 512 predicted binding site nodes each are used for the decoder, reducing the high compute and memory required by the decoders' tensor-product convolution layers.

Decoder The decoder works on the joint PPI graph consisting of the top 512 predicted binding site nodes of both the ligand and the receptor. The first component of the decoder is a DCGNN (with vector neuron layers) (Wang et al., 2019; Deng et al., 2021) which takes coordinates and normal vectors as input and outputs higher-dimensional vector embeddings. The vector features, surface coordinates, and surface scalar features are then provided, to an E(3)-equivariant graph convolution layer constructed using the E3nn library (Geiger & Smidt, 2022). The final output of the decoder is the prediction of the translation and rotation required for the ligand coordinates.

Losses The combined denoising score loss and auxiliary binding loss is:

$$\mathscr{L} = \lambda \text{BCELoss}(\hat{c}, c) + S(s_\theta(x(t), \text{argkmax}_{512}(\hat{c}), \phi, \psi),$$

where $\hat{c}$ and c are the predicted contact probabilities and ground truth contacts respectively, S is the denoising score loss used in DiffDock-PP, s_θ is the decoder model and ϕ and ψ are the true rotation and translation scores sampled at time step t from $p(x_\phi(t)|x_\phi(0))$ and $p(x_\psi(t)|x_\psi(0)$. Instead of

	DockQ CAPRI classification								
	Apo			Holo			Predicted		
Method	Acceptable	Medium	High	Acceptable	Medium	High	Acceptable	Medium	High
FRODOCK	**35.41**	**20.7**	**8.48**	96.58	95.56	91.12	38.22	**31.81**	11.68
HDOCK	25.94	14.96	8.23	**98.01**	**97.72**	**97.04**	36.25	30.41	**13.4**
PatchDock	15.21	7.73	2.74	80.75	77.9	62.98	26.41	20.83	5.46
GeoDock	2.99	0.0	0.0	12.76	1.37	0.0	5.84	0.76	0.0
DockGPT	14.46	6.48	1.25	42.2	33.54	21.3	31.24	20.83	4.19
EquiDock	1.25	0.75	0.75	0.17	0.0	0.0	1.59	1.21	0.76
DiffDock-PP	21.45	4.24	0.5	52.39	30.81	10.31	25.52	12.7	1.71
DIFFMASIF	22.94	7.48	2.49	58.83	40.72	16.69	**47.11**	25.84	3.11

Table 1: Table shows the complex prediction metrics for traditional physics docking tools FRODOCK, HDOCK and PatchDock, and machine learning tools EquiDock, DiffDock-PP, DockGPT and DIFFMASIF on the PINDER (PINDER-xl) benchmark set. We report the percentage of systems covered by DockQ CAPRI hit categories as "Acceptable" (or higher), "Medium" (or higher) and "High" solutions. We report the metrics as *oracles* for the generative machine learning tools, such as DiffDock-PP and DIFFMASIF and the traditional docking tools after generating 40 samples for each system.

the full noise-perturbed coordinates, the decoder model receives perturbed coordinates masked to only include the 512 most likely contacts at time step $x(t)$. We balance these two losses with λ, a weighting term.

Training We trained DIFFMASIF using the holo protein-protein pairs obtained from the PINDER training set (236,128 systems), after subjecting them to various PINDER dataloader filters in order to remove outliers. Specifically, we removed systems with elongated structures, chains with over 800 residues, and less than 4 atom types (all available filters are detailed in Appendix section **3**). This resulted in 134,278 total holo pairs. We adapt the dynamic noise sampling from DiffDock-PP, where we sample noise for each system once per epoch. In addition, we also compute the auxiliary loss on the binding site predictions, detailed in Appendix Algorithm **1**.

Inference We performed denoising diffusion and generated 40 poses for each PINDER-xl and PINDER-af2 system. Prior to this, we optimized the magnitude of translation and rotation perturbations at each step by maximizing the DockQ scores of 10 PINDER validation systems. This optimization resulted in scaling the perturbations by a factor of 6 for translation and 3 for rotation.

3 Results

We benchmarked DIFFMASIF against popular physics-based and deep learning-based methods using the PINDER-xl test set (Appendix section **3.**). The deep learning-based methods used were retrained on the PINDER training set, as described in Akdel et al. (2023). As it is difficult to retrain the AF2MM co-folding method with our data splits, we compare to this method using the PINDER-af2 hold-out set (Appendix Table **2.**). Each pose generated by a particular docking method was superposed to the reference pose and evaluated using the CAPRI classification of DockQ scores, based on the composite score that also encompasses Fnat – the fraction of native contacts, interface RMSD and ligand RMSD (Basu & Wallner, 2016). For DIFFMASIF, we generated 40 poses per complex with a reverse ODE using 40 steps. For all comparisons of DIFFMASIF, EquiDock, and DiffDock-PP, we used the pose with the best DockQ out of the top 40 poses (defined as the *oracle* approach). These results are shown in 1. A confidence model trained to rank these poses is an important future direction to better enable DIFFMASIF application.

3.1 DIFFMASIF outperforms co-evolution based docking methods

Similar to DIFFMASIF, we generated 40 poses per test complex with DiffDock-PP with a reverse SDE using 40 steps, and low temperature sampling with default parameters. As seen in Table 1, DIFFMASIF shows consistently better scores than DiffDock-PP and returns more acceptable complexes (see Figure 3A for an example).

In addition, Figure 2 demonstrates the percentage of acceptable poses returned by DiffDock-PP and DIFFMASIF as a function of the average number of effective sequences (N_{eff}, as calculated by HHSuite (Steinegger et al., 2019)) available per complex. This can be seen as a measure of co-evolution, as complexes with a high detectable co-evolutionary signal would be expected to have deep MSAs. Both DiffDock-PP and AF2MM leverage this information implicitly or explicitly. DiffDock-PP leverages ESM2-pretrained embeddings which learn co-evolution implicitly by learning on large sequence databases, while AF2MM directly learns to predict complexes given paired MSAs. From Figure 2, it is clear that DiffDock-PP performance drops when the average N_{eff} goes below 5, while DIFFMASIF is not affected. Thus, DIFFMASIF is a highly complementary approach to rigid protein docking without reliance on co-evolutionary signals.

Unlike the rigid body docking methods described here, AlphaFold2-Multimer (Evans et al., 2021) takes the sequences of both proteins as input and co-folds them into a complex structure, with a heavy reliance on co-evolutionary signals between the interfaces involved. We ran the default AF2MM pipeline (with the latest template date set to September 2022) for the PINDER-af2 benchmark set of 72 complexes (Appendix Table **1.**). This resulted in 5 models per complex from which we took the model with the highest DockQ score, i.e an oracle approach. AF2MM returned unacceptable models for 23 of these complexes. DIFFMASIF predictions were acceptable or better for 8 of these, again demonstrating the complementarity of our approach for difficult interfaces (see Figure 3B for an example, and additional PINDER-af2 benchmark in Appendix Table **2.**.

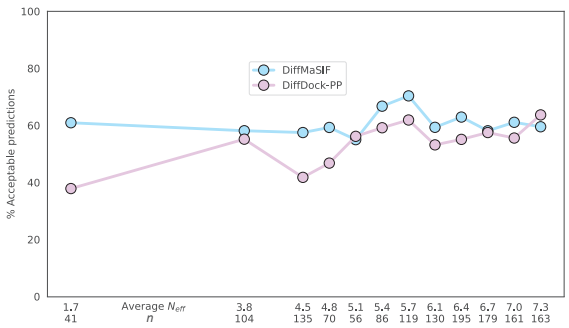

Figure 2: Percentage of acceptable DiffDock-PP and DIFFMASIF complexes across different N_{eff}.

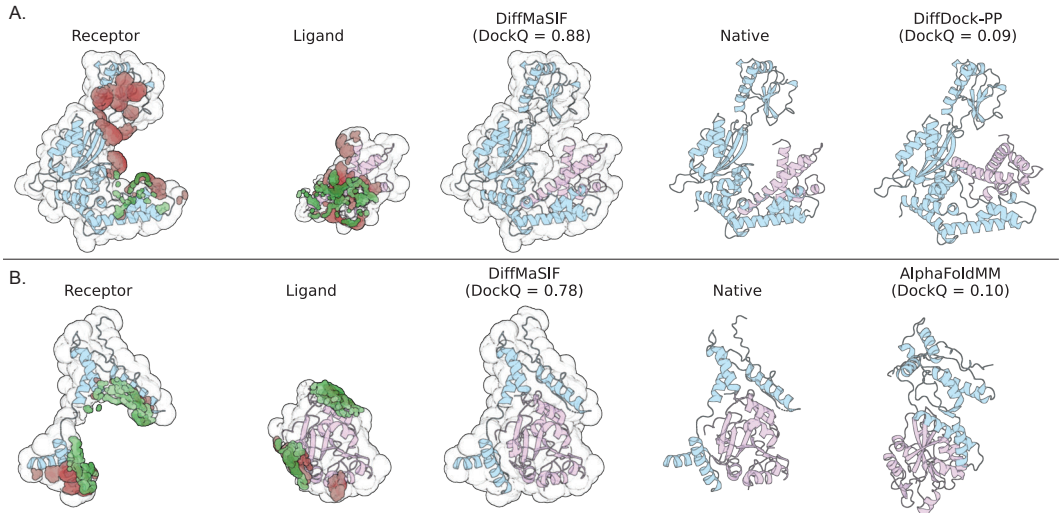

Figure 3: Docking of PDB ID: 6K3B (**A**) and 8FZZ (**B**). The two leftmost columns show DIFFMASIF's ability to identify binding sites, with correctly identified interface points in green, and the rest in red. The third and fourth columns show the DIFFMASIF predicted docking pose and the ground truth pose, while the fifth column shows the DiffDock-PP predicted pose for **A** and the AF2MM predicted complex for **B**.

3.2 DIFFMASIF outperforms other tools in docking of predicted monomers

As seen in Table 1, the traditional physics-based docking tools (FRODOCK and HDOCK and PatchDock) perform very well on docking holo structures, showcasing their ability to accurately predict the rigid complex when the proteins already in bound conformation. However, these methods show a sharp decrease in performance when predicted AlphaFold monomers are used as input for docking, a more relevant use-case, falling behind DIFFMASIF. DIFFMASIF is able to generalise also to predicted structures better than other machine learning based methods.

3.3 DIFFMASIF learns relevant protein-protein complex characteristics

Binding site prediction Using the dMaSIF contact probabilities, the DIFFMASIF encoder selects the 512 most probable contacts for both the receptor and the ligand to use in further steps for pose prediction. This ranking is based on the binding site auxiliary loss and thus is trained to cover the interfaces of both proteins. From the 512 top ranked DIFFMASIF surface points of holo systems in the PINDER-xl test set, 53% lie directly at the correct interface, a significant improvement over randomly sampling 512 surface points, which would result in merely 12% of interface sites. Note that sampling 100% of the 512 points from the binding site is unlikely to be desirable as the model might want to leverage reference points not directly at the binding site to avoid clashing of two proteins at peripheral sites. Figure 4A shows how the DockQ score of the predicted pose improves with increasing binding site accuracy. In addition, systems with $> 50\%$ accurate binding site predictions (n=873) resulted in considerably higher DockQ CAPRI hit rate with 81.01% ("Acceptable" or better) using holo as monomers and 64.16% using predicted monomers, compared to the base performance of 58.83% and 46.11% respectively (Table 1). Thus, DIFFMASIF learns to recognize binding surfaces thanks to the cross-attention between learned surface embeddings. This two-step scheme also has the added benefit that known binding site information from either or both proteins can be utilized at inference time to further improve interface specificity.

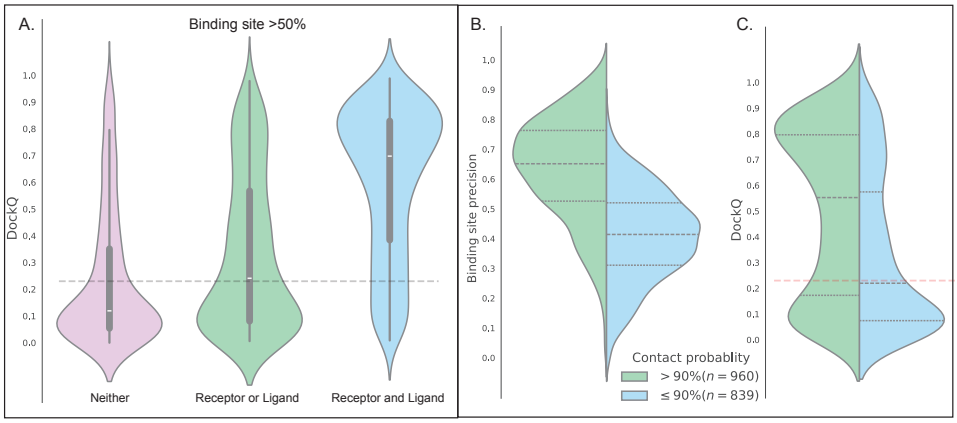

Figure 4: **A** DockQ distributions of complexes with accurate binding site prediction, i.e >50% of the predicted surface points lie in the interface, for neither ligand nor receptor, either one of ligand or receptor, and both. **B)** and **C)** Binding site precision and DockQ distributions for complexes with PRODIGY (Vangone & Bonvin) contact probability $> 90\%$ and $\leq 90\%$.

Performance on physiological interfaces. As deep learning methods are often seen to be biased due to the inherent biases in their training data, we wanted a complementary approach to verify that DIFFMASIF learns biologically relevant surface and structural complementarity signals. We demonstrate this in Figure 4 comparing the binding site precision (B.) and DockQ (C.) distributions of interfaces with high and low contact probabilities predicted by PRODIGY (Vangone & Bonvin). The differing distributions clearly favour dimers with higher probability physiological interfaces, and confirms that DIFFMASIF learns structural complementarity without overfitting, despite the unavoidable levels of noise present in protein complex experimental structure data used in training.

4 Conclusion

We present the first purely structure- and surface-based diffusion deep learning model for protein-protein docking. Our results demonstrate that ML models can achieve comparable results without the use of co-evolutionary information, and out-perform in situations where such information is scarce or not expected. In addition, DIFFMASIF is able to perform better than traditional physics based algorithms on rigid docking of predicted monomers, which is a more realistic scenario. This effort expands our toolbox for leveraging physico-electrochemical surface characteristics of proteins and lends well to future efforts where the right combination of co-evolution and structural complementarity can be learned across protein-protein space. In addition, we demonstrate the power of learning joint interface prediction and pose generation, also enabling the use of knowledge-based priors to improve prediction specificity. Overall, DIFFMASIF represents a step forward for protein representation learning especially in the context of generative modeling.

References

Mehmet Akdel, Alexander Goncearenco, and Yusuf Adeshina. Pinder: The protein interface dataset and resource, Nov 2023. URL `https://www.moml.mit.edu/`.

Sankar Basu and Björn Wallner. Dockq: a quality measure for protein-protein docking models. *PloS one*, 11(8):e0161879, 2016.

Gabriele Corso, Hannes Stärk, Bowen Jing, Regina Barzilay, and Tommi Jaakkola. Diffdock: Diffusion steps, twists, and turns for molecular docking. *arXiv preprint arXiv:2210.01776*, 2022.

Congyue Deng, Or Litany, Yueqi Duan, Adrien Poulenard, Andrea Tagliasacchi, and Leonidas J Guibas. Vector neurons: A general framework for so (3)-equivariant networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12200–12209, 2021.

Israel T Desta, Kathryn A Porter, Bing Xia, Dima Kozakov, and Sandor Vajda. Performance and its limits in rigid body protein-protein docking. *Structure*, 28(9):1071–1081, 2020.

Richard Evans, Michael O'Neill, Alexander Pritzel, Natasha Antropova, Andrew W Senior, Timothy Green, Augustin Žídek, Russell Bates, Sam Blackwell, Jason Yim, et al. Protein complex prediction with alphafold-multimer. *BioRxiv*, 2021.

Pablo Gainza, Freyr Sverrisson, Frederico Monti, Emanuele Rodola, D Boscaini, MM Bronstein, and BE Correia. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nature Methods*, 17(2):184–192, 2020.

Pablo Gainza, Sarah Wehrle, Alexandra Van Hall-Beauvais, Anthony Marchand, Andreas Scheck, Zander Harteveld, Stephen Buckley, Dongchun Ni, Shuguang Tan, Freyr Sverrisson, et al. De novo design of protein interactions with learned surface fingerprints. *Nature*, pp. 1–9, 2023.

Mario Geiger and Tess Smidt. e3nn: Euclidean neural networks. 2022. doi: 10.48550/ARXIV.2207.09453. URL `https://arxiv.org/abs/2207.09453`.

Ameya Harmalkar and Jeffrey J Gray. Advances to tackle backbone flexibility in protein docking. *Current opinion in structural biology*, 67:178–186, 2021.

Susan Jones and Janet M Thornton. Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences*, 93(1):13–20, 1996.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Ephraim Katchalski-Katzir, Isaac Shariv, Miriam Eisenstein, Asher A Friesem, Claude Aflalo, and Ilya A Vakser. Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proceedings of the National Academy of Sciences*, 89(6):2195–2199, 1992.

Mohamed Amine Ketata, Cedrik Laue, Ruslan Mammadov, Hannes Stärk, Menghua Wu, Gabriele Corso, Céline Marquet, Regina Barzilay, and Tommi S Jaakkola. Diffdock-pp: Rigid protein-protein docking with diffusion models. *arXiv preprint arXiv:2304.03889*, 2023.

Michael C Lawrence and Peter M Colman. Shape complementarity at protein/protein interfaces, 1993.

Nicholas A Marze, Shourya S Roy Burman, William Sheffler, and Jeffrey J Gray. Efficient flexible backbone protein–protein docking for challenging targets. *Bioinformatics*, 34(20):3461–3469, 2018.

Matt McPartlon and Jinbo Xu. Deep learning for flexible and site-specific protein docking and design.

Burcu Ozden, Andriy Kryshtafovych, and Ezgi Karaca. The impact of AI-based modeling on the accuracy of protein assembly prediction: Insights from CASP15. *BioRxiv*, 2023.

Martin Steinegger, Markus Meier, Milot Mirdita, Harald Vöhringer, Stephan J Haunsberger, and Johannes Söding. Hh-suite3 for fast remote homology detection and deep protein annotation. *BMC bioinformatics*, 20(1):1–15, 2019.

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*, pp. 15272–15281, 2021.

Raphael Townshend, Rishi Bedi, Patricia Suriana, and Ron Dror. End-to-end learning on 3d protein structure for interface prediction. *Advances in Neural Information Processing Systems*, 32, 2019.

Sandor Vajda and Dima Kozakov. Convergence and combination of methods in protein–protein docking. *Current opinion in structural biology*, 19(2):164–170, 2009.

Anna Vangone and Alexandre M. J. J. Bonvin. PRODIGY: A contact-based predictor of binding affinity in protein-protein complexes. 7(3):e2124. ISSN 2331-8325. doi: 10.21769/BioProtoc.2124. URL `https://doi.org/10.21769/BioProtoc.2124`. Publisher: Bio-protocol LLC.

Thom Vreven, Iain H Moal, Anna Vangone, Brian G Pierce, Panagiotis L Kastritis, Mieczyslaw Torchala, Raphael Chaleil, Brian Jiménez-García, Paul A Bates, Juan Fernandez-Recio, et al. Updates to the integrated protein–protein interaction benchmarks: docking benchmark version 5 and affinity benchmark version 2. *Journal of molecular biology*, 427(19):3031–3041, 2015.

Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (tog)*, 38(5): 1–12, 2019.