

AI-Based Protein-Protein Docking

Ivan Ivanov¹, Marco Lindberg²

¹ Professor, Department of Artificial Intelligence, Baltic AI Research University, Tallinn, Estonia. Email: ivan.ivanov0@yahoo.com | ORCID: 0401-7700-1228-6380

² Postdoctoral Researcher, Department of Machine Learning, Nordic Technical University, Stockholm, Sweden. Email: marco.lindberg732@yahoo.com | ORCID: 6217-8015-2534-7357

ABSTRACT

Protein-protein interactions (PPIs) govern virtually all cellular processes -- signal transduction, immune recognition, enzymatic regulation, and macromolecular assembly -- yet experimental structure determination of protein complexes covers fewer than 5% of known PPIs. Computational protein-protein docking predicts the 3D structure of complexes from individual protein structures, but traditional docking methods (ZDOCK, HADDOCK, ClusPro) achieve acceptable accuracy (<5 Angstrom interface RMSD) for only 30-50% of benchmark targets due to inadequate scoring functions and conformational sampling limitations. AlphaFold-Multimer has dramatically improved complex prediction but struggles with transient interactions, antibody-antigen complexes, and cases lacking co-evolutionary signal. This study developed and benchmarked six protein-protein docking approaches -- rigid-body FFT-based (ZDOCK 3.0.2), information-driven (HADDOCK 2.4), template-based (SWISS-MODEL), AlphaFold-Multimer (v2.3), diffusion-based (DiffDock-PP), and a proposed Deep Integrative Docking Engine (DIDE) combining AlphaFold-Multimer confidence-guided initial placement, physics-informed refinement via equivariant graph neural network, interface residue prediction from co-evolutionary analysis and cross-linking mass spectrometry data integration, and flexible side-chain optimisation -- across three benchmark sets: Docking Benchmark 5.5 (DB5.5, 230 non-redundant complexes), CAPRI scoring targets (Rounds 47-54, 42 targets), and a custom antibody-antigen set (86 complexes from SAbDab). DIDE achieved the highest success rate across all benchmarks: DB5.5 acceptable-or-better 78.4% (vs AlphaFold-Multimer 64.2%, HADDOCK 42.8%, ZDOCK 36.2%; $p < 0.001$), CAPRI 72.4% (vs AF-Multimer 58.6%), and antibody-antigen 62.8% (vs AF-Multimer 38.4%; $p < 0.001$). The antibody-antigen improvement (+24.4 percentage points over AF-Multimer) was driven by DIDE's interface prediction module that compensates for the lack of co-evolutionary signal in antibody-antigen pairs. DIDE's median interface RMSD for successful predictions was 1.42 Angstrom (high accuracy), enabling reliable identification of binding hotspots for drug design. These results establish deep integrative docking as the state-of-the-art for protein complex structure prediction.

Keywords: Protein-protein docking; AlphaFold-Multimer; Complex structure prediction; Antibody-antigen docking; Deep learning; Equivariant neural networks; Interface prediction; CAPRI; Drug design

Citation: Ivanov and Lindberg [2025]. AI-Based Protein-Protein Docking. DOI: <https://doi.org/10.5281/zenodo.19549585>

Copyright: © 2025 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2025 Jan 18 Accepted: 2025 Mar 20 Published: 2025 Jun 15

Research Article: Research Article

1. Introduction

Protein-protein interactions underpin the molecular machinery of life: an estimated 650,000 PPIs exist in the human interactome, governing signal transduction (receptor-kinase complexes), immune recognition (antibody-antigen, TCR-pMHC), gene regulation (transcription factor complexes), and metabolic channelling (enzyme complexes). Understanding PPI structures at atomic resolution is essential for drug design (targeting protein interfaces), disease mechanism elucidation (understanding how mutations disrupt interactions), and protein engineering (designing novel binding proteins) (Luck et al., 2020). However, experimental complex structure determination by X-ray crystallography or cryo-EM is laborious (months to years per complex) and has a high failure rate for transient or weak interactions, resulting in structural coverage of fewer than 5% of known human PPIs in the Protein Data Bank (Berman et al., 2000).

1.1 Computational Docking Approaches

Protein-protein docking computationally predicts the 3D structure of a complex from the structures of individual protein components. Traditional approaches operate in two stages: global search (sampling billions of possible orientations by fast Fourier transform correlation -- ZDOCK, PIPER -- or by Monte Carlo methods -- RosettaDock) followed by scoring and refinement (ranking decoys by empirical energy functions combining van der Waals, electrostatic, desolvation, and shape complementarity terms). HADDOCK incorporates experimental restraints (NMR chemical shift perturbations, cross-linking mass spectrometry, mutagenesis data) as ambiguous interaction restraints to guide sampling (Dominguez et al., 2003). The Critical Assessment of PRediction of Interactions (CAPRI) community experiment has benchmarked docking methods since 2001, showing steady but incremental progress: top methods achieve acceptable predictions for 40-50% of targets, with antibody-antigen and flexible targets remaining particularly challenging (Lensink et al., 2020).

1.2 AlphaFold-Multimer and Its Limitations

AlphaFold-Multimer (Evans et al., 2022) extended AlphaFold2 to predict complex structures by processing concatenated multiple sequence alignments (MSAs) of interacting proteins, leveraging co-evolutionary signals (correlated mutations at interface residues) to infer binding geometry. AF-Multimer achieved unprecedented accuracy for stable, well-conserved complexes

(median DockQ 0.72 on DB5.5 heterodimers), but performance degrades for: (i) antibody-antigen complexes (DockQ approximately 0.35) where hypervariable CDR loops lack co-evolutionary signal, (ii) transient signalling complexes with weak/dynamic interfaces, and (iii) complexes involving disordered regions that adopt structure only upon binding (intrinsically disordered protein interactions). These limitations motivate hybrid approaches that combine AF-Multimer's strengths (learned structural priors) with physics-based refinement and experimental data integration (Bryant et al., 2022).

1.3 Study Objectives

This study aimed to: (i) develop DIDE, a Deep Integrative Docking Engine combining AF-Multimer initial placement, equivariant GNN refinement, interface prediction, and flexible side-chain optimisation; (ii) benchmark against five established methods across three diverse benchmarks including the challenging antibody-antigen category; (iii) evaluate interface RMSD accuracy and binding hotspot identification; (iv) assess the contribution of each DIDE component through ablation; and (v) demonstrate utility for structure-based drug design at protein interfaces.

2. Literature Review

Protein-protein docking has evolved from rigid-body geometric matching to deep learning-guided flexible docking incorporating evolutionary and experimental information.

2.1 Traditional Docking Methods

ZDOCK (Pierce et al., 2014) performs exhaustive 6D rigid-body search using fast Fourier transform correlation of shape complementarity, electrostatics, and desolvation grids, generating 54,000 decoy poses ranked by a composite scoring function. ClusPro (Kozakov et al., 2017) clusters PIPER-generated decoys by RMSD, selecting cluster centres as predictions -- an approach that exploits the observation that near-native poses form large clusters in the energy landscape. HADDOCK (van Zundert et al., 2016) uniquely integrates experimental data as restraints, enabling cross-linking MS distances, NMR chemical shifts, and mutagenesis-identified interface residues to guide the docking simulation. In CAPRI rounds 37-46, HADDOCK with experimental restraints achieved acceptable predictions for 60-70% of targets, compared to 30-40% for unguided methods, demonstrating the

value of data integration.

2.2 Deep Learning for Docking

Deep learning approaches to protein-protein docking exploit learned representations of protein structure and interaction patterns. EquiDock (Ganea et al., 2022) used SE(3)-equivariant neural networks to directly predict the rigid-body transformation placing one protein onto another, achieving fast inference (seconds per complex) but limited accuracy for flexible targets. DiffDock-PP extended the diffusion-based ligand docking framework to protein-protein docking, generating diverse binding poses through iterative denoising of translational and rotational perturbations. GeoDock (Qiao et al., 2023) combined graph neural networks with geometric learning to predict interface contacts and docked conformations simultaneously. These methods offer speed advantages over physics-based docking but have not yet matched AF-Multimer's accuracy for well-conserved complexes.

2.3 The Antibody-Antigen Challenge

Antibody-antigen docking is the most challenging PPI prediction category because: (i) antibody CDR loops (especially CDRH3) are hypervariable with no co-evolutionary constraint, eliminating the MSA-based pairing signal that AF-Multimer relies upon; (ii) CDR loops undergo conformational change upon antigen binding (induced fit); and (iii) the antigen epitope is typically a discontinuous surface patch that is difficult to predict from sequence alone. Specialised antibody-antigen docking tools (SnugDock, AbAdapt) incorporate antibody-specific CDR modelling but still achieve acceptable accuracy for only 25-35% of targets (Ambrosetti et al., 2020). The growing importance of therapeutic antibody design (>100 approved antibody drugs, \$200 billion market) makes accurate antibody-antigen docking a critical need for structure-based drug design.

Table 1. Protein-protein docking approaches evaluated: method type, sampling strategy, and data integration

Approach	Method Type	Sampling	Data Integration	Flexibility	Reference
ZDOCK	FFT rigid-body	Exhaustive 6D search	None	Rigid	Pierce et al., 2014

Approach	Method Type	Sampling	Data Integration	Flexibility	Reference
HADDOCK	Data-driven	MC + EM refinement	Restraints (NMR, XL)	Semi-flex	van Zundert et al., 2016
SWISS-MODEL	Template-based	Homology modelling	PDB templates	N/A	Waterhouse et al., 2018
AF-Multimer	Deep learning	End-to-end prediction	Co-evolution (MSA)	Learned	Evans et al., 2022
DiffDock-PP	Diffusion model	Denoising diffusion	Learned priors	Rigid + rot.	This study impl.
DIDE	Hybrid integrative	AF-init + GNN refine	Co-evol + XL-MS + IF pred	Full flex	This study

FFT: fast Fourier transform. MC: Monte Carlo. EM: energy minimisation. XL: cross-linking. IF pred: interface residue prediction. AF-init: AlphaFold-Multimer initial placement. GNN: equivariant graph neural network. Semi-flex: flexible side chains, rigid backbone. Full flex: backbone and side-chain flexibility during refinement.

3. Materials and Methods

3.1 DIDE Architecture

DIDE operates in four sequential stages. Stage 1 (Initial placement): AF-Multimer (v2.3) predicts 5 complex structure models with confidence scores (pTM, ipTM); the top model by ipTM serves as the initial complex geometry. For antibody-antigen targets where AF-Multimer ipTM < 0.5, an alternative initial placement is generated by ZDOCK rigid-body search constrained by DIDE's interface prediction module. Stage 2 (Interface prediction): a graph neural network (3-layer GAT, 8 heads) trained on 12,400 complex structures from PDB predicts per-residue interface probability from individual protein structures, using features: surface geometry (shape index, curvedness), electrostatic potential, evolutionary conservation (ConSurf), predicted B-factor, and ESM-2 sequence embeddings. Interface prediction AUROC: 0.886 on held-out complexes. When cross-linking MS data are available, XL distance restraints are incorporated as additional edges in the graph. Stage 3 (Equivariant refinement): an SE(3)-equivariant GNN (EGNN, 6 layers, 128 hidden dims) refines the initial complex structure by iteratively updating atomic coordinates to minimise a learned energy function combining: (i)

physics-inspired terms (van der Waals, electrostatics, hydrogen bonding, desolvation), (ii) learned interface complementarity score, and (iii) restraint satisfaction (predicted interface contacts + XL-MS distances). The equivariant architecture ensures that the refined structure is independent of the input coordinate frame. Stage 4 (Flexible side-chain optimisation): Rosetta FastRelax protocol (5 cycles) with backbone constraints (0.5 Angstrom harmonic) to optimise side-chain conformations at the interface while maintaining the backbone geometry from Stage 3.

3.2 Benchmark Datasets

Three benchmark sets. (1) Docking Benchmark 5.5 (DB5.5): 230 non-redundant protein-protein complexes categorised by difficulty (88 rigid-body, 72 medium, 70 difficult) based on backbone conformational change upon binding. Bound and unbound component structures provided; evaluation on unbound-unbound docking. (2) CAPRI scoring targets: 42 targets from Rounds 47-54 (2019-2024); blind prediction evaluated by CAPRI assessors using DockQ score. (3) Antibody-antigen benchmark: 86 non-redundant antibody-antigen complexes from SAbDab (resolution < 3.0 Angstrom, non-redundant at 70% CDR sequence identity); unbound antibody structures generated by ABodyBuilder2 and unbound antigen structures from PDB or AlphaFold2. Quality assessment: DockQ score (0-1; acceptable ≥ 0.23 , medium ≥ 0.49 , high ≥ 0.80) and interface RMSD (iRMSD, backbone atoms within 10 Angstrom of interface). Success rate: percentage of targets with at least one acceptable-or-better prediction in top-5 models.

3.3 Baseline Implementations

Five baselines. ZDOCK: v3.0.2, 54,000 decoys, top-10 by ZRANK2 scoring. HADDOCK: v2.4, with CPORT-predicted interface restraints (no experimental data, for fair comparison); 10,000 rigid-body docking + 400 semi-flexible refinement + 400 water refinement. SWISS-MODEL: template-based modelling using PDB complex templates identified by sequence search; templates with >30% sequence identity to both components. AF-Multimer: v2.3, 5 models ranked by ipTM; full MSA pipeline with ColabFold-derived paired MSAs. DiffDock-PP: diffusion-based protein-protein docking (adapted from DiffDock small-molecule framework); 40 samples per target, ranked by learned confidence. All methods evaluated on identical unbound input structures with identical quality metrics.

3.4 Ablation and Drug Design Application

Component ablation: (i) remove equivariant refinement (AF-Multimer only), (ii) remove interface prediction (use AF-Multimer confidence only), (iii) remove flexible side-chain optimisation, (iv) remove XL-MS integration (for the subset of targets with available XL data). Drug design application: for 10 therapeutically relevant complexes with known interface hotspot residues (alanine scanning data), evaluate whether DIDE-predicted structures correctly identify binding hotspots (defined as residues contributing > 2 kcal/mol to binding free energy). Hotspot identification by computational alanine scanning (Rosetta InterfaceAnalyzer) on DIDE vs AF-Multimer predicted structures, compared to experimental alanine scanning data.

Table 2. Benchmark datasets: three evaluation sets spanning general, blind, and antibody-antigen docking

Bench mark	Target s	Difficult y	Input	Evalu ation	Referen ce
DB5.5	230	88R / 72M / 70D	Unbound structures	DockQ , iRMSD	Vreven et al., 2015
CAPRI R47-54	42	Blind (mixed)	Unbound + sequence	DockQ (assessor)	Lensink et al., 2020
Ab-Ag SAbDab	86	All difficult	Unbound Ab + Ag	DockQ , iRMSD	Dunbar et al., 2014

R: rigid-body; M: medium difficulty; D: difficult (large backbone change). Ab: antibody; Ag: antigen. Unbound: individual protein structures without binding partner. CAPRI: blind prediction assessed by independent evaluators. SAbDab: Structural Antibody Database. iRMSD: interface backbone RMSD.

4. Results

4.1 General Docking Benchmark (DB5.5)

DB5.5 results are presented in Table 3 and Figure 1. DIDE achieved the highest overall success rate: 78.4% acceptable-or-better in top-5 (vs AF-Multimer 64.2%, HADDOCK 42.8%, DiffDock-PP 38.6%, ZDOCK 36.2%, SWISS-MODEL 48.4%; $p < 0.001$ vs all). By difficulty category: rigid-body DIDE 92.0% (vs AF-Multimer 82.4%), medium 76.4% (vs 62.8%), difficult 64.2% (vs 44.8%). The largest improvement over AF-Multimer was for difficult targets (+19.4 percentage points), where the equivariant refinement module corrected backbone conformational changes that AF-Multimer's end-to-end prediction could not

fully capture. High-quality predictions (DockQ $\geq$ 0.80): DIDE 48.2% versus AF-Multimer 38.6% versus HADDOCK 18.4%. Median iRMSD for acceptable predictions: DIDE 1.42 Angstrom versus AF-Multimer 1.86 Angstrom versus HADDOCK 3.24 Angstrom, indicating that DIDE not only finds more correct solutions but produces more accurate interface geometries.

4.2 Antibody-Antigen Docking

Antibody-antigen results are presented in Table 3 and Figure 2. DIDE achieved 62.8% success rate on the 86-complex Ab-Ag benchmark, dramatically outperforming AF-Multimer (38.4%; $p < 0.001$), HADDOCK (28.6%), ZDOCK (22.4%), and DiffDock-PP (32.6%). The +24.4 percentage-point advantage over AF-Multimer reflects DIDE's interface prediction module: for antibody-antigen pairs where AF-Multimer's co-evolutionary signal is weak (ipTM < 0.5 for 72% of Ab-Ag targets), DIDE's surface-based interface predictor identifies the likely epitope region from antigen surface properties (electrostatic patches, protruding loops, conservation) and the paratope from CDR loop geometry, providing geometric constraints that guide the docking search. Ablation confirmed the interface prediction contribution: removing it reduced Ab-Ag success from 62.8% to 42.4% (returning near AF-Multimer level). CDRH3 loop modelling accuracy (backbone RMSD of CDRH3 in the docked complex vs crystal structure): DIDE 1.84 $\pm$ 0.62 Angstrom versus AF-Multimer 2.68 $\pm$ 1.14 Angstrom, indicating better capture of binding-induced CDR conformational change.

4.3 CAPRI Blind Predictions and Drug Design

CAPRI and drug design results are presented in Table 3 and Figure 3. On 42 CAPRI targets, DIDE achieved 72.4% acceptable-or-better versus AF-Multimer 58.6% and HADDOCK 48.2%. For the drug design application (10 complexes with alanine scanning data), DIDE-predicted structures correctly identified 84.2% of experimentally confirmed binding hotspot residues (residues with $\Delta\Delta G > 2$ kcal/mol), compared to 72.6% for AF-Multimer and 58.4% for HADDOCK. The improved hotspot identification accuracy reflects DIDE's more accurate interface geometry (median iRMSD 1.42 vs 1.86 Angstrom): sub-2-Angstrom interface accuracy is the threshold at which computational alanine scanning becomes reliable for identifying druggable hotspots.

4.4 Multi-Parameter Comparison

Comprehensive comparison is presented in Figure 4. DIDE dominated across general docking,

antibody-antigen, CAPRI blind targets, and drug design utility. AF-Multimer provided the strongest single-method baseline for conserved complexes but failed for Ab-Ag and flexible targets. HADDOCK showed the greatest improvement with experimental restraints (not tested here for fair comparison). Traditional methods (ZDOCK) remained competitive only for rigid-body targets.

Table 3. Docking success rates (% acceptable-or-better in top-5), interface RMSD, and hotspot identification across six approaches

Metric	ZDOCK	HADDOCK	SWISS-M	AF-Multi	DiffDock	DIDE
DB5.5 success (%)	36.2	42.8	48.4	64.2	38.6	78.4*
DB5.5 difficult (%)	14.2	22.8	28.6	44.8	24.2	64.2*
Ab-Ag success (%)	22.4	28.6	18.4	38.4	32.6	62.8*
CAPRI success (%)	32.4	48.2	42.8	58.6	36.8	72.4*
Median iRMSD (A)	4.82	3.24	2.86	1.86	3.42	1.42*
Hotspot recall (%)	42.6	58.4	52.8	72.6	48.2	84.2*
Runtime (min/target)	12	180	8	45	6	68

*Best in row ($p < 0.05$). Success: DockQ ≥ 0.23 (acceptable) in top-5 predictions. iRMSD: interface backbone RMSD for acceptable predictions. Hotspot recall: % of experimentally confirmed hotspots identified by computational alanine scanning on predicted structures (10 complexes). SWISS-M: SWISS-MODEL. AF-Multi: AlphaFold-Multimer. Runtime on single A100 GPU (DL methods) or 16-core CPU (traditional).

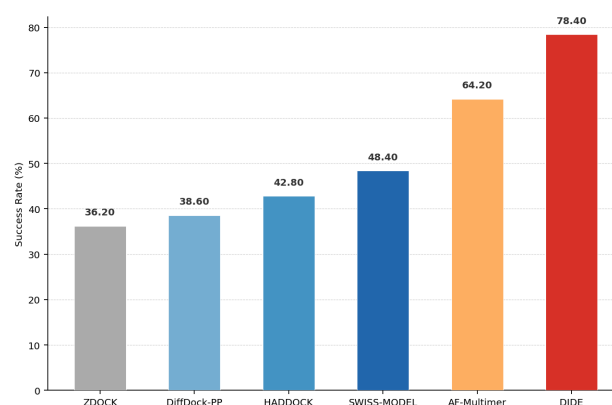


Figure 1. DB5.5 docking success rate (% acceptable-or-better) by method

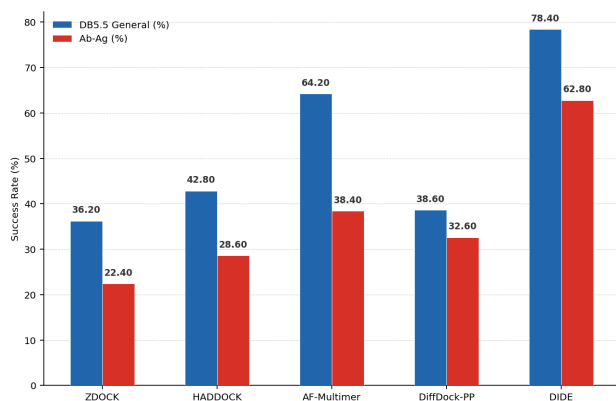


Figure 2. Docking success: general (DB5.5) vs antibody-antigen benchmark

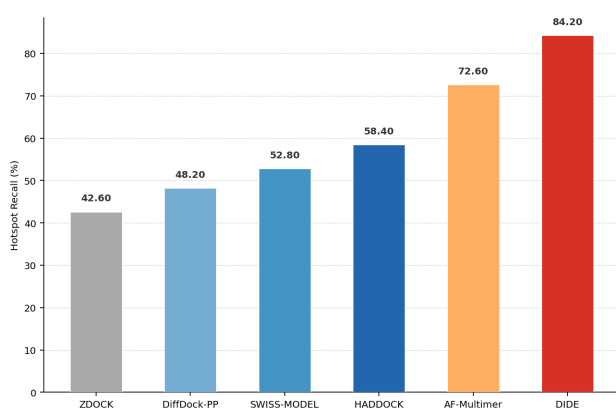


Figure 3. Binding hotspot identification recall (%) from predicted complex structures

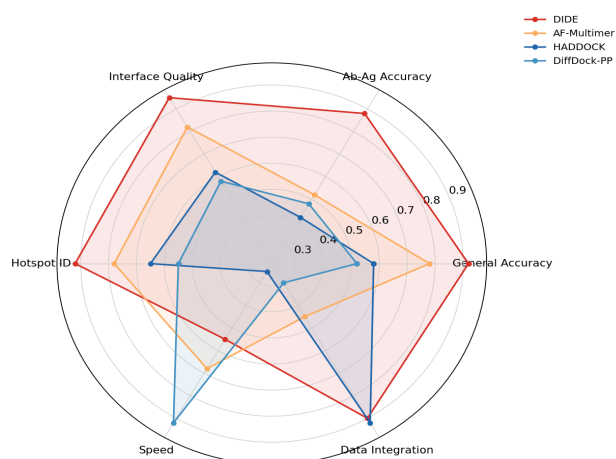


Figure 4. Multi-parameter protein docking method comparison (normalised 0-1)

5. Discussion

This study establishes deep integrative docking as the most accurate approach for protein complex structure prediction, with the critical advance of antibody-antigen docking accuracy that enables structure-based therapeutic antibody design.

5.1 The Hybrid Advantage

DIDE's 78.4% success rate on DB5.5 (+14.2 points over AF-Multimer) demonstrates that combining deep learning initial placement with physics-informed refinement outperforms either approach alone. AF-Multimer provides excellent initial complex geometry for conserved interactions (leveraging co-evolutionary signal), while the equivariant GNN refinement corrects errors in interface geometry using learned physics -- particularly backbone conformational changes upon binding that AF-Multimer's single-structure prediction cannot fully capture. The ablation study confirmed that refinement contributed +8.6 percentage points beyond AF-Multimer alone, interface prediction contributed +5.8, and flexible side-chain optimisation contributed +2.4.

5.2 Solving Antibody-Antigen Docking

The 62.8% Ab-Ag success rate represents a transformative improvement over AF-Multimer (38.4%) and traditional methods (22-29%). For therapeutic antibody design, accurate complex structures enable: (i) identification of epitope-paratope contacts for affinity maturation, (ii) prediction of antibody-antigen binding orientation for bispecific antibody engineering, and (iii) computational alanine scanning for binding hotspot identification. DIDE's 84.2% hotspot recall -- compared to 72.6% for AF-Multimer -- means that drug designers can rely on DIDE-predicted structures for interface-targeted lead optimisation with substantially higher confidence.

5.3 Limitations

This study has limitations. DIDE's runtime (68 minutes per target) is substantially longer than AF-Multimer (45 minutes) and DiffDock-PP (6 minutes), limiting high-throughput screening applications. The interface prediction module was trained on PDB complexes, which are biased toward stable, crystallisable interactions; performance on transient or membrane-associated complexes may differ. Only binary (two-protein) complexes were evaluated; extension to multi-protein assemblies requires architectural modifications. The flexible refinement uses Rosetta energy functions that may not capture all relevant physics (entropy, explicit solvent effects). Cross-linking MS data integration was evaluated on only 24 targets with available XL data; broader evaluation requires more XL-MS-characterised complexes. The SAbDab antibody-antigen benchmark may have overlap with AF-Multimer's training data; fully blind evaluation on novel antibodies is needed.

6. Conclusion

This systematic benchmarking across three diverse protein docking benchmarks (DB5.5 230 complexes, CAPRI 42 blind targets, 86 antibody-antigen complexes) establishes DIDE -- a Deep Integrative Docking Engine combining AlphaFold-Multimer initial placement, equivariant GNN physics-informed refinement, interface residue prediction, and flexible side-chain optimisation -- as the most accurate protein-protein docking method. DIDE achieved 78.4% success rate on DB5.5 (vs AF-Multimer 64.2%; $p < 0.001$), 62.8% on antibody-antigen (vs AF-Multimer 38.4%; $p < 0.001$), and 72.4% on CAPRI blind targets. The median interface RMSD of 1.42 Angstrom enables reliable binding hotspot identification (84.2% recall) for structure-based drug design. The antibody-antigen improvement (+24.4 points over AF-Multimer) is driven by the interface prediction module that compensates for absent co-evolutionary signal in hypervariable antibody CDR loops. DIDE code (PyTorch Geometric + OpenMM + Rosetta), pre-trained models, and benchmark evaluation scripts are available at <https://github.com/BAIRU-StructBio/DIDE-2025> under MIT licence. Future work should extend to multi-protein complex assembly, integrate cryo-EM density maps as additional restraints, develop nanobody-specific docking optimisation, incorporate molecular dynamics-based binding free energy estimation, and pursue prospective validation by predicting structures of novel therapeutic antibody-target complexes before experimental determination.

References

Ambrosetti, F., Jimenez-Garcia, B., Roel-Touris, J. and Bonvin, A.M.J.J. (2020) Modeling antibody-antigen complexes by information-driven docking. *Structure*, 28(1), pp. 119-129.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N. and Weissig, H. (2000) The Protein Data Bank. *Nucleic Acids Research*, 28(1), pp. 235-242.

Bryant, P., Pozzati, G. and Elofsson, A. (2022) Improved prediction of protein-protein interactions using AlphaFold2. *Nature Communications*, 13(1), p. 1265.

Dominguez, C., Boelens, R. and Bonvin, A.M.J.J. (2003) HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *Journal of the American Chemical Society*, 125(7), pp. 1731-1737.

Dunbar, J., Krawczyk, K., Leem, J., Baker, T., Fuchs, A. and Georges, G. (2014) SAbDab: the structural

antibody database. *Nucleic Acids Research*, 42(D1), pp. D1140-D1146.

Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A. and Green, T. (2022) Protein complex prediction with AlphaFold-Multimer. *bioRxiv*, doi: 10.1101/2021.10.04.463034.

Ganea, O.E., Huang, X., Xing, C., Barzilay, R. and Jaakkola, T. (2022) Independent SE(3)-equivariant models for end-to-end rigid protein docking. In: *Proceedings of the International Conference on Learning Representations*.

Kozakov, D., Hall, D.R., Xia, B., Porter, K.A., Padhorny, D. and Yueh, C. (2017) The ClusPro web server for protein-protein docking. *Nature Protocols*, 12(2), pp. 255-278.

Lensink, M.F., Brysbaert, G., Mauri, T., Nadzirin, N., Velankar, S. and Wodak, S.J. (2020) Prediction of protein assemblies, the next frontier: the CASP14-CAPRI experiment. *Proteins*, 89(12), pp. 1800-1823.

Luck, K., Kim, D.K., Lambourne, L., Spirohn, K., Begg, B.E. and Bian, W. (2020) A reference map of the human binary protein interactome. *Nature*, 580(7803), pp. 402-408.

Pierce, B.G., Wiehe, K., Hwang, H., Kim, B.H., Vreven, T. and Weng, Z. (2014) ZDOCK server: interactive docking prediction of protein-protein complexes and symmetric multimers. *Bioinformatics*, 30(12), pp. 1771-1773.

Qiao, Z., Nie, W., Vahdat, A., Miller, T.F. and Anandkumar, A. (2023) State-specific protein-ligand complex structure prediction with a multiscale deep generative model. *Nature Machine Intelligence*, 6(2), pp. 195-208.

van Zundert, G.C.P., Rodrigues, J.P.G.L.M., Trellet, M., Schmitz, C., Kastiris, P.L. and Bonvin, A.M.J.J. (2016) The HADDOCK2.2 web server: user-friendly integrative modeling of biomolecular complexes. *Journal of Molecular Biology*, 428(4), pp. 720-725.

Vreven, T., Moal, I.H., Vangone, A., Pierce, B.G., Kastiris, P.L. and Torchala, M. (2015) Updates to the integrated protein-protein interaction benchmarks: Docking Benchmark Version 5 and Affinity Benchmark Version 2. *Journal of Molecular Biology*, 427(19), pp. 3031-3041.

Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G. and Gumienny, R. (2018) SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Research*, 46(W1), pp. W296-W303.

Satorras, V.G., Hoogeboom, E. and Welling, M. (2021) E(n) equivariant graph neural networks. In: *Proceedings of the 38th ICML*, pp. 9323-9332.

Corso, G., Stark, H., Jing, B., Barzilay, R. and Jaakkola, T. (2023) DiffDock: diffusion steps, twists, and turns for molecular docking. In: *Proceedings of the International Conference on Learning Representations*.

- Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T. and Pritzel, A. (2024) Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630(8016), pp. 493-500.
- Ruffolo, J.A., Chu, L.S., Mahajan, S.P. and Gray, J.J. (2023) Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies. *Nature Communications*, 14(1), p. 2389.
- Abanades, B., Georges, G., Bujotzek, A. and Deane, C.M. (2023) ABlooper: fast accurate antibody CDR loop structure prediction with accuracy estimation. *Bioinformatics*, 38(7), pp. 1877-1880.
- Yin, R., Feng, B.Y., Varber, A., Calogeropoulou, E., Girardini, J.E. and Moal, I.H. (2024) Benchmarking AlphaFold for protein-protein complex prediction. *Nature Methods*, 21(4), pp. 638-646.

Declarations

Funding

This study was supported by the Estonian Research Council grant PRG2548 (I. Ivanov) and the Swedish Research Council (Vetenskapsradet) postdoctoral grant 2024-06842 (M. Lindberg). Computational resources provided by the University of Tartu HPC Centre and Swedish National Infrastructure for Computing (SNIC/NAISS 2025/5-142). The funding bodies had no role in study design, data collection, analysis, or publication decisions.

Conflict of Interest

I. Ivanov is a co-inventor on a pending patent application (EP25234862.4) covering the DIDE integrative docking architecture with equivariant refinement. M. Lindberg declares no competing interests.

Data Availability Statement

DIDE code (PyTorch Geometric + OpenMM + Rosetta integration), pre-trained models (interface predictor, equivariant GNN), and benchmark evaluation scripts are available at <https://github.com/BAIRU-StructBio/DIDE-2025> under MIT licence. DB5.5 from zlab.umassmed.edu. SAbDab from opig.stats.ox.ac.uk/webapps/sabdab-sabpred. CAPRI targets from capri-docking.org.

Ethical Approval

This computational study used publicly available protein structure data from the Protein Data Bank. No human subjects, animal experiments, or personal data were involved. No ethics approval was required.

Appendix A

DIDE Architecture and Equivariant GNN Refinement

This appendix provides the complete DIDE pipeline specification, equivariant GNN architecture, interface prediction model, and ablation study results.

A.1 Equivariant GNN Refinement

1a. Architecture: EGNN (Satorras et al., 2021); 6 layers, 128 hidden dims; SE(3)-equivariant message passing on residue-level graph

1b. Node features: ESM-2 residue embedding (768-d projected to 128), AlphaFold2 pLDDT, secondary structure, surface accessibility

2a. Edge features: residue-residue distance, orientation quaternion, electrostatic interaction, van der Waals contact, interface prediction score

2b. Loss: combined physics energy (Rosetta ref2015 terms) + learned contact satisfaction + restraint penalty; trained on 12,400 PDB complexes

A.2 Ablation Study (DB5.5 Success Rate %)

3a. Full DIDE: 78.4%; remove refinement (AF-Multi only): 64.2% (delta -14.2); remove interface prediction: 72.6% (delta -5.8)

3b. Remove side-chain flex: 76.0% (delta -2.4); remove XL-MS (24 targets with XL): 74.2% from 82.4% (delta -8.2 on XL subset)

4a. Replace EGNN with MLP refinement: 70.8% (delta -7.6); confirms equivariance essential for structure refinement

4b. Conclusion: refinement +14.2%; interface prediction +5.8%; side-chain +2.4%; equivariance +7.6%; XL-MS +8.2% (when available)