What has AlphaFold3 learned about antibody and nanobody docking, and what remains unsolved?

Fatima N. Hitawala, Jeffrey J. Gray
Department of Chemical and Biomolecular Engineering
Johns Hopkins University
Baltimore, MD. 21287
Correspondence to jgray@jhu.edu

Abstract

The development of antibody therapeutics is a major focus in healthcare, owing to their high binding affinity and specificity. To accelerate drug development, significant efforts have been directed toward the design and screening of antibodies. For effective *in silico* development, high modeling accuracy is necessary. To probe the improvement and limitations of AlphaFold3 (AF3), we tested the capability of AF3 to capture the fine details and interplay between antibody structure prediction and antigen docking accuracy. AF3 achieves an 10.2% and 11.4% high-accuracy docking success rate for antibodies and nanobodies, respectively, and a median unbound CDR H3 RMSD accuracy of 2.67 Å and 2.30 Å; CDR H3 accuracy also boosts complex prediction accuracy, and antigen context helps improve CDR H3 accuracy for loops that are greater than 15 residues long. However, AF3's 60% failure rate for antibody and nanobody docking (when sampling a single seed) leaves room for improvement to equip antibody design endeavors.

1 Introduction

Antibodies (Abs) play a critical role in the immune system, and the development of antibody and nanobody therapeutics has become a major interest due to their ability to target cancer, autoimmune, cardiovascular, and infectious diseases, their soluble nature, tunable affinity, and high tolerance by the human body [15]. The antigen (Ag) binding interface of an antibody (nanobody) is composed of six (three) hypervariable loops, called the complementarity determining region (CDR) loops. The third loop on the heavy chain of the antibody (CDR H3) is particularly diverse and known to have the highest number of contacts with the epitope [29]. These CDR loops sometimes undergo conformational changes upon binding to an antigen [14]. Designing antibodies is challenging, primarily due to potential off-target effects [11] and the substantial time and resources required for developability testing [15]. To address these limitations, significant effort has gone into developing antibody and antibody-antigen complex structure predictors [42] (see Related Work in Appendix 4).

Due to the flexibility and importance of antibody CDR loops, being able to model structural movement and docking is highly valuable. Until AlphaFold3 (AF3), the highest success rate achieved for antibody docking was 43% by AlphaRED [22], a hybrid model using AlphaFold2-Multimer (AF2-M) predicted complexes and confidence measures with Rosetta-based replica exchange docking. Despite being trained on the same antibody dataset as AF2-M [9, 27], in May 2024, DeepMind reported a notable 60% success rate for AF3 when 1,000 seeds were sampled.

To understand the source of improvement and where AF3 still has limitations, here we thoroughly assess AF3's ability to dock antibody-antigen and nanobody-antigen complexes and predict unbound antibody and nanobody structures. To understand the limitations of learning from the limited experimental structures provided in the PDB, we study the interplay of the CDR H3 loop and Ab-Ag

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

docking using structures from a redundancy-filtered bespoke dataset after the 2021 AF3 training cutoff.

2 Experiments

2.1 AF3 outperforms previous state-of-the-art antibody docking methods.

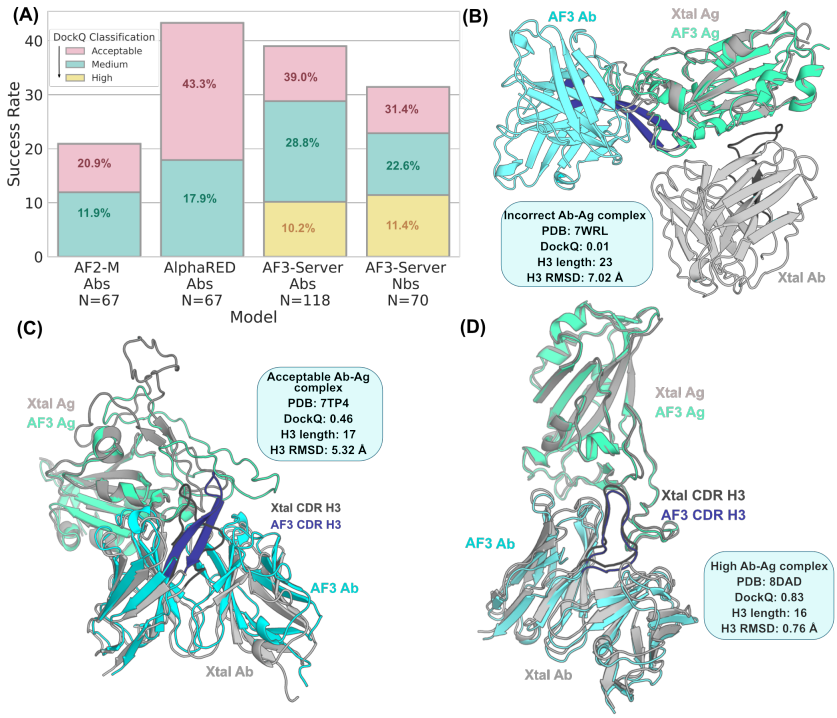

Figure 1: **The success rates in antibody (Ab) and nanobody (Nb) docking for state-of-the-art models. (A)** Performance of AF3 on antibody-antigen docking (N=118) and nanobody-antigen docking (N=70) with our curated dataset against AF2-M and AlphaRED (N=67) using evaluation sets from Harmalkar et al. [22]. DockQ scores for the bound antibodies and nanobodies are binned into incorrect, acceptable, medium, and high categories based on CAPRI classifications [16]. **(B, C, D)** Protein complex structures of example incorrect, acceptable, and high accuracy predictions.

To compare AF3 to previous state-of-the-art models, we first curated a benchmark set of bound and unbound antibodies and nanobodies (Appendix 5). Then we ran three seeds for every target in AF3, choosing the top-ranked decoy in the first seed for comparison consistency to the AlphaRED evaluation set. As seen in Fig. 1, while AlphaRED improves the percent of acceptably docked structures to 43% compared to AF2-M, AF3 improves overall docking quality by increasing the number of high-accuracy structures. AF2-M and AlphaRED both have negligible success rates for high-accuracy docking, while AF3 has a considerably high accuracy success rate of 10.2%, and overall success rate of 39.0% for antibodies. AF3 has a slightly lower success rate (31.4%) for nanobodies and achieves a 11.4% success rate for highly accurate complexes (structures in Fig. 6).

2.2 Antibody structure prediction and antibody-antigen docking interdependently improve overall complex accuracy.

The antibody maturation process serves to improve binding affinity for an expressed antigen [32] so that an unbound antibodies can target the antigen by complementing the paratope [17]. As the hypervariable H3 loop often makes the majority of contacts between the antibody and antigen [29], correctly modeling the CDR H3 loop is pivotal in improving docking quality.

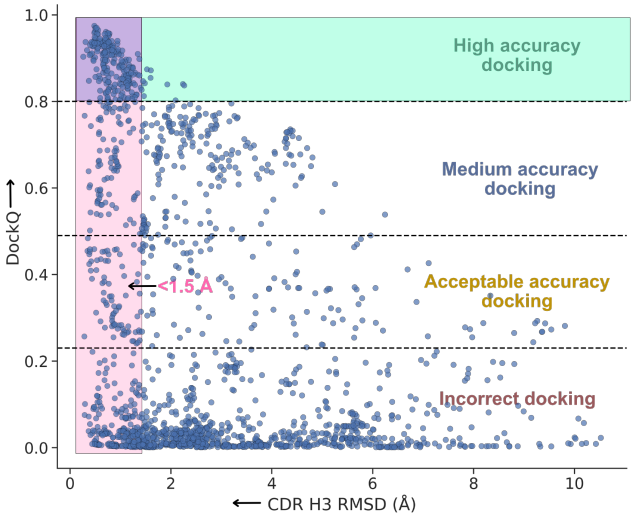

Figure 2: **Distribution of DockQ scores versus CDR H3 loop RMSD of predicted antibody-antigen complexes.** CAPRI classification zones marked, with the high-accuracy docking complex region shaded in green, the less than 1.5 Å CDR H3 loop RMSD region shaded in pink, and the intersection shaded in purple. The conditional probability of the CDR H3 loop RMSD being less than 1.5 Å given a highly accurate complex is the number of points in the intersection of both events (purple) over the number of total points with highly accurate docking (green). The conditional probability of a highly accurate complex given a less than 1.5 Å H3 loop RMSD is the number of points in the intersection (purple) over the number of points in the sub-angstrom H3 loop RMSD region (pink).

To understand the correlation between modeling the CDR H3 loop and docking accuracy, we measured CDR H3 RMSD for AF3 predictions of antibody-antigen complexes and then compared the conditional probabilities between DockQ and CDR H3 RMSD at each level of docking accuracy (method in Fig. 2's caption). As shown in Fig. 2, for highly accurate complexes, the $p(\text{DockQ} > 0.8 \mid \text{ CDR H3 RMSD} \leq 1.5 \text{ Å})$ is 27.8%, while the $p(\text{ CDR H3 RMSD} \leq 1.5 \text{ Å} \mid \text{DockQ} > 0.8)$ becomes very high: 96.9%, implying that a correct CDR H3 loop is critical for high-quality docking predictions. The inequality between the conditional probabilities flips when comparing the effect of CDR H3 RMSD less than 1 Å on complexes with acceptable or better accuracy (Tables 1, 2).

2.3 AF3 outperforms AF2.3-M, AF2-M, and IgFold in predicting unbound Fv structures.

Considering the impact of CDR H3 loop accuracy on overall docking success, we sought to evaluate AF3's predictive accuracy for unbound CDR H3 loops. To compare AF3 against previous state-of-the-art structure prediction models, we use IgFold's curated benchmark of 196 unbound antibody variable fragments, 70 unbound nanobodies, and published results for IgFold and AF2.0-M [36]. We predicted one seed from AF3 (seed = 1) and used the top-ranked antibody from the five decoys predicted for the comparison, with ranking determined by AF3's combination of ipTM, pTM, and disorder confidence measures [9] (Fig. 4).

While AF2.3-M has the lowest median CDR H3 RMSD of the three previous state-of-the-art models at 2.73 Å, AF3 achieves 2.02 Å. IgFold and AF2.3-M perform similarly (2.87 Å, 2.73 Å respectively), with AF2-M performing the worst (2.99 Å). Thus, AF3 has improved CDR H3 loop prediction by 0.71Å ($p \leq 0.02$). The accuracy plateau reached by IgFold, AF2-M, and AF2.3-M led to questions about the limit of accuracy for the loop considering its flexibility. In a recent survey of 177 pairs of bound-unbound antibody complexes, Liu et al. [14] found that in 70.6% of antibody CDR H3 loops, binding-induced conformational changes are under 1 Å (Xtal B-U column in 3). Thus, despite the loop's potential flexibility, in principle, models may someday be able to reach sub-angstrom predictive accuracy, implying room for improvement.

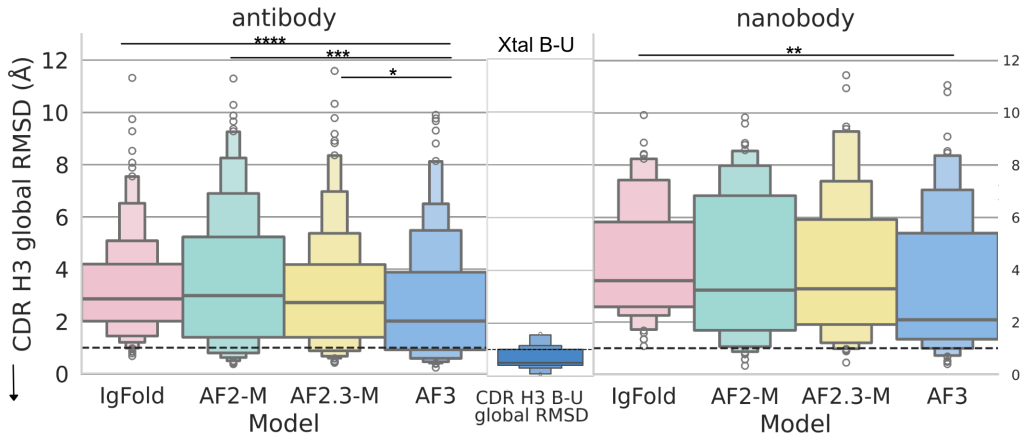

Figure 3: **Performance of AF3 on predicting unbound CDR H3 loop structures of 196 antibodies and 70 nanobodies compared to previous models. (A)** AF3 improves upon the average median to 2.02 Å. For reference, unbound and antigen-bound antibody CDR H3 loops vary by a median of 0.5 Å. **(B)** While IgFold, AF2.0-M, and AF2.3-M have a median H3 RMSD for nanobodies above 3 Å, AF3 improves the median to 2.08 Å.

2.4 Antigen context affects antibody CDR H3 loop prediction accuracy.

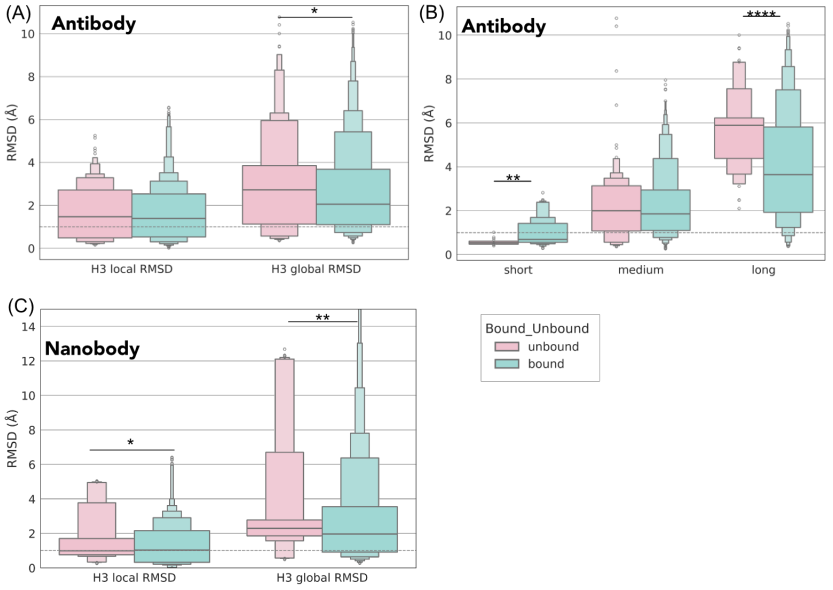

Figure 4: **Effect of antigen context on antibody CDR H3 loop position and shape prediction accuracy and CDR H3 loop prediction per length, and effect of antigen context on nanobody CDR H3 loop position and shape prediction accuracy. (A)** Bound structures (N = 1,770) predicted with antigen, unbound structures (N = 300) predicted with the antibody alone. Global RMSD calculated after superposition of the VH domain; local RMSD calculated by superposing the loop residues only. **(B)** Bound structures predicted performed with antigen, unbound structures predicted with the antibody alone. Short loops (N = 165) are defined as less than 10 residues, medium loops (N = 1,335) between 10 to 15 residues, and long loops (N = 570) longer than 15 residues. **(C)** Bound structures (N = 1,005) predicted with antigen, unbound structures (N = 150) predicted with the nanobody alone. Global RMSD calculated after superposition of the VH domain; local RMSD calculated by superposing the loop residues only.

To further observe how the H3 loop's structure is affected by antigen context, we compared the RMSD of all bound antibody H3 loops against all unbound antibody H3 loops. We choose to keep all five diffusion decoys instead of filtering for top-ranked decoys as it reduces the power of the Mann-Whitney U test, despite the trend in the difference between bound and unbound RMSDs staying consistent. Additionally, the effect of antigen context should remain a consistent trend for all decoys, not just the top-ranked. We divide the RMSDs into the global H3 RMSD, calculated by superposing heavy chains, and the local H3 RMSD, calculated by superposing only the CDR H3 residues. The global RMSD captures the loop shape and placement, while the local RMSD only represents the loop's shape. As shown in Fig. 4 panel A, the local RMSD of AF3 is not significantly different if antigen context is provided, but has a lower median (1.47 Å for unbound, 1.40 Å for bound). The global RMSD does significantly (p=0.02) improve when given antigen context, from 2.67 Å to 2.05 Å.

Previous antibody structure prediction studies report H3 loop RMSD increasing as loop length increases [35] due to the degrees of freedom granted to the loop. As seen in Fig. 4 panel B, CDR H3 accuracy is improved significantly (p=4.12e-8) from 5.89 Å to 3.64 Å when antigen context is present to help constrain longer loops. As the nanobody dataset is smaller, we simply compare the significance of antigen context between the local and global CDR H3 prediction accuracy. We find local loop structure prediction improves (from 1.04 Å to 0.99 Å) given antigen context (p=0.05), as does global loop structure prediction (from 2.30 Å to 1.97 Å) (p=0.008).

3 Discussion

AF3 builds on established performance-enhancing methods for AF2 (see Related Work 4) and protein data representation [41]. In our work, we demonstrate that AF3 improves the rate of high-quality docked antibody-antigen from 0 to 10.4%; however, AF3 (with one seed) still leaves 60% of the targets incorrect. The accuracy of the model has been reported to increase to 60% when evaluating 1,000 seeds in non-reproducible work [9]. The model appears to have difficulty finding the correct antigen interface for a given antibody interface. AF3 struggles the most in global docking and may be enhanced by a global sampling protocol such as AlphaRED [22]. AF3's confidence measures correlate with docking accuracy, with a high overall linear correlation between DockQ and ipTM between the heavy chain and antigen (ipTM-HA) (Fig. 7).

Generative models learn continuous distributions instead of discrete points, making them suited to simulate protein plasticity [26, 25]. Thus, AF3's generative framework, combined with simultaneous structure and docking prediction, led us to study whether AF3 can capture the relationships between CDR H3 loop prediction and Ab-Ag docking. We find that AF3 learns the dependence between the antibody binding interface and docking accuracy, perhaps because the model predicts structure and the docked complex simultaneously [9]. AF3 demonstrates a 2.05 Å CDR H3 RMSD when given the antigen sequence. Another generative network, AbODE, achieves a 1.73 Å average antibody CDR H3 loop RMSD when conditioned on the antigen sequence and structure using a dataset from SabDab with a CDR sequence redundancy of 40% through MMseq2 (the same as AF3). The multi-resolution structures of AF3 and AbODE can be interpreted to have initial steps corresponding to rigid backbone sampling and later steps learning local docking [18]. The success of both models may be due to the concurrent structure and docking prediction task. As per Yin et al. [43], recycling plays an important part in docking accuracy in AF2.3-M. Recycling may also bolster AF3, especially with antigen context, as the model iteratively refines the local and global structure of the complex.

Our study was limited by the number of jobs available per day on the AF3 server, and the small dataset size due to data contamination concerns. Our results show that while AF3 is considerably more accurate in modeling antibody and nanobody structures and docked complexes than previous approaches, there remains room for improvement of the 60% failure rate for both antibody and nanobody docking for single seed predictions. The impressive CDR H3 loop structure prediction accuracy for bound (2.05 Å) and unbound structures (2.67 Å) can also in principle still be improved relative to the limited overall bound-to-unbound conformational changes in antibodies measured by Liu et al. [14].

References

[1] AFsample: improving multimer prediction with AlphaFold using massive sampling | Bioinformatics | Oxford Academic, . URL https://academic.oup.com/bioinformatics/article/39/9/btad573/7274860.

[2] ANARCI: antigen receptor numbering and receptor classification | Bioinformatics | Oxford Academic, . URL https://academic.oup.com/bioinformatics/article/32/2/298/1743894?login=false.

[3] Biopython: freely available Python tools for computational molecular biology and bioinformatics | Bioinformatics | Oxford Academic, . URL https://academic.oup.com/bioinformatics/article/25/11/1422/330687.

[4] IgLM: Infilling language modeling for antibody sequence design - ScienceDirect, . URL https://www.sciencedirect.com/science/article/pii/S2405471223002715.

[5] PROTEINS: Structure, Function, and Bioinformatics | Protein Science Journal | Wiley Online Journal, . URL https://onlinelibrary.wiley.com/doi/10.1002/prot.10092.

[6] SciPy 1.0: fundamental algorithms for scientific computing in Python | Nature Methods, . URL https://www.nature.com/articles/s41592-019-0686-2.

[7] B. Abanades, W. K. Wong, F. Boyles, G. Georges, A. Bujotzek, and C. M. Deane. ImmuneBuilder: Deep-Learning models for predicting the structures of immune proteins. *Communications Biology*, 6(1):1–8, May 2023. ISSN 2399-3642. doi: 10.1038/s42003-023-04927-7. URL https://www.nature.com/articles/s42003-023-04927-7. Publisher: Nature Publishing Group.

[8] K. R. Abhinandan and A. C. R. Martin. Analysis and improvements to Kabat and structurally correct numbering of antibody variable domains. *Molecular Immunology*, 45(14):3832–3839, Aug. 2008. ISSN 0161-5890. doi: 10.1016/j.molimm.2008.05.022. URL https://www.sciencedirect.com/science/article/pii/S0161589008002046.

[9] J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A. J. Ballard, J. Bambrick, S. W. Bodenstein, D. A. Evans, C.-C. Hung, M. O'Neill, D. Reiman, K. Tunyasuvunakool, Z. Wu, A. Žemgulytė, E. Arvaniti, C. Beattie, O. Bertolli, A. Bridgland, A. Cherepanov, M. Congreve, A. I. Cowen-Rivers, A. Cowie, M. Figurnov, F. B. Fuchs, H. Gladman, R. Jain, Y. A. Khan, C. M. R. Low, K. Perlin, A. Potapenko, P. Savy, S. Singh, A. Stecula, A. Thillaisundaram, C. Tong, S. Yakneen, E. D. Zhong, M. Zielinski, A. Žídek, V. Bapst, P. Kohli, M. Jaderberg, D. Hassabis, and J. M. Jumper. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630 (8016):493–500, June 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07487-w. URL https://www.nature.com/articles/s41586-024-07487-w. Publisher: Nature Publishing Group.

[10] F. Ambrosetti, B. Jiménez-García, J. Roel-Touris, and A. M. J. J. Bonvin. Modeling Antibody-Antigen Complexes by Information-Driven Docking. *Structure*, 28(1):119–129.e2, Jan. 2020. ISSN 0969-2126. doi: 10.1016/j.str.2019.10.011. URL https://www.cell.com/structure/abstract/S0969-2126(19)30352-1. Publisher: Elsevier.

[11] P. Chames, M. Van Regenmortel, E. Weiss, and D. Baty. Therapeutic antibodies: successes, limitations and hopes for the future. *British Journal of Pharmacology*, 157(2):220–233, May 2009. ISSN 0007-1188. doi: 10.1111/j.1476-5381.2009.00190.x. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2697811/.

[12] S. Chaudhury, S. Lyskov, and J. J. Gray. PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics*, 26(5):689, Mar. 2010. doi: 10.1093/bioinformatics/btq007. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2828115/.

[13] L.-S. Chu, J. A. Ruffolo, A. Harmalkar, and J. J. Gray. Flexible protein–protein docking with a multitrack iterative transformer. *Protein Science*, 33(2):e4862, 2024. ISSN 1469-896X. doi: 10.1002/pro.4862. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/pro.4862. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pro.4862.

[14] Chu'nan Liu, Lilian M Denzler, O. E. Hood, and Andrew C R Martin. Do antibody CDR loops change conformation upon binding? *mAbs*, 2024. doi: 10.1080/19420862.2024.2322533. S2ID: eef3a70d7719a01bfa1dd6a6f99f420b2c1597aa.

[15] M. F. Chungyoun and J. J. Gray. AI models for protein design are driving antibody engineering. *Current Opinion in Biomedical Engineering*, 28:100473, Dec. 2023. ISSN 2468-4511. doi: 10.1016/j.cobme.2023.100473. URL https://www.sciencedirect.com/science/article/pii/S2468451123000296.

[16] K. W. Collins, M. M. Copeland, G. Brysbaert, S. J. Wodak, A. M. J. J. Bonvin, P. J. Kundrotas, I. A. Vakser, and M. F. Lensink. CAPRI-Q: The CAPRI resource evaluating the quality of predicted structures of protein complexes. *Journal of Molecular Biology*, 436(17):168540, Sept. 2024. ISSN 0022-2836. doi: 10.1016/j.jmb.2024.168540. URL https://www.sciencedirect.com/science/article/pii/S0022283624001359.

[17] S. Conti, V. Ovchinnikov, J. G. Faris, A. K. Chakraborty, M. Karplus, and K. G. Sprenger. Multiscale affinity maturation simulations to elicit broadly neutralizing antibodies against HIV. *PLoS Computational Biology*, 18(4):e1009391, Apr. 2022. ISSN 1553-734X. doi: 10.1371/journal.pcbi.1009391. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9020693/.

[18] G. Corso, A. Deng, N. Polizzi, R. Barzilay, and T. S. Jaakkola. The Discovery of Binding Modes Requires Rethinking Docking Generalization.

[19] G. Corso, H. Stärk, B. Jing, R. Barzilay, and T. Jaakkola. DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking, Feb. 2023. URL http://arxiv.org/abs/2210.01776. arXiv:2210.01776 [physics, q-bio].

[20] J. Dunbar, K. Krawczyk, J. Leem, T. Baker, A. Fuchs, G. Georges, J. Shi, and C. M. Deane. SAbDab: the structural antibody database. *Nucleic Acids Research*, 42(Database issue):D1140–D1146, Jan. 2014. ISSN 0305-1048. doi: 10.1093/nar/gkt1043. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3965125/.

[21] R. Evans, M. O'Neill, A. Pritzel, N. Antropova, A. Senior, T. Green, A. Žídek, R. Bates, S. Blackwell, J. Yim, O. Ronneberger, S. Bodenstein, M. Zielinski, A. Bridgland, A. Potapenko, A. Cowie, K. Tunyasuvunakool, R. Jain, E. Clancy, P. Kohli, J. Jumper, and D. Hassabis. Protein complex prediction with AlphaFold-Multimer, Oct. 2021. URL http://biorxiv.org/lookup/doi/10.1101/2021.10.04.463034.

[22] A. Harmalkar, S. Lyskov, and J. J. Gray. Reliable protein-protein docking with AlphaFold, Rosetta, and replica-exchange. *eLife*, 13, Feb. 2024. doi: 10.7554/eLife.94029.1. URL https://elifesciences.org/reviewed-preprints/94029. Publisher: eLife Sciences Publications Limited.

[23] B. Jiménez-García, J. Roel-Touris, M. Romero-Durana, M. Vidal, D. Jiménez-González, and J. Fernández-Recio. LightDock: a new multi-scale approach to protein–protein docking. *Bioinformatics*, 34(1):49–55, Jan. 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/btx555. URL https://doi.org/10.1093/bioinformatics/btx555.

[24] W. Jin, J. Wohlwend, R. Barzilay, and T. Jaakkola. Iterative Refinement Graph Neural Network for Antibody Sequence-Structure Co-design, Jan. 2022. URL http://arxiv.org/abs/2110.04624. arXiv:2110.04624 [cs, q-bio].

[25] B. Jing, E. Erives, P. Pao-Huang, G. Corso, B. Berger, and T. Jaakkola. EigenFold: Generative Protein Structure Prediction with Diffusion Models. *ArXiv*, page arXiv:2304.02198v1, Apr. 2023. ISSN 2331-8422. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10104185/.

[26] B. Jing, B. Berger, and T. Jaakkola. AlphaFold Meets Flow Matching for Generating Protein Ensembles, Sept. 2024. URL http://arxiv.org/abs/2402.04845. arXiv:2402.04845 [cs, q-bio].

[27] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, Aug. 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2. URL https://www.nature.com/articles/s41586-021-03819-2. Publisher: Nature Publishing Group.

[28] D. Kozakov, D. R. Hall, B. Xia, K. A. Porter, D. Padhorny, C. Yueh, D. Beglov, and S. Vajda. The ClusPro web server for protein-protein docking. *Nature protocols*, 12(2):255–278, Feb. 2017. ISSN 1754-2189. doi: 10.1038/nprot.2016.169. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5540229/.

[29] Liang Zhao, Limsoon Wong, and Jinyan Li. Antibody-Specified B-Cell Epitope Prediction in Line with the Principle of Context-Awareness | IEEE Journals & Magazine | IEEE Xplore, Mar. 2011. URL https://ieeexplore.ieee.org/document/5728794.

[30] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, M. Fazel-Zarandi, T. Sercu, S. Candido, and A. Rives. Language models of protein sequences at the scale of evolution enable accurate structure prediction.

[31] M. Mirdita, K. Schütze, Y. Moriwaki, L. Heo, S. Ovchinnikov, and M. Steinegger. ColabFold: making protein folding accessible to all. *Nature Methods*, 19(6):679–682, June 2022. ISSN 1548-7105. doi: 10.1038/s41592-022-01488-1. URL https://www.nature.com/articles/s41592-022-01488-1. Publisher: Nature Publishing Group.

[32] A. K. Mishra and R. A. Mariuzza. Insights into the Structural Basis of Antibody Affinity Maturation from Next-Generation Sequencing. *Frontiers in Immunology*, 9:117, Feb. 2018. ISSN 1664-3224. doi: 10.3389/fimmu.2018.00117. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5799246/.

[33] J. Peng and J. Xu. Raptorx: Exploiting structure information for protein alignment by statistical inference. *Proteins: Structure, Function, and Bioinformatics*, 79(S10):161–171, 2011. ISSN 1097-0134. doi: 10.1002/prot.23175. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.23175. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.23175.

[34] R. M. Rao, J. Liu, R. Verkuil, J. Meier, J. Canny, P. Abbeel, T. Sercu, and A. Rives. MSA Transformer. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8844–8856. PMLR, July 2021. URL https://proceedings.mlr.press/v139/rao21a.html. ISSN: 2640-3498.

[35] J. A. Ruffolo, J. J. Gray, and J. Sulam. Deciphering antibody affinity maturation with language models and weakly supervised learning, Dec. 2021. URL http://arxiv.org/abs/2112.07782. arXiv:2112.07782 [cs, q-bio].

[36] J. A. Ruffolo, L.-S. Chu, S. P. Mahajan, and J. J. Gray. Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies. *Nature Communications*, 14(1):2389, Apr. 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-38063-x. URL https://www.nature.com/articles/s41467-023-38063-x.

[37] Schrödinger, LLC. The PyMOL molecular graphics system, version 1.8. November 2015.

[38] A. Sircar and J. J. Gray. SnugDock: Paratope Structural Optimization during Antibody-Antigen Docking Compensates for Errors in Antibody Homology Models. *PLoS Computational Biology*, 6(1):e1000644, Jan. 2010. ISSN 1553-734X. doi: 10.1371/journal.pcbi.1000644. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2800046/.

[39] Y. Verma, M. Heinonen, and V. Garg. AbODE: Ab Initio Antibody Design using Conjoined ODEs, May 2023. URL http://arxiv.org/abs/2306.01005. arXiv:2306.01005 [cs, q-bio].

[40] M. L. Waskom. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60): 3021, 2021. doi: 10.21105/joss.03021. URL https://doi.org/10.21105/joss.03021.

[41] J. L. Watson, D. Juergens, N. R. Bennett, B. L. Trippe, J. Yim, H. E. Eisenach, W. Ahern, A. J. Borst, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, N. Hanikel, S. J. Pellock, A. Courbet, W. Sheffler, J. Wang, P. Venkatesh, I. Sappington, S. V. Torres, A. Lauko, V. De Bortoli, E. Mathieu, S. Ovchinnikov, R. Barzilay, T. S. Jaakkola, F. DiMaio, M. Baek, and D. Baker. De novo design of protein structure and function with RFdiffusion. *Nature*, 620(7976):1089–1100, Aug. 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06415-8. URL https://www.nature.com/articles/s41586-023-06415-8. Publisher: Nature Publishing Group.

[42] B. D. Weitzner, J. R. Jeliazkov, S. Lyskov, N. Marze, D. Kuroda, R. Frick, J. Adolf-Bryfogle, N. Biswas, R. L. Dunbrack, and J. J. Gray. Modeling and docking of antibody structures with Rosetta. *Nature Protocols*, 12(2):401–416, Feb. 2017. ISSN 1750-2799. doi: 10.1038/nprot.2016.180. URL https://www.nature.com/articles/nprot.2016.180. Publisher: Nature Publishing Group.

[43] R. Yin and B. G. Pierce. Evaluation of AlphaFold Antibody-Antigen Modeling with Implications for Improving Predictive Accuracy. *bioRxiv*, page 2023.07.05.547832, July 2023. doi: 10.1101/2023.07.05.547832. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10349958/.

[44] F. Zong, C. Long, W. Hu, S. Chen, W. Dai, Z.-X. Xiao, and Y. Cao. Abalign: a comprehensive multiple sequence alignment platform for B-cell receptor immune repertoires. *Nucleic Acids Research*, 51(W1):W17–W24, May 2023. ISSN 0305-1048. doi: 10.1093/nar/gkad400. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10320167/.

4 Related Work

Traditional docking algorithms Traditional Rosetta-based antibody-antigen docking algorithms use ensembles of homology models of antibodies, and sampling of rigid backbones, loop conformation, and VH-VL relative orientations [42]. This protocol has a 20% success rate for antibody-antigen docking, defined by a DockQ score > 0.23 [10]. Several other physics- and structure-based methods have been published with similar performance [38, 23, 5, 28]. While these methods are independent of data, the calculations are very time-consuming. Like here, success rates were also limited by the ability to accurately model CDR H3.

ML antibody-antigen structure and complex prediction Machine learning methods have significantly improved antibody-antigen and nanobody-antigen complex prediction and need less time. There are a variety of methods, with some focusing on only structure prediction or docking [36, 7, 19, 25, 13], while others combine the tasks [27, 39, 24]. Tested architectures have included convolutional neural networks, transformers, diffusion models, and normalizing flow models for structure and complex prediction [33, 34, 36, 19, 24, 39, 25, 26], with the dataset and loss functions being adjusted for the task. While these methods have improved protein-protein docking success rates, especially for complexes that have co-evolved, the lack of co-evolution context between antibody-antigen complexes [1] has stunted overall docking success rates, with AF2.3-M having a 20% success rate for antibody-antigen docking [10, 22]. In addition to docking and structure prediction tasks, context extraction from query sequences has also been improved through neural networks that process MSAs [27, 21, 9, 34], and protein language models (PLMs) [30, 4].

Methods enhancing AF2 docking performance After the release of AF2, Mirdita et al. found that adding glycine linkers between monomeric chains allowed the model to predict multimers [31], essential to model both chains of an antibody variable fragment. Evans et al. then published a combined structure and complex prediction model: AF2-M [21]. The updated AF2.3-M model was trained with an expanded dataset (training date cutoff change from 04/30/2018 to 09/30/2021). Many

methods have sought to improve the docking performance of AF2.3-M, such as large sampling rates with increased diversity via tuned dropout rates and generative frameworks [1, 25], replica exchange docking applied to AF2.3-M predicted complexes [22], and MSA sub-sampling [34]. MSA sub-sampling and generative frameworks have been proven to extract conformational change information from sequence data, while massive sampling and AlphaRED have been shown to improve sampling diversity for multimers with little co-evolutionary context. When comparing AF2 and ESMFold adapted into generative networks via a normalizing flow objective, Jing et al [26] found that context from MSA sub-sampling combined with AlphaFold provided richer context for extracting structural conformation data than ESM's embeddings. This model has not been extended to multimer prediction yet, but would likely further improve antibody-antigen docking prediction accuracy when considering the effect of seeds for antibody-antigen complexes seen in [9]. AF3 is a culmination of many of these established methods.

5 Benchmark dataset

5.0.1 Individual structure evaluation set

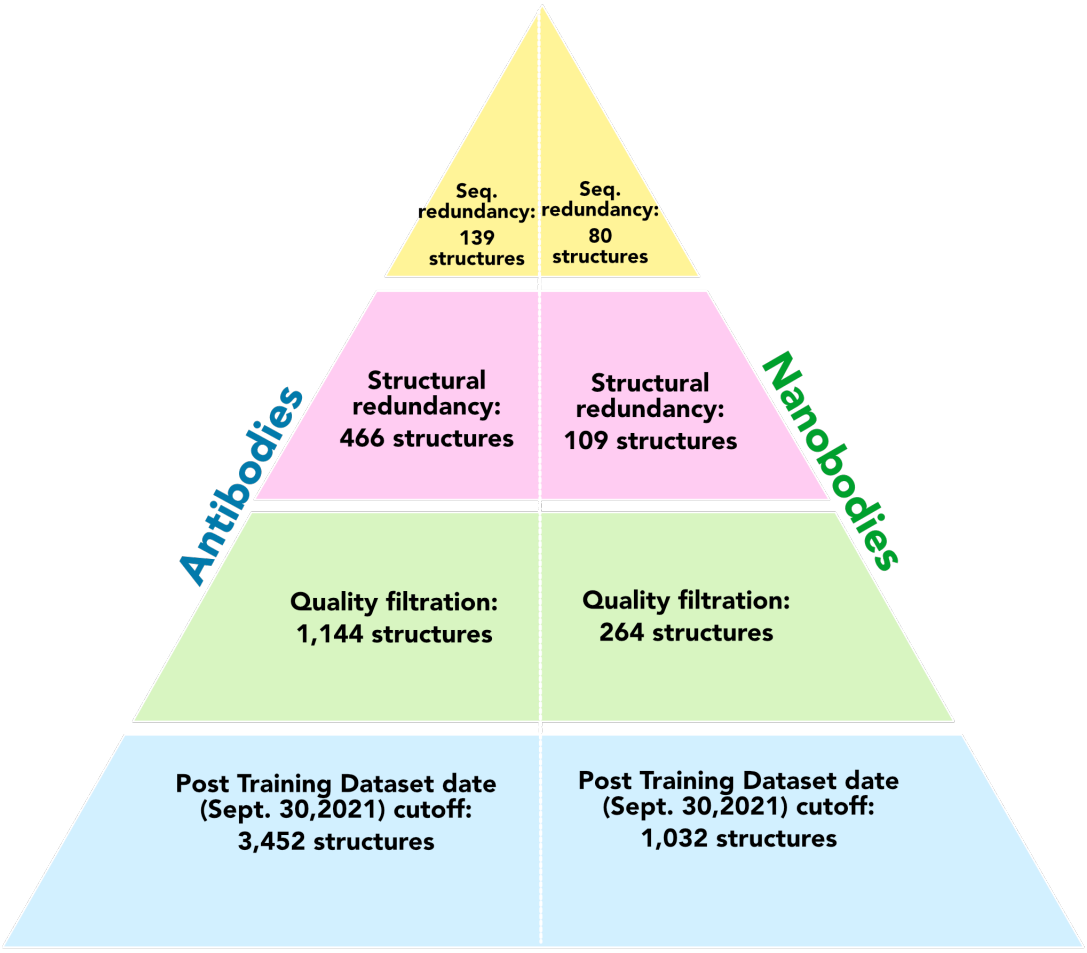

Figure 5: **The process of dataset curation for evaluating antibody structure prediction and how many structures came out of each step.**

To create an immunoglobulin structure prediction dataset, we pulled structures from SabDab (May 31, 2024 for Abs, June 4, 2024 for Nbs) [20], and temporally separated evaluation structures using the Sept. 30, 2021 training dataset date cutoff by AF3. Firstly, we separated all antibody copies in the remaining PDBs, then conducted a quality filtration similar to [14], where we removed PDBs with a

resolution $\geq$ 2.8 Å, and had missing residues in the CDR loops by comparing atomic sequence and sequence residues using the Bio.Seq python package [3]. We used the Kabsch alignment algorithm to calculate the structural redundancy of variable heavy and light chains between pairs of structures [36], and kept pairs of structures that had heavy and light chain RMSD > 1Å, but only one representative out of a pair of redundant structures. Finally, we filtered the remaining structures based on sequence redundancy against each other and AF3's training set with a sequence identity cutoff of 99% and 95% respectively. We conducted MSA alignment on the heavy and light chains separately via Abalign [44], and then a custom Python function to calculate the sequence identities. The number of structures at each step of this process is shown in Figure 5. We then cropped to the variable fragment region using a custom function to prevent the RMSD calculations from being confounded by small hinge movements in the loop connecting the variable and constant regions. Finally, we renumbered the structures using the AbNum webserver [8] using the Chothia scheme.

6 Evaluation Methods

6.0.1 AF2.3-M predictions

We used a local ColabFold installation downloaded from (https://github.com/YoshitakaMo/localcolabfold) with AlphaFold-Multimer version 2.3. We predicted a single decoy for each target and did not use templates. Similar to the curated benchmark, we then cropped the predictions to the variable fragment region using a custom function, and renumbered the structures using the AbNum webserver [8], using the Chothia scheme.

6.0.2 AF3 predictions

We used the AF3 server (https://alphafoldserver.com) to generate decoys. The server generates five decoys (diffusion samples) per seed. To test the diversity produced by seeds, we predicted three seeds per target, with the seed number pre-set to either one, two, or three by using the JSON file upload option. We cropped the predictions to the variable fragment region using a custom function, and renumbered the structures using the AbNum webserver [8] using the Chothia scheme

6.0.3 RMSD calculations

To calculate RMSD, we used the PyRosetta AntibodyInfo class [12]. For calculating the H3 global RMSD, we used the CDR backbone function, and for H3 local RMSD, we extracted sub-poses for the H3 loops also using the AntibodyInfo class, then calculated the Cα loop RMSD using the Rosetta Scoring class.

6.0.4 Figures and statistical analysis

We conducted Mann-Whitney-U and Pearson correlation statistical analyses through the Scipy package [6] in Python (https://www.python.org), and generated boxen plots, scatter plots, and regression plots using the Seaborn package [40] also from Python. We used PyMol 3.0 (Schrodinger, Inc.) [37] for protein structure visualization images.

6.0.5 Average H3 pLDDT calculations

We used AbNumber [2] (IMGT scheme) and heavy chain sequences to extract pLDDTs of Cα residues for each H3 loop from AF3's confidence summary files, then averaged them in Python.

7 Additional experiments

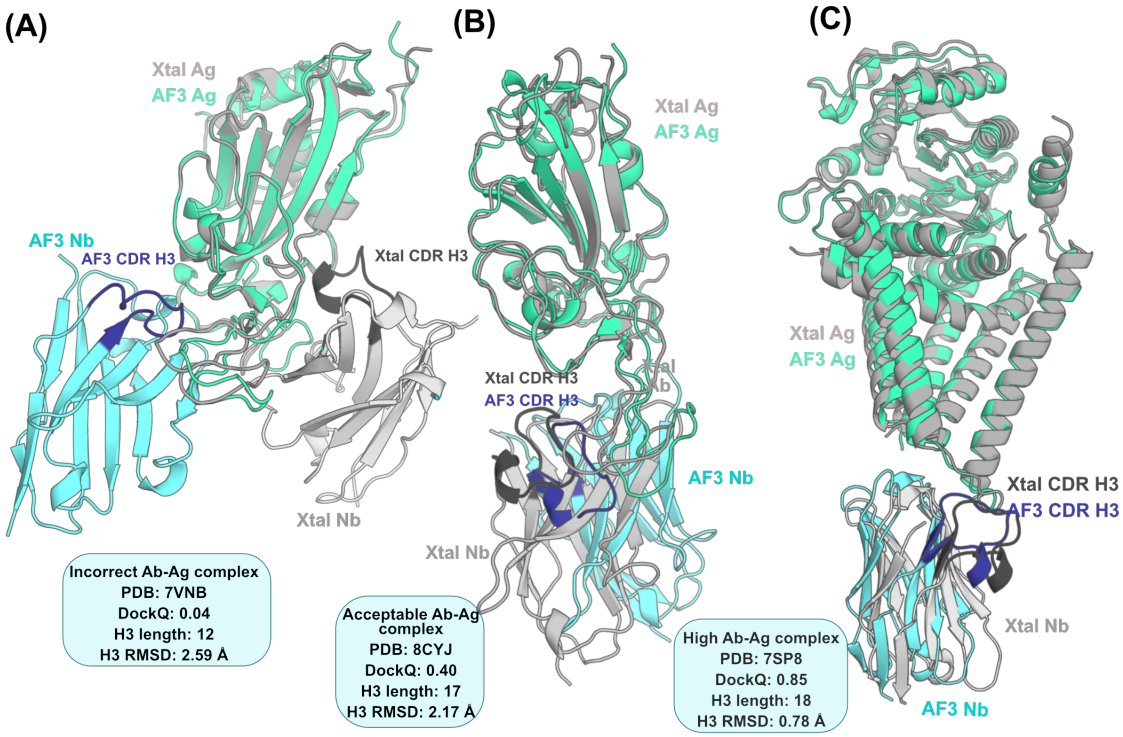

Figure 6: **The figure shows examples of incorrect, acceptable, and high accuracy docked nanobody-antigen complexes.** **(A)** shows an incorrect nanobody-antigen complex compared to the crystal structure. The prediction samples the incorrect antigen interface and has a lower antigen structure prediction accuracy. **(B)** shows an acceptably docked nanobody-antigen complex, where the antigen structure prediction accuracy is low, affecting the sampled paratope. The nanobody CDR H3 loop's shape is correct, but the positioning is not. **(C)** The antigen and CDR H3 loop of the nanobody are correctly predicted, and the correct binding interface is sampled.

Table 1: **Conditional probability of docking accuracies given varying CDR H3 loop accuracies**

DockQ Class \ H3 RMSD	<1 Å	<1.5 Å	<2 Å
High	0.34	0.28	0.22
High and Medium	0.40	0.42	0.38
High, Medium, and Acceptable	0.62	0.53	0.48
Incorrect	0.26	0.38	0.40

Table 2: **Conditional probability of CDR H3 loop accuracies given varying docking accuracies**

DockQ Class \ H3 RMSD	<1 Å	<1.5 Å	<2 Å
High	0.67	0.97	0.99
High and Medium	0.43	0.63	0.73
High, Medium, and Acceptable	0.38	0.58	0.66
Incorrect	0.32	0.37	0.43

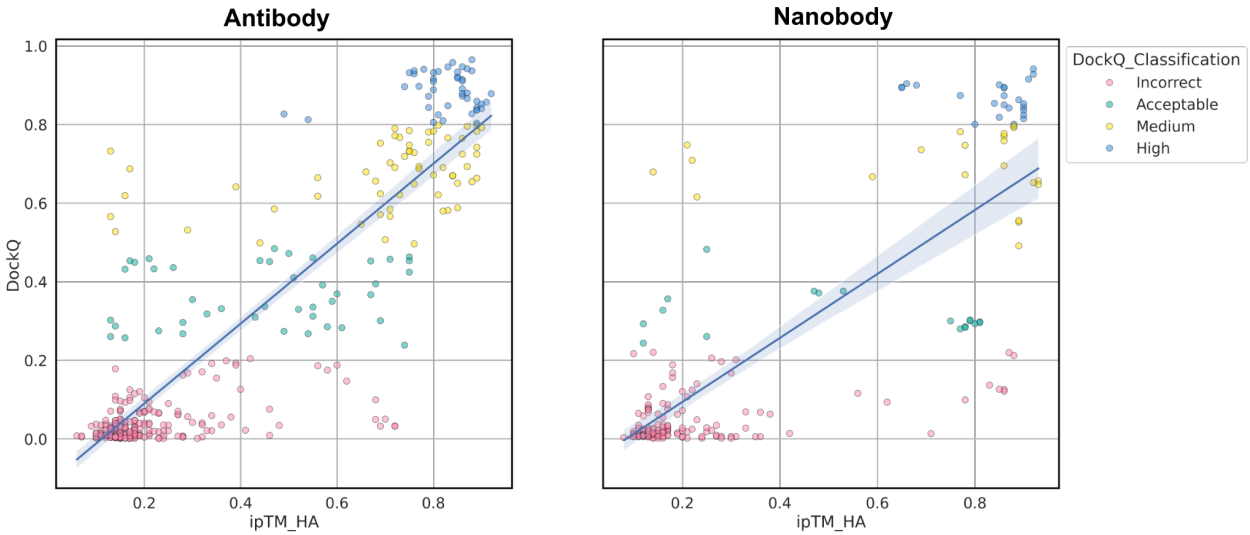

Figure 7: **Scatterplot of ipTM confidence measures against all docked antibody complexes. (A)** shows a linear correlation (R=0.91) between the DockQ scores of all evaluated antibody-antigen complexes and the predicted ipTM-HA.