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

Abstract

Predicting protein-protein complexes is a central challenge of computational structural biology. However, 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 SOTA among Deep Learning methods for rigid body docking, in particular on structurally novel interfaces and low sequence conservation. This provides a significant advance towards accurate modelling of protein interactions with low co-evolution 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. Protein 3D structure 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 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

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

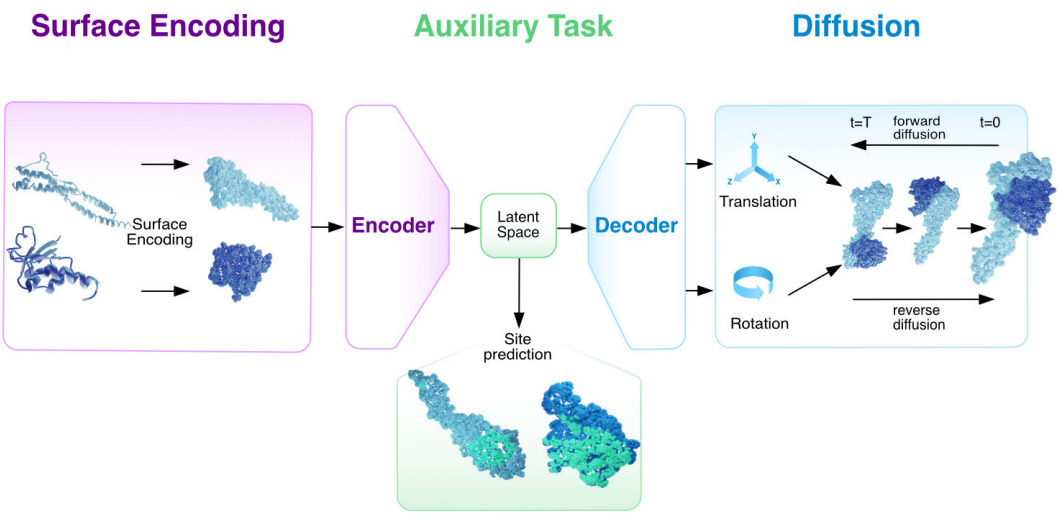

Figure 1: Overview of the DIFFMASIF method. Protein surface point clouds are generated and fed into a encoder-decoder network. The model learns both binding site prediction via an encoder, and denoising a reverse diffusion process over rotations and translations via a decoder.

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 such as UniRef (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 introduce a new dataset curation and splitting strategy intended for rigid-body docking by using many more protein complexes from the PDB and a structural interface clustering approach. This is consistent with our desire to reflect the performance of current methods by their ability to learn structural complementarity and generalize towards novel (potentially weakly co-evolved) interfaces. This resulted in the PINDER-xl test set 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. See appendix section **1.1** for more details on this data curation.

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 (vector neuron layers) (Wang et al., 2019; Deng et al., 2021) which takes as input coordinates and normal vectors 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 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.

3 Results

3.1 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. To evaluate the accuracy of binding site prediction, we evaluated the trained model on holo complexes from the PINDER-xl test set (appendix section **1.1**). We show that from the 512 top ranked DIFFMASIF surface points, 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 2A shows how the DockQ score of the predicted pose improves with increasing binding site accuracy. 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.

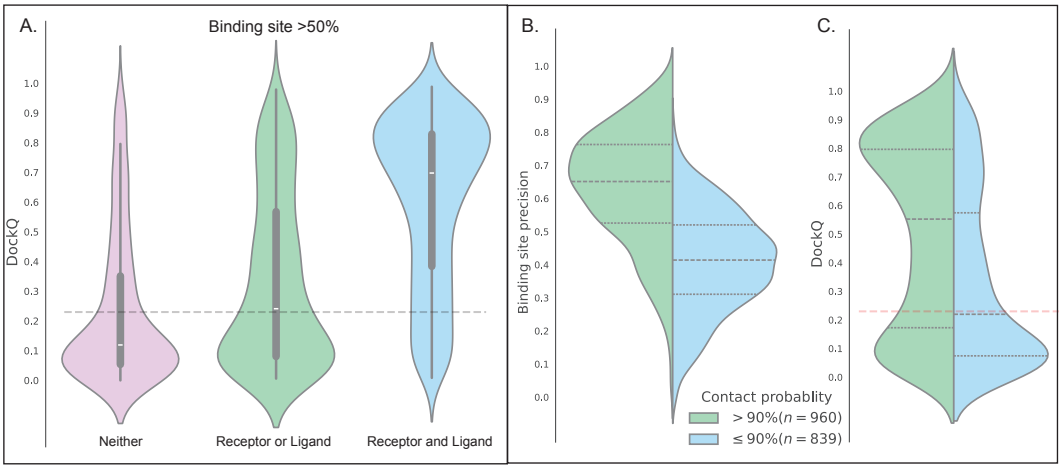

Figure 2: **A** DockQ distributions of complexes with accurate binding site prediction, i.e >50% of the predicted surface points lie in the interface, for neither ligand nor receptor, either one of ligand or receptor, and both. **B)** and **C)** Binding site precision and DockQ distributions for complexes with Prodigy contact probability > 90% and ≤90%.

Physiological interface prediction. 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 is Figure 2B and C by comparing the binding site precision and pose DockQ distributions of interfaces with high and low PRODIGY (Vangone & Bonvin) contact probabilities. The differing distributions clearly favour physiological interfaces and confirms that DIFFMASIF learns structural complementarity without overfitting, despite the unavoidable levels of noise present in protein complex experimental structure data.

3.2 DIFFMASIF outperforms co-evolution based docking methods

We benchmarked DIFFMASIF against popular physics-based and deep learning-based methods using the PINDER-xl test set appendix section **1.1** . 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 section **1.1**). Each pose generated by a particular docking method was superposed to the reference pose and evaluated based on three metrics: iRMSD (interface RMSD), LRMSD (ligand RMSD), and DockQ (a composite score that also encompasses Fnat – the fraction of native contacts). 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 (an *oracle* approach). These results are shown in 1.

Table 1: Comparison of Docking Methods

Method	**Holo**			**Predicted**		
	$> Acceptable$	$> Medium$	$High$	$> Acceptable$	$> Medium$	$High$
Frodock	96.8	95.78	91.32	38.42	31.97	11.74
Hdock	98.01	97.72	97.04	36.25	30.41	13.4
EquiDock	0.17	0	0	1.59	1.21	0.76
DiffDock-PP 40	53.5	31.83	10.76	-	-	-
DIFFMASIF 40	58.83	40.72	16.69	47.11	25.84	3.11
DIFFMASIF 40 correct binding (n=874)	81.01	61.56	26.89	64.16	40.52	4.56

Table 2: Complex prediction oracle metrics for traditional physics docking tools Frodock and Hdock, and machine learning tools EquiDock, DiffDock-PP and DIFFMASIF on the PINDER (PINDER-xl) benchmark set. Number next to method give number of poses oracle metrics are generated from for methods with 40 poses. We report the percentage of systems covered by DockQ hit categories as "$> Acceptable$", "$> Medium$" and "$High$" solutions.

Physics-based docking tools As seen in Table 1, the traditional physics-based docking tools (FroDock and HDOCK) perform very well on docking holo structures. They are able 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, falling behind DIFFMASIF which is able to generalise also to predicted structures, a more relevant use-case.

DiffDock-PP, We retrained DiffDock-PP on our novel data splits, after having ensured reproducibility of results previously reported using the original splits, and generated 40 poses per test complex with a reverse SDE using 40 steps. As seen in Table 1 and Figure 3A, DIFFMASIF shows consistently better scores than DiffDock-PP and returns more acceptable complexes (see Figure 3.2A for an example).

In addition, Figure 3B 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 as well as co-folding methods such as AF2MM include representations learned via masked language modelling objective on this information. From Figure 3A, it is clear that DiffDock-PP performance drops when the average N_{eff} goes below 5, while DIFFMASIF is not effected. Thus, DIFFMASIF is a highly complementary approach to rigid protein docking without reliance on co-evolutionary signals. Due to time constraints, we could not report the DiffDock-PP performance on predicted monomers, however the results will be submitted and uploaded to PINDER leaderboard in future.

Alphafold2-Multimer. 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 for the PINDER-af2 benchmark set of 72 complexes (appendix **Section 3**, Table **1**), with the exception of removing template structures dating September 2022 and onward. 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 (DockQ < 0.23) 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 3.2B for an example, and additional PINDER-af2 holo leaderboard in appendix **Section 4**, Table **2**.

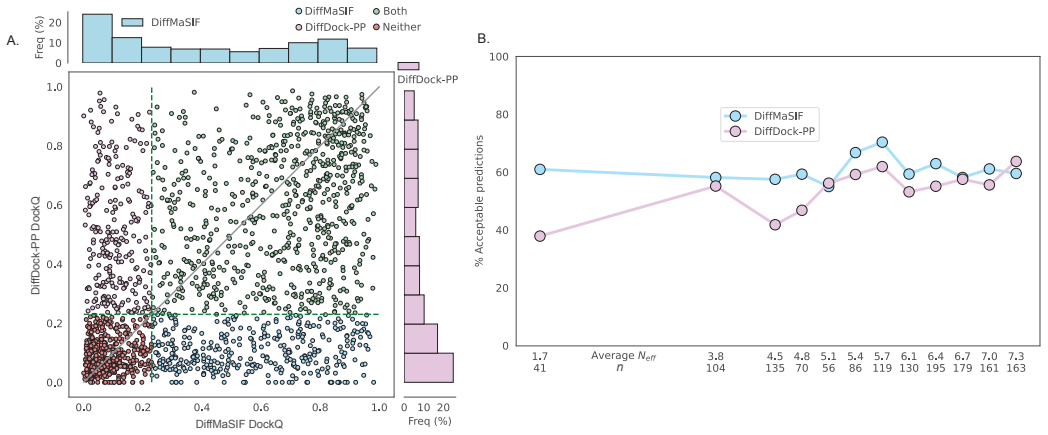

Figure 3: **A** DIFFMASIF and DiffDock-PP DockQ distributions. **B** Percentage of acceptable DiffDock-PP and DIFFMASIF complexes across different N_{eff}.

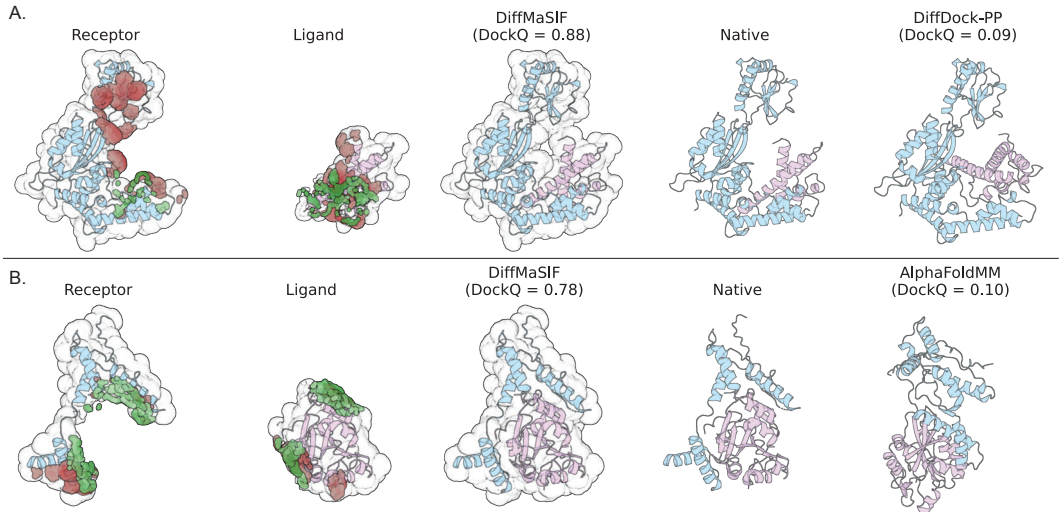

Figure 4: Docking of PDB ID: 6K3B (**A**) and 8FZZ (**B**). The two leftmost columns show DIFFMASIF's ability to correctly identify binding sites, with correctly identified interface points in green, and the rest in red. The third and fourth columns show the DIFFMASIF predicted docking pose and the ground truth pose, while the fifth column shows the DiffDock-PP predicted pose for **A** and the AF2MM predicted complex for **B**.

4 Conclusion

We present the first purely structure- and surface-based 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, our surface point graph and atom-to-surface pooling approach represent a step forward for protein representation learning especially in the context of generative modeling.

References

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.

1 Background

1.1 Deep learning on Protein Surfaces

MaSIF (Gainza et al. (2020)) is a geometric Riemannian convolutional architecture (Monti et al. (2017)) used to learn protein surface descriptors for predicting interaction properties. In its faster and end-to-end differentiable successor dMaSIF (Sverrisson et al. (2021)), points are generated in an iterative procedure, where for a sampled set of points $x_i \in X$ their distance to a level set of a signed distance function (SDF) is minimized, leading to surface points roughly equidistant from each surface atom. For each point, a local coordinate system $\hat{n}_i, \hat{u}_i, \hat{v}_i \in \mathbb{R}^3$ is constructed using the gradient of the distance function and subsequent calculation of tangent vectors. This local coordinate system is then used to define quasi-geodesic convolutions along the surface, which are used to propagate a 16-dimensional feature vector for each point. The 16 input features of each point are comprised of 10 geometric descriptors (mean- and Gaussian curvatures at 5 different scales) and a 6-dimensional vector stemming from its 16 nearest atoms. Each atom is one-hot encoded by their respective atom types $[C, H, O, N, S, Se]$ and concatenated to the atom-point distance, creating an embedding in $\mathbb{R}^7$ which is pooled to the closest point via an MLP. With DIFFMASIF we further build on these representations.

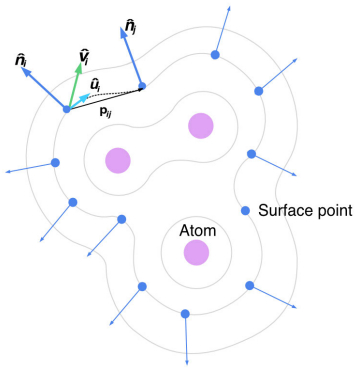

Figure 1: dMaSIF surface point cloud featurization. Level sets (gray) around atoms (pink) are used to find surface points (blue). Each point is equipped with a coordinate system comprised of normals $\hat{n}_i$ (blue) and orthogonal tangent vectors $\hat{v}_i$, $\hat{u}_i$.

1.2 Contact Prediction as Implicit Confidence Models

In this section, we show that we can rank generated poses using the proximity of contact prediction scores to the interface. To do this, we round the surface contact probabilities to binary labels. Next, we take the sum of these binary values at the interface of the generated pose. We define the generated interface as receptor and ligand surface nodes within 3 angstroms. We refer to this metric as the EvoScore.

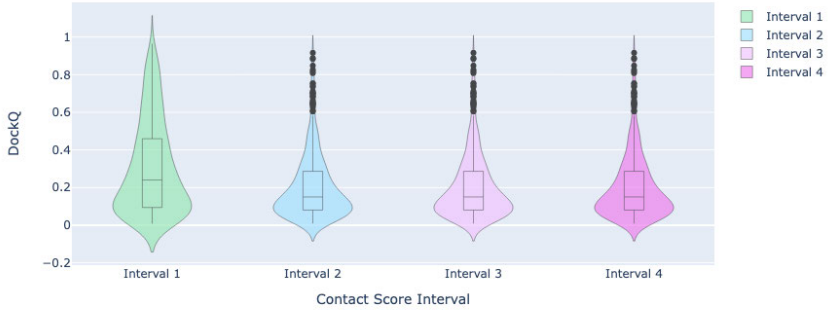

Figure 2: Interval 1 is oracle metrics of top 10 highest EvoScore poses, Interval 2 is oracle metrics of 11th-20th, highest scoring poses, etc. We see that this simple metric is able to rank high quality and low quality bins highlighting that the score model and contact ranking model learn complementary features.

1.3 Score-based Diffusion Models

Score-based diffusion models integrate techniques from both score-based generative models and diffusion models into a unified framework (Song & Ermon). In score-based generative modeling, the

score function $\mathbf{s}_\theta(\mathbf{x}) \approx \nabla_\mathbf{x} \log p(\mathbf{x})$ represents gradients of the data log-density $p(\mathbf{x})$. The score can be estimated via denoising score matching on noise-corrupted samples, without needing to compute intractable normalizing constants. Langevin dynamics can then sample from the estimated score model. Diffusion models perturb data $\mathbf{x}_0$ through Markov chains of added Gaussian noise to obtain $\mathbf{x}_t$ at noise level t. The forward diffusion process can be represented as a stochastic differential equation (SDE):

$$\mathrm{d}\mathbf{x} = f(t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w} \tag{1}$$

where $f(t)$ and $g(t)$ represent drift and diffusion coefficients respectively, and $\mathrm{d}\mathbf{w}$ is Gaussian noise. The reverse process is modeled by learning an approximate conditional distribution $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$. Score-based diffusion models leverage score functions to parameterize the generative diffusion process. The forward SDE incrementally adds noise to the data distribution $p_0(\mathbf{x})$. Critically, the reverse-time SDE is:

$$\mathrm{d}\mathbf{x} = [f(t) - g(t)^2 \nabla_\mathbf{x} \log p_t(\mathbf{x})]\mathrm{d}t + g(t)\mathrm{d}\mathbf{w} \tag{2}$$

The score functions $\nabla_\mathbf{x} \log p_t(\mathbf{x})$ can be estimated by a time-dependent score-based model $\mathbf{s}_\theta(\mathbf{x}, t)$ trained via score matching. This results in an estimated reverse SDE that can be numerically solved to sample from $p_0(\mathbf{x})$. Alternatively, the estimated reverse SDE can be converted to a probability flow ODE, enabling exact likelihood computation (Song et al.).

2 DIFFMASIF Methods

2.1 Diffusion Process

For the pose generation component of our model, our study adopts a diffusion process approach akin to the methodologies presented in Corso et al. (2022); Ketata et al. (2023). The key observation of DiffDock was that instead of modelling the high-dimensional distribution of protein and ligand coordinates, the problem of rigid-body docking can instead be formulated as learning a conditional distribution of centroid translations and rotations of the ligand given the receptor coordinates as conditioning variables or $P(R_L, X_L|r^*)$ where r^* is encodes the receptor. This requires formulation of the diffusion process over the combined manifold formed by the group of rigid-body rotations $(SO(3))$ and translations denoted as $SE(3)$. Since $SE(3)$ is a lie group, it is by definition a Riemannian manifold. Therefore methods for diffusion on Riemannian manifolds can be lifted to this setting.

Our model arbitrarily chooses one of the protein partners as the ligand. The receptor coordinates are kept fix and serves only as a conditioning variable for the ligands modeled score.

Translations: For generating the translational degrees of freedom of our ligand we utilize the variance exploding SDE:

$$d\mathbf{x} = \sqrt{d\sigma_{tr}^2(t)/dt}\, d\mathbf{w} \tag{3}$$

Where $d\mathbf{x}$ signifies the change in position, $\sigma_{tr}^2(t)$ is the a variance of the Brownian motion at a specific time t.

Rotations: For the rotational degrees of freedom, a two step process is used:

We initially select a random axis, represented as $\hat{\omega}$, and a random angle ω constrained between 0 and π to avoid degeneracy. The likelihood of opting for a particular angle of rotation ω about $\hat{\omega}$ is expressed by:

$$p(\omega) = \frac{1 - \cos\omega}{\pi} f_\epsilon(\omega) \tag{4}$$

Where $f_\epsilon(\omega)$ is a truncated series expression:

$$f_\epsilon(\omega) = \sum_{l=0}^{\infty} (2l+1) \exp\left(-l(l+1)\epsilon^2\right) \frac{\sin\left((l+\frac{1}{2})\omega\right)}{\sin(\frac{\omega}{2})} \tag{5}$$

Similar to translations, we utilize a variance exploding SDE to model the change in rotations via:

$$d\mathbf{r} = \sqrt{d\epsilon(t)/dt}\, d\mathbf{w} \tag{6}$$

2.2 Model Architecture

The architecture of DIFFMASIF has two main components: an encoder and a decoder. The encoder and decoder were designed to aggregate surface level descriptions of our systems in an equivariant fashion. While introducing surfaces features explicitly encodes geometric complementarity, it also greatly increases the memory overhead. The introduction of the joint binding site training task was an easy way to limit the cardinality of surface points while also incorporating a useful biological prior into the diffusion process. By utilizing the predicted binding site as a mask on surface nodes, we can greatly reduce the number of points required for the costly operations of tensor-product convolutions in our decoder network. Additionally, the task of pose prediction allows for useful additional signal to be passed to the binding prediction module. Finally, by converting our model to a more local problem, we can use a shallower number of tensor-product layers as the problem requires a smaller receptive field. This allowed us to drop to only 2 layers of tensor-product convolution enabling a mean inference time for 40 poses of 12.3 seconds when running inference on one H100.

Representation, Encoder and Binding-Site Auxiliary Task DIFFMASIF makes use of surface point cloud based representations. Point cloud coordinate and normal vectors are generated by dMaSIF method, in addition to the one-hot encoding of atom-types. These surface point based features are then fed into dMaSIF layers separately for ligand and receptor proteins.

In DockGPT (McPartlon & Xu), it was shown that including only one contact point during generation greatly improves complex prediction. This led us to hypothesize that, while related, binding-site prediction and pose generation are fundamentally distinct tasks. Therefore we sought to construct a loss that differentiates interaction site prediction from pose prediction. We define true binding site nodes are defined as surface nodes < 3Å from the other molecule's surface points in the ground truth pose. To predict the binding sites on protein and receptor, dMaSIF features are first computed for each surface independently. For the purposes of docking, we want our predicted interface to be localized to the site that the given partner will bind to rather than any possible interaction location. Therefore we needed a way to update the dMasif features based on the features of the binding partner in a way that is invariant to translations and translations of the ligand chain. **?** introduced one possible solution for invariant conditioning on a partner by updating surface features via alternating steps of self- and cross-attention on the paired surface features. Since both the ligand and receptor surface embeddings are invariant to rigid rotations and translations, cross attention will also be invariant.

Decoder. After the binding-site auxiliary task, we take the top 512 predicted binding surface nodes of both the ligand and the receptor. These 512 ligand and receptor surface nodes are then combined into a joint surface graph using a radius graph. Similar to DiffDock, the radius of this graph is dynamically set to be a multiple of the current translation σ, allowing for sparser graphs at lower perturbations while still enabling connections at large perturbations. The first component of this encoder after the masking is DGCN (vector neuron layers) (Wang et al. (2019); Deng et al. (2021)) which takes as input coordinates and normal vectors and outputs higher-dimensional vector embeddings. These vector features are then provided, along with surface coordinates, and surface scalar features to an E(3)-equivariant graph convolution layer constructed using the e3nn library (Geiger & Smidt (2022)). The use of simple vector neurons allows us to obtain equivariant features from a large receptive field without requiring the cost of tensor product calculations performed by e3nn. The final output of the decoder is the prediction of the translation and rotation required for the ligand coordinates. The combination of DGCNN and E(3)-equivariant graph convolutions ensure the process respects the geometric constraints of the protein surface structure and enhances the model's ability to predict complex protein interactions.

Below we define a single training step for DIFFMASIF. The capital letter denotes the chain (L for ligand R for receptor). The variable x denotes coordinates, f denotes raw features, and $\hat{n}$ denotes normals.

3 Dataset details

We leverage PINDER (**?**), summarized in table 1, for training, validation and testing. PINDER, short for (**P**rotein **IN**terface **D**atas**E**t and **R**esource), is a dataset and resource for training and evaluation of

Algorithm 1 DIFFMASIF Train Step

Input : $x_L, \hat{n}_L, f_L, x_R, \hat{n}_R, f_R$
Pretransform:
Center all coordinates about receptor surface x_R
Sample and apply random rotation to both chains to ensure equivariant learning
Compute ground truth contacts c_L, c_R
Sample rotation and translation noise x' and R'
Noise coordinates and normals to obtain $x'_L, \hat{n}'_L$
Predict Binding Site:
Compute dMasif features $h_L = \Omega(x'_L, \hat{n}'_L, f_{Ls})$ and $h_R = \Omega(x_R, \hat{n}_{Rs}, f_R)$
Perform cross attention by $A = \text{softmax}\left(\frac{f_R \cdot h_L^T}{\sqrt{d_k}}\right)$ and take $h'_L = A \cdot h_L$. Repeat for h_R
Feed h'_L and h'_R through linear layers to obtain surface logits c'_L, c'_R
Take argkmax(c'_L) and argkmax(c'_L) to obtain predicted interface indices
Mask according to indices to obtain $h'_{L'}, x'_{L'}, \hat{n}'_{L'}$ and $h_{R'}, x_{R'}, \hat{n}_{R'}$
Compute Score:
Predicted scores using masked coordinates $\omega, \hat{x} = \Phi(h'_{L'}, x'_{L'}, \hat{n}'_{L'}, h_{R'}, x_{R'}, \hat{n}_{R'})$
Compute Loss:
Compute contact loss C $= \frac{1}{2}(\text{BCELoss}(\sigma(c'_L), c_L) + \text{BCELoss}(\sigma(c'_R), c_R))$
Compute score loss $\mathcal{L} = ||\omega - \nabla \log p_t^{rot}(\Delta R|0)||^2 + ||\hat{x} - \nabla \log p_t^{tr}(\Delta x|0)||^2$
Compute full loss $C + \mathcal{L}$

protein-protein docking algorithms. We summarize the details briefly below but refer to the pre-print for a full description.

PINDER provides redundancy- and quality filtered validation and test sets with paired, pre-rotated and translated, unbound and AlphaFold2-predicted monomers next to bound monomers, allowing to assess performance on both rigid and flexible systems. It is also the only dataset that provides a large training dataset comprised of redundant and unfiltered bound complexes and a large amount of paired unbound (apo) and predicted (AlphaFold2) monomers to train on the flexible docking task specifically and select different training data mixes from. In this work, all reported machine learning methods were trained only on bound complexes sampled randomly from each interface cluster, while metrics are provided for bound (holo), and predicted. We leave training on unbound and predicted structures and selection of different data mixes via quality filters as highly interesting directions for future work. PINDER also provides results from other State-of-the-Art Deep Learning methods that are retrained on holo structures of PINDER to allow for fair comparison.

PINDER provides 3 benchmark sets:

- PINDER-xl: Comprised of 4297 structures (1756 holo structures, with 868 paired apo and 1673 paired predicted structures). This constitutes high quality representatives of 10% of the interface clusters.
- PINDER-s: A subset of PINDER-xl with complete apo coverage, diversity and highest resolution. It comprises 351 structures (117 holo, 117 apo, 117 predicted).
- PINDER-af2: Comprised of test clusters where no cluster member was released before AlphaFold-Multimer 2.3 (Evans et al. (2021)) training cutoff. This constitutes an Alphafold2-Multimer de-leaked benchmark set, albeit smaller in size with a total of 150 structures (72 holo, 21 apo, 57 predicted).

In our benchmarks, we made use of PINDER-xl holo and predicted, and PINDER-af2 holo datasets.

While we only train on holo structures, we also perform docking with predicted chains to assess DIFFMASIF against flexible alternatives like AlphaFold2 Multimer.

PINDER was created as follows:

The PINDER database was constructed using the RCSB NextGen database (as of 01.09.2023) as the foundation. The mmCIF files were obtained and representative biological assemblies were generated.

Set	Size	Holo	Apo	Predicted	Clusters	Easy/Medium/Hard
PINDER-xl	4297	1756	868	1673	1124	57/18/25
PINDER-s	351	117	117	117	100	63/11/26
PINDER-af2	150	72	21	57	72	55/36/9
validation	3751	1777	317	1657	1098	N/A
train	506617	236128	71970	198519	10752	N/A

Table 1: Dataset counts, including size, number of pairs, clusters, difficulty level and purpose. Difficulty is assigned via DB4 criteria. Train and validation set are not assigned canonical apo structures and have no single difficulty label

Protein-Protein Interactions (PPIs) were identified as all pairs of chains with a heavy atom in contact at a 5Å threshold. For apo structures, up to 10 monomeric PDB entries with the same UniProt ID as PPI entries were aligned using the UniProt numbering to the corresponding PPI entry. The entry with the highest alignment overlap was selected, and entries that did not align with more than 5 residues were discarded. For AlphaFold2 structures, AFDB entries with the same UniProt ID as PPI entries were aligned using UniProt numbering to the corresponding PPI entry.

All-vs-all structural alignments of complete holo protein chains were performed with FoldSeek utilizing both sequence and structure similarity. Interfaces were delineated by selecting residues within 5Å Cα distance between aligned chains. A graph was constructed with protein chains as nodes and edges between chains with over 75% alignment coverage of interface residues, thereby connecting chains with similar interfaces. Community clustering via asynchronous label propagation partitioned the chain graph into clusters. During the later training/validation/test assignment, to ensure no leakage between train, validation and test splits derived from these clusters, further filtering during test/validation set selection excluded protein-protein interaction pairs with more than 30% sequence similarity as a second layer of checks. This did not remove any further structures confirming the clustering is strict. The final paired-interface clusters represent each protein-protein interaction by the pair of cluster identifiers $\{c_a, c_b\}$ for the two constituent chains.

The total dataset comprised 281,756 holo systems, of which 87,665 were matched with apo systems and 232,557 were matched AlphaFold2 predicitons. The coverage of apo and AlphaFold2 clusters in the total dataset was 46% and 84% respectively, with a total of 601,978 systems.

Annotations were obtained from the RCSB NextGen database, including: (1) oligomeric state of the protein complex such as homodimer, heterodimer, oligomer or higher-order complexes; (2) structure determination method such as X-Ray diffraction, cryo-electron microscopy (CryoEM), or nuclear magnetic resonance (NMR) spectroscopy; (3) resolution of the structure; (4) interfacial gaps defined as structurally-unresolved segments on protein-protein interaction (PPI) interfaces; (5) number of distinct atom types where earlier CryoEM structures often contain only a few atom types such as Cα or backbone atoms; (6) annotation by PRODIGY-cryst (Vangone & Bonvin) on whether an interface is likely physiological or a crystal contact; (7) structural elongation defined by the maximum variance of coordinates projected onto the largest principal component to detect long end-to-end stacked complexes with small repetitive interfaces; (8) planarity defined by the deviation of interfacial Cα atoms from the fitted plane to quantify interfacial shape complementarity where transient complexes have smaller, more planar interfaces than permanent and structural scaffold complexes; (9) number of connected components in a 10Å Cα radius graph to detect structurally discontinuous domains; and (10) intermolecular contacts labeled as polar or apolar.

Selection of test and validation set was done by filtering via these annotations. Filtering criteria included physiological contact as determined by PRODIGY-cryst Vangone & Bonvin, selection of only dimers, a resolution of less than 4Å, the presence of more than four atom types, an elongation score of less than 0.95, and the requirement of a single component. Clusters that contained at least one member meeting these criteria were retained to sample validation and test clusters from. This filtering process yielded 45,628 protein-protein interactions (PPIs), distributed across 1098 and 1124 clusters for train and test, respectively. PINDER-XL/S are subsets of these test sets. PPIs within these sampled clusters that met the criteria were ranked by resolution and cluster size, starting from median sized clusters and extending in both directions until 10% of clusters are sampled. Two PPIs

were randomly sampled from each cluster, with the exception of singleton clusters, from which only one PPI was sampled. The representative samples were divided equally between clusters where the top-ranking member was a homodimer and those where the top-ranking member was a heterodimer.

Test clusters which contain only members released after 01.01.2022 (the AF-MM 2.3 training cutoff) were separated into 150 clusters. From these, 1 dimer per cluster was selected as the PINDER-AF2 dataset.

The remaining 236'128 systems from 10'752 clusters were used to sample from during training. The cluster coverage of apo and AlphaFold2 in the training set was 45% and 84% respectively, with the total size of the training set being 506'617. However, we only sampled from apo, as previously mentioned.

We assign target difficulity similar to DB5 Vreven et al. (2015) via the following criteria:

- The difficulty is "Rigid-body" if the irmsd is less than or equal to 1.5 and the fnonnat is less than or equal to 0.4.
- The difficulty is "Medium" if either of the following conditions is true: (a) The irmsd is greater than 1.5 and less than or equal to 2.2. (b) The irmsd is less than or equal to 1.5 and the fnonnat is greater than 0.4.
- The difficulty is "Difficult" if none of the above conditions are true.

To estimate the level of co-evolutionary signals present in the test-set, PINDER ran the multiple sequence alignment (MSA) creation step of ColabFold (Mirdita et al.) which generates an MSA of the MMSeqs2 hits from UniRef100, PDB70 and an environmental sequence set. These MSAs were used to calculate the average number of effective sequences (N_{eff}) at each position, a measure of availability of homologous sequences that can be utilized by evolution-based approaches.

4 Additional results

Table 2: Comparison of Docking Methods for pinder-af2 holo benchmark set

Method	**Holo**		
	>Acceptable	**>Medium**	**High**
Frodock	90.28	87.5	86.11
Hdock	94.44	93.06	93.06
EquiDock	0	0	0
DiffDock-PP 40	52.78	37.5	12.5
DIFFMASIF 40	55.56	37.5	16.67

Table 3: Complex prediction oracle metrics for traditional physics docking tools Frodock and Hdock, and machine learning tools EquiDock, DiffDock-PP and DIFFMASIF on the PINDER-af2 benchmark. The number next to the method gives the number of poses oracle metrics are generated from for methods with 40 poses. We report the percentage of systems covered by DockQ hit categories as ">Acceptable", ">$Medium$" and "$High$" solutions.

References

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.

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.

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

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.

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

Milot Mirdita, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger. ColabFold: making protein folding accessible to all. 19(6):679–682. ISSN 1548-7105. doi: 10.1038/s41592-022-01488-1. URL https://doi.org/10.1038/s41592-022-01488-1.

Federico Monti, Davide Boscaini, Jonathan Masci, Emanuele Rodola, Jan Svoboda, and Michael M Bronstein. Geometric deep learning on graphs and manifolds using mixture model cnns. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5115–5124, 2017.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. (arXiv:1907.05600). URL http://arxiv.org/abs/1907.05600.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. (arXiv:2011.13456). URL http://arxiv.org/abs/2011.13456.

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.

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.