A Details of Dataset

Background of Antibodies Antibodies are vital components of the immune system and are classified into various classes, including IgG, IgM, IgA, IgD, and IgE. Among them, IgG antibodies are the most abundant in the bloodstream and play a primary role in immune responses against pathogens.

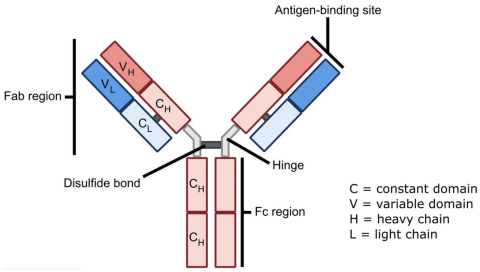

Figure 6: Structure of an IgG antibody. The heavy chain is colored orange, while the light chain is blue.

As depicted in Figure 6, IgG antibodies exhibit a Y-shaped structure composed of two identical light chains and two identical heavy chains, where heavy chains provide structural stability. Each antibody chain is further divided into distinct regions. (1) The variable regions, referred to as the variable heavy (VH) and variable light (VL) regions, are located at the tips of the Y arms. These regions contribute to the specificity of antibodies in recognizing and binding to antigens. The VH and VL regions collaborate to form the fragment antigen-binding (Fab) region. (2) At the base of the Y structure, the constant regions, also known as the fragment crystallizable (Fc) region, are important in the effector functions of antibodies. The Fc region interacts with immune cells and triggers immune responses, such as the activation of complement proteins for pathogen destruction and the promotion of phagocytosis.

Given this background knowledge of antibodies, it becomes clear that antibody-antigen docking is fundamental in immune responses, therapeutic applications, vaccine development, and drug discovery. Therefore, our study places a particular emphasis on antibody-antigen docking, contributing to this field by curating a high-quality benchmark. This dataset will serve as a valuable resource for evaluating computational models in predicting antibody-antigen interactions, ultimately facilitating the development of novel therapeutics and immunological interventions.

Antibody-antigen Benchmark The training set comprises 4,890 complexes of antibody-antigen pairs, each consisting of proteins with a minimum of 30 residues. These complexes encompass three chains, including the light and heavy chains of the antibody, along with one antigen chain. All complexes were released before January 2022. Similarly, the test set consists of 68 antibody-antigen complexes with three chains, released after October 2022. Thus, we ensure that neither baselines nor our proposed model was trained using the test set and avoid data leakage.

In practical applications, obtaining the ground truth structures of antibody-antigen complexes poses significant challenges. Researchers often turn to existing folding models to predict them. To simulate real-world scenarios, we employ a specialized antibody model called xTrimoABFold [52] to predict the conformations of antibodies and AlphaFold2 [28] for antigens. Given these predicted structures as rigid structures, we construct training and test datasets essential for further analysis and investigation. The PDB identifiers of the test set are listed here.

{8dls, 8dlr, 8dfi, 8dfh, 8dcc, 8dad, 7zr8, 7zf8, 7xxl, 7xh8, 7x26, 7wsl, 7wsi, 7ws6, 7ws2, 7wrz, 7wrv, 7wro, 7wrl, 7wrj, 7wog, 7wlc, 7wef, 7wee, 7wed, 7wcr, 7wbz, 7urq, 7uaq, 7tty, 7ttx, 7ttm, 7tpj, 7tp4, 7tp3, 7tlz, 7the, 7tc9, 7t8w, 7t7b, 7t01, 7swp, 7su1, 7str, 7sem, 7sd5, 7sbu, 7sbg, 7sbd, 7sa6, 7s5p, 7rxp, 7rxi, 7rbu, 7qtk, 7n0a, 7lo8, 7lo7, 7kql, 7fjc, 7f7e, 7f6z, 7f6y, 7eng, 7ek0, 7ejz, 7ejy, 7e9p}

B Details of Implementation

Baselines ZDOCK[1], ClusPro[2], and HDOCK[3] are user-friendly local packages suitable for automated experiments or web servers for manual submissions. We select the top-1 predicted structure from each of these methods for subsequent evaluation. For Equidock[4] and Multimer[5], we utilize their pretrained models available on GitHub for the inference. It is worth emphasizing that all methods

[1] `https://zdock.umassmed.edu`
[2] `https://cluspro.org`
[3] `http://hdock.phys.hust.edu.cn`
[4] (MIT license) `https://github.com/octavian-ganea/equidock_public`
[5] (Apache-2.0 license) `https://github.com/aqlaboratory/openfold`

except Multimer are designed for docking two chains. Therefore, during the evaluation, we employ a sequential docking strategy. This entails initially docking the light chain and heavy chain together, followed by treating them as a unified entity for docking with the antigen. And we calculate evaluation metrics using the tools USalign[6] and DockQ[7].

MSA Extraction We utilize the heuristic approach described in [23] to pair sequences from per-chain multiple sequence alignments (MSAs). Initially, the per-chain MSA sequences are grouped based on species, with the species labels obtained from UniProt's idmapping [8]. Within each specific species group, the sequences are paired together. We match the chain MSAs by minimizing the base-pair distance between the chains for prokaryotic species. While in terms of eukaryotic species, we order them based on sequence identity to the target sequence [58]. To reduce computational and memory costs, we employ the MSA clustering approach from AlphaFold2 [28]. We randomly select $N_{cluster} = 252$ sequences as the MSA cluster centers, with the primary protein sequence always set as the first cluster center. The remaining sequences are assigned to their closest cluster based on the Hamming distance.

Sequence-modal Input The sequence modality incorporates information derived from the primary sequence itself and co-evolutionary information obtained from MSAs. Following prior research [28, 23], we extract two types of features: type features $F^{typ} \in \mathbb{R}^{N_{res} \times 21}$ and primary pair features $F^{pp} \in \mathbb{R}^{N_{res} \times N_{res} \times 73}$ from the primary sequence, where N_{res} represents the number of residues. Regarding MSAs, we utilize cluster MSA features $F^{msa} \in \mathbb{R}^{N_{cls} \times N_{res} \times 49}$, where N_{cls} denotes the number of cluster centers. Specifically,

- The *type feature* $F^{typ} \in \mathbb{R}^{N_{res} \times 21}$ comprises one-hot representations of the amino acid types, encompassing the 20 known amino acids and one additional category for unknown types.
- The *primary pair feature* $F^{pp} \in \mathbb{R}^{N_{res} \times N_{res} \times 73}$ contains positional information within or across chains, including three components. (1) The *relative positional feature* of size $[N_{res}, N_{res}, 66]$ represents the relative residue indices, which are clipped between $[-32, 32]$. The 66-th index is used to indicate cross-chain pairs. (2) The *entity indicator* of size $[N_{res}, N_{res}, 1]$ identifies whether residues i and j originate from the same chain. (3) The *relative index feature* of size $[N_{res}, N_{res}, 6]$ introduces the relative *sym_id*[9] indices clipped between $[-2, 2]$. The 6-th index is assigned to pairs where the two residues have different *sym_id*s.
- The *cluster MSA feature* $F^{msa} \in \mathbb{R}^{N_{cls} \times N_{res} \times 49}$ consists of five components. (1) The *one-hot representation of the amino acid types* with size [$N_{cluster}$, N_{res}, 23], including 20 amino acids, one unknown type, one gap or missing residue, and one mask token as introduced in Section 3.1. (2) The *amino acid distribution* of size $[N_{cluster}, N_{res}, 23]$ represents the distribution of amino acid types within each MSA cluster. (3) The *deletion indicator* of size $[N_{cluster}, N_{res}, 1]$ indicates whether there is a deletion to the left of each residue. (4) The *deletion value* of size $[N_{cluster}, N_{res}, 1]$ is calculated using the formula $\frac{2}{\pi} \arctan \frac{c}{3}$, where c refers to the number of deletions to the left of each position. (5) The *mean deletion value* of size $[N_{cluster}, N_{res}, 1]$ is computed as $\frac{2}{\pi} \arctan \frac{\bar{c}}{3}$, where $\bar{c}$ represents the average number of deletions to all residues on the left of each position.

Structure-modal Input For the structure modality, we extract angle features $F^{ang} \in \mathbb{R}^{N_{res} \times 57}$ and pair features $F^{p} \in \mathbb{R}^{N_{res} \times N_{res} \times 88}$ from the rigid protein structures. These features capture important structure-modal information and are used as input for our docking model. Specifically,

- The *angle feature* $F^{ang} \in \mathbb{R}^{N_{res} \times 57}$ comprises three components. (1) The *one-hot representation of the amino acid types* with a size of $[N_{res}, 22]$, including 20 amino acids, one unknown type, and one gap or missing residue. (2) The *angle representations* of size $[N_{res}, 28]$ use sine and cosine to encode three backbone torsion angles, four side-chain torsion angles, and alternative torsion angles

[6](MIT license) `https://github.com/pylelab/USalign`

[7](GPL-3.0 license) `https://github.com/bjornwallner/DockQ`

[8]`https://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping`

[9]The *sym_id* is used to distinguish chains with the same sequence. For example, we consider a complex comprising five chains $\{A, B, B, C, C\}$, where A, B, and C represent three unique chains. The corresponding *sym_id*s for each chain would be $\{1, 1, 2, 1, 2\}$, respectively.

Table 6: Impacts of noisy structures on the docking performance of classical software and BiDock. (bold: best; underline: runner-up)

	Ground Truth			Predicted Structure		
	RMSD ↓	*TM-score* ↑	*DockQ* ↑	*RMSD* ↓	*TM-score* ↑	*DockQ* ↑
ZDOCK	11.830±5.227	0.738±0.120	0.095±0.130	12.491±6.294	0.689±0.114	0.084±0.113
ClusPro	11.486±7.993	0.780±0.133	0.204±0.256	14.135±8.153	0.702±0.118	0.118±0.192
HDOCK	**3.464**±7.394	**0.935**±0.144	**0.815**±0.364	11.328±8.073	0.742±0.167	0.314±0.390
BiDock	6.173±8.825	0.892±0.156	0.648±0.432	**7.280**±8.117	**0.847**±0.158	**0.564**±0.369

with 180° rotation symmetry for each local frame of residue. (3) The *angle indicator* with size $[N_{res}, 7]$ indicates the presence or absence of torsion angles.

- The *pair feature* $F^p \in \mathbb{R}^{N_{res} \times N_{res} \times 88}$ comprises five components. (1) The *distogram feature* of size $[N_{res}, N_{res}, 39]$ represents the discretized distances between Cβ atoms. In the case of glycine, which lacks Cβ atoms, Cα is used instead. The distances are discretized into 38 bins of equal width ranging from 3.25 to 50.75Å, with an additional bin accounting for larger distances. (2) The *residue type feature* of size $[N_{res}, N_{res}, 44]$ is derived from expanding one-hot representations of residue types with dimensions of $[N_{res}, 1, 22]$ and $[N_{res}, 22, 1]$. (3) The *backbone feature* of size $[N_{res}, N_{res}, 3]$ is obtained by constructing the unit vector of the local frame through the Gram-Schmidt process based on the original N-Cα-C coordinates. (4) The *residue indicator* with size $[N_{res}, N_{res}, 1]$ is expanded from the indicator of residue existence. (5) The *pair indicator* of size $[N_{res}, N_{res}, 1]$ indicates whether the pair is masked.

MSA Mask Policy Reflecting on Section 3.1, we design a masked MSA loss to supervise the learning of evolution representations and the integration of cross-modal information. Specifically, we randomly mask each position in an MSA cluster center with a 15% probability. Each masked token is replaced according to the following policies:

- 70% probability of substitution with a special token $\star$
- 10% probability of substitution with a randomly selected amino acid from a uniform distribution
- 10% probability of substitution with an amino acid sampled from the MSA profile that corresponds to the position
- 10% probability of no substitution

Hyperparameter Settings We initialize specific parameters of the cross-modal transformer with the checkpoint of Multimer and implement bi-level optimization using TorchOpt[10] library. The crop size is set to 412, and the batch size is set to 1. The coefficients in Equation (12) are $\lambda_1 = 0.2$, $\lambda_2 = 2.0$, and $\lambda_3 = 10.0$. For optimization, we employ the Adam optimizer with a learning rate of 10^{-4} and integrate learning rate warmup, gradually increasing the learning rate from 0 to 10^{-4} within the first 100 steps. The exponential moving average (EMA) strategy applies a decay rate of $\beta = 0.999$ and undergoes updates every 200 steps. The environment where we run experiments is:

- Operating system: Linux version 5.13.0-30-generic
- CPU information: AMD EPYC 7742 64-Core Processor
- GPU information: NVIDIA A100-SXM4-80GB

C Additional Results

Effects of Noisy Structures Classical software rely on score functions derived from statistics in the protein data bank. This dependency renders them susceptible to noise. When using folding algorithms to predict unbounded proteins, the performance of these software can degrade significantly. To validate this intuition, we conduct a docking performance analysis on the DB5.5 dataset using ground truth and predicted structures from folding models as unbounded structures, respectively. As

[10](Apache-2.0 license) `https://github.com/metaopt/torchopt`

shown in Table 6, these results illustrate that although HDOCK performs exceptionally well with ground truth, minor noise in predicted structures leads to a substantial decline in its performance. On the contrary, BiDock consistently generates acceptable predictions regardless of the input type, showcasing its robustness to noise. In real-world applications, reliance on the availability of ground truth structures is impractical. The ability of BiDock to maintain high prediction quality when confronted with noisy structures makes it an invaluable tool.