A Systematic Evaluation Framework for Universal Antibody–Antigen Binding Affinity Prediction and Candidate Recommendation

Yunrui Li[‡2*], Yue Zhao[1], Kemal Sonmez[1], Luca Giancardo[1], Lan Guo[1], Pengyu Hong[2], Nina Cheng[1], Melih Yilmaz[1*]

[1]Amazon Web Services, Seattle, WA, USA.
[2]Computer Science Department, Brandeis University, Waltham, MA, USA.
[‡]Work done during internship at Amazon Web Services.

*Corresponding author(s). E-mail(s): yunruili@brandeis.edu; melihyz@amazon.com;
Contributing authors: yuezhaom@amazon.com; ksonmez@amazon.com; lugianca@amazon.com; languo@amazon.com; hongpeng@brandeis.edu; cxinyun@amazon.com;

Abstract

Predicting antigen–antibody binding is essential to drug discovery and protein engineering. For *de novo* antibody design, generalizable binding prediction models are crucial for efficient *in silico* screening. However, existing affinity predictors lack generalization, with performance deteriorating for antibodies targeting antigens absent from training data or datasets lacking non-binders. To address this, we establish a benchmarking framework for evaluating universal antibody–antigen binding affinity prediction. Our framework compares sequence- and structure-based methods across diverse antigens, introducing standardized evaluation protocols based on pairwise accuracy and retrieval metrics. We propose **MochiBind**, a sequence-only pairwise binding affinity predictor, and benchmark it against structure-derived baselines such as Boltz-2, GeoDock, and Graphinity. Results show **MochiBind** achieves comparable or superior performance in pairwise accuracy and retrieval, suggesting sequence-based approaches can match or surpass structure-based models in generalization. The proposed benchmark provides a foundation for fair comparison and future development, enabling scalable, sequence-driven solutions to binding affinity prediction.

1 Introduction

Monoclonal antibodies (mAbs) have become indispensable therapeutics, and their discovery pipeline has undergone a profound transformation. The classical approach to antibody generation, pioneered by the hybridoma technology[1, 2], relies entirely on the natural immune system. This foundational method involves immunizing a host animal, typically a mouse, and fusing its antibody-producing B cells with immortal myeloma cells to create a hybridoma. This process yields fully affinity-matured binders but necessitates subsequent complex and costly humanization to reduce immunogenicity in human patients. Following hybridomas, techniques like phage display[3–7] emerged, offering an in vitro alternative that still relies on selecting binders from large, naturally-derived or randomized libraries. Both the hybridoma and display methods fundamentally operate on the principle of selection from a vast pool of potential binders.

In contrast, the *de novo* antibody design pipeline[8–10] represents a critical paradigm shift, moving from immune-driven discovery to rational engineering. *De novo* design aims to create functional antibodies or antibody-like scaffolds without relying on an in vivo immune response[11, 12]. This approach leverages structural biology and computational methods to overcome the inherent limitations of classical methods, such as the requirement for animal immunization, the difficulty in targeting non-immunogenic epitopes, and the need for downstream humanization[13, 14]. The core goal is to computationally predict and engineer molecules with optimized properties, including enhanced stability, reduced immunogenicity, and high specificity for challenging targets, right from the initial design phase. This modern methodology is primarily divided into two sophisticated approaches: Structure-Based methods[15] and Sequence-Based modeling[8, 16]. Structure-based antibody design uses computational techniques that incorporate three-dimensional information about antibodies and antigens, allowing models to reason about spatial complementarity and interaction patterns that support effective binding. Sequence-based design complements this by working directly with amino acid sequences, leveraging statistical, evolutionary, and machine-learning methods trained on large antibody repertoires to identify or generate promising candidates with desirable biophysical and functional properties.

Antigen–antibody interactions lie at the core of immune recognition and therapeutic antibody design. Binding affinity serves as a fundamental criterion for evaluating antibody quality and guiding subsequent engineering cycles[17]. However, experimental affinity measurement remains laborious and costly[18–20]. As a result, computational prediction has emerged as a scalable first-pass screening strategy that complements and accelerates experimental discovery[21–24]. The progression from classical, immune-driven selection to rational, in silico *de novo* design has amplified the necessity of accurate predictive modeling.

Despite significant progress in computational approaches, most existing models focus on in-sample prediction, which performs well on antibody and antigen types similar to those in the training sets, while their ability to generalize to unseen antigen–antibody complexes remains limited. For practical antibody discovery, the capability to make accurate out-of-sample predictions is crucial, as it eliminates the

need for antigen or antibody framework-specific retraining and enables universal antibody design evaluation. More importantly, no standardized benchmark exists for assessing generalization in antibody–antigen binding prediction. Studies in this domain differ widely in dataset curation, data splitting strategies, and evaluation metrics (see Table 1), making it difficult to compare models or quantify methodological progress.

In this work, we introduce a unified and systematic framework for evaluating the antibody–antigen binding affinity prediction in a large pool of antibody candidates (Figure 1), given an antigen. First, the framework assesses both sequence-based and structure-based methods with pairwise learning, where the model is asked to compare two given antibodies of the same antigen and determine which antibody offers higher binding affinity. We propose a sequence-only model, **M**odels for **O**ptimized **C**omputational **H**andling of **I**mmunoglobulins - **Bind** (**MochiBind**), shown in Figure 1B, and systematically compare it against structure-derived baselines including Boltz-2[25], GeoDock[26], and Graphinity[27]. The **MochiBind** approach encodes antibody and antigen sequences using a pretrained protein language model, projects them into a shared latent space, and predicts relative binding strength through a lightweight pairwise scoring module. Second, to better reflect real-world antibody discovery pipelines, we use the pairwise prediction to derive the global ranking within the candidate pool (Figure 1C). Finally, we design new evaluation metrics using two retrieval-based strategies (accuracy and precision), which measure a model's ability to prioritize high-affinity candidates from a large pool of antibody candidates, thereby capturing practical performance in antibody design campaigns faithfully (Figure 1D). We used the AlphaBind[28] dataset, which contains four antigen systems (PP489, PEM, VHH72, and TRA), each with about 30,000 experimentally characterized antibody variants spanning a wide affinity range. The dataset provides antigen-level and antibody-level diversity. First, the antigens exhibit low sequence similarity to one another. Pairwise antigen similarity was quantified using Needleman-Wunsch global alignment with BLOSUM62 substitution matrix, with the gap opening penalty set to -10 and gap extension penalty set to -0.5. These are standard setups for algorithms (e.g. BLAST) and many protein alignment tools. Scores were normalized by the maximum self-alignment score to account for sequence length differences. The off-diagonal scores among the four antigen systems were all close to zero or negative, ranging from -0.053 to 0.034, indicating very low global sequence similarity between these antigens. Second, each antigen system consists of a large pool of antibody variants, generated from a distinct seed antibody framework through diverse multi-point mutations that vary in number and location across candidates. This setup compels models to learn true binding determinants rather than memorizing antigen or antibody identities, enabling rigorous assessment of model generalization and robustness.

2 Related Work

Computational prediction of antibody–antigen binding affinity has attracted growing attention, with diverse modeling paradigms proposed across sequence-[29], structure-[25–27, 30], and generative-based approaches[31]. These methods differ in their reliance on structural information, the diversity of their training data, and the evaluation

Table 1: Common evaluation and dataset issues in prior antibody–antigen studies. A universal predictor should avoid these pitfalls to enable out-of-sample discrimination on unseen complexes without case-specific retraining.

Related Work	Non-binders included[a]	Large pool of antibodies[b]	Out-of-sample test set[c]	Large test set[d]	Thorough binding-affinity evaluation[e]
Yuan et.al[29]	×	×	✓	✓	×
Shrestha et.al[30]	✓	×	✓	×	×
Hummer et.al[27]	✓	✓	✓	✓	×
Chen et.al[31]	×	×	✓	×	×
Passaro et.al[25]	✓	✓	×	✓	×
Chu et.al[26]	✓	✓	×	✓	×
MochiBind (ours)	✓	✓	✓	✓	✓

Notes: (a) The dataset consists primarily of strong and weak binders; non-binders are not included or under-represented. (b) For every antigen, more than 100 antibody variants are available, ensuring adequate sampling for evaluating antigen-specific binding affinity. (c) The test set is independent of the training and evaluation dataset. (d) Test set contains more than 100 samples. (e) Binding affinity is predicted and evaluated directly using a broad suite of metrics, such as accuracy, ranking, and retrieval, etc.

strategies. Below we summarize representative studies and discuss their key limitations in terms of generalization, data coverage, and applicability to real-life antibody discovery, especially for *in silico* screening.

Some recent efforts focus on models that directly predict binding affinity from sequence or structure information. DG-Affinity [29] introduced a sequence-based framework that employs pretrained language model embeddings as input to a ConvNeXt-style network. The model achieves strong performance (Pearson correlation >0.65) on independent test sets without requiring structural data. NanoBinder [30] similarly aimed to predict nanobody–antigen binding using Rosetta-derived energy features with a Random Forest classifier trained on experimentally validated complexes. While both studies demonstrate promising accuracy, they share several limitations. First, their evaluation datasets are small (26 complexes for DG-Affinity and 49 complexes for NanoBinder), restricting assessment of cross-antigen generalization. Additionally, DG-Affinity data lack non-binders, while the NanoBinder dataset uses synthetically generated negative samples, making them more easily distinguishable from weak or strong binders. Moreover, antigens in these datasets are associated with highly uneven numbers of antibodies: some only have a handful, whereas others may have thousands. Because datasets are not curated to produce comparable distributions of antigen–antibody complexes across antigens, evaluating a model's capability by antigen system becomes challenging. These limitations underscore the need for antigen-diverse datasets with large antibody pools and consistent evaluation protocols.

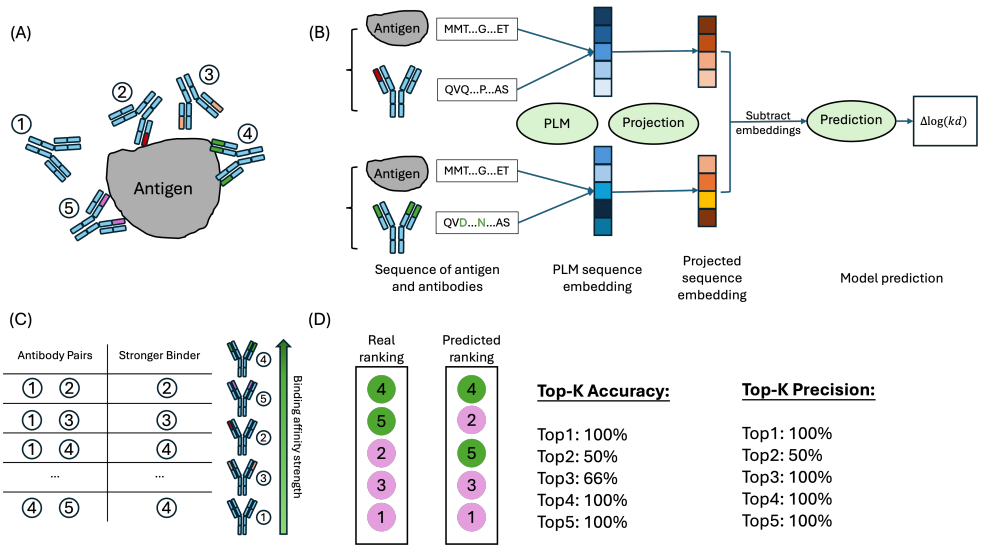


Fig. 1: MochiBind model architecture and evaluation framework. (A) We used the AlphaBind dataset that contains a large pool of antibodies, per antigen. The antibodies have different multi-point mutations and binding affinities. (B) The architecture of MochiBind, which processes a pair of antibody sequences with the same antigen sequence through a pre-trained protein language model (PLM) to obtain embeddings. These embeddings are projected into a shared latent space, their difference is computed, and the resulting representation is used to predict relative binding affinity strength. (C) From the pairwise prediction, the stronger binder is determined for each pair. This information is used to derive a global ranking of the binding affinity strength within the candidate pool. (D) Using the global ranking, the model is able to recommend the top candidates for further investigation. To evaluate the retrieval performance, we propose two metrics in our evaluation framework: retrieval accuracy and retrieval precision. Retrieval accuracy calculates the percentage of antibodies from the K highest-ranked candidates that are among the true top-binding antibodies, while retrieval precision measures the percentage of good binders captured within the model's recommendations, relative to the total number of good binders possible in this group. See Section 4.4 for a detailed description.

As generalization across antigens has become a well-known challenge in the field, some work has begun to address this challenge in antibody–antigen binding prediction. Graphinity [27] represents one of such efforts, employing an equivariant graph neural network to predict $\Delta\Delta G$ (changes in binding free energy) after mutation using structural inputs. In particular, this work highlights the importance of data diversity and data volume needed for robust generalization. The model reports high correlations (up to $\sim$0.85) on large-scale simulated datasets. However, its reliance on simulated single-point mutations limits applicability to experimental or multi-mutation settings, and the study does not evaluate the model's capability for discriminating good binders

from weak ones in a large pool of antibodies, which is critical for practical candidate ranking.

Some studies use binding affinity prediction as a surrogate objective for other tasks, such as sequence optimization or generative design. AffinityFlow [31] applies a flow-based generative modeling approach to guide antibody sequence generation toward higher affinity in latent space. In this setup, the method does not provide direct affinity prediction or quantitative validation against experimental data.

Finally, computational antibody design has long relied on physics-based models to estimate antigen–antibody binding affinity. Molecular Dynamics (MD) simulations provide atomistic insight into conformational flexibility[32], while post-processing approaches offer tractable approximations of binding free energy from MD trajectories[33–35]. More rigorous alchemical methods, including Free Energy Perturbation (FEP)[36], Thermodynamic Integration (TI)[37, 38], and Bennett Acceptance Ratio (BAR)[39, 40], provide higher-fidelity estimates of ΔG but remain computationally expensive for large-scale antibody screening. Faster alternatives, such as protein–protein docking[41] and Rosetta's interface energy functions[42], supply physics-inspired scoring metrics, though often with limited treatment of backbone rearrangement and entropic contributions. Complementary to these physics-based methods, several recent structure prediction and docking frameworks generate confidence or stability scores that are frequently interpreted as surrogates for binding affinity. Models such as Boltz-2 [25] and GeoDock [26] exemplify this trend, producing interface plausibility metrics (e.g., ipTM, ipLDDT) that correlate with structural correctness but do not directly reflect thermodynamic interaction strength. Moreover, these systems are typically trained and evaluated on datasets containing only strong or weak binders, without explicit inclusion of non-binders, limiting their ability to support reliable binder ranking in realistic discovery settings, where the vast majority of antibody candidates are non-binders, especially in *de novo in silico* campaigns.

Taken together, these studies illustrate the field's rapid progress yet also highlight critical gaps in standardized evaluation of model performance and generalization power. Existing methods vary widely in dataset composition, data-splitting schemes, and target definitions, making cross-model comparison difficult. Our proposed framework addresses these issues by establishing a unified evaluation protocol that systematically compares both sequence- and structure-based models across diverse antigens under retrieval-oriented settings that resembles real antibody discovery campaigns.

3 Results

3.1 Retrieval Performance

A key goal in antibody–antigen modeling, especially for *in silico* design, is to prioritize high-affinity candidates from large sequence libraries before committing to costly and time-consuming experimental screening. To better mirror the antibody discovery workflow, we introduce a retrieval-based evaluation metric designed to assess a model's practical ranking capability. In contrast to absolute regression metrics such as mean absolute error (MAE) or correlation, retrieval focuses on the relative ordering of candidates based on predicted affinity. This perspective provides a more robust

and application-relevant measure of model utility, particularly under conditions where affinity measurements are noisy or heterogeneous across assays. Retrieval performance thus serves as a direct indicator of how effectively a model can guide candidate recommendation in antibody discovery pipelines.

We evaluate retrieval performance using the AlphaBind dataset [28], in which each unique antigen is associated with 5,000–10,000 experimentally characterized antibody variants (see Section 4.1 for data and preparation details). To train and evaluate in a consistent and scalable setting, we formulate the task as a pairwise ranking problem. We sample 200,000 pairs across the antibody pool for each antigen (see Section 4.1.1 for details). The training, evaluation and test datasets cover distinct antigen systems, ensuring no overlap between data splits. Specifically, we use 2 antigen dataset for training, 1 for validation and 1 for testing, resulting in 400,000 pairs of antibodies sampled for training, 200,000 for validation, and 200,000 for testing. While the sampled pairs do not cover all possible antibody combinations, they are sufficient to capture the distribution of candidate diversity and enable efficient model training. For ranking, we applied the TrueSkill[43] algorithm (see Section 4.4 for details) to infer global rankings across all candidates, allowing the identification of top binders within each antigen pool.

We evaluated two retrieval strategy: Top-K retrieval precision, and Top-K retrieval accuracy (see Section 4.4 for details), to assess the model's capability in recommending the most promising candidates from the large antibody pool. Across all retrieval-based evaluations shown in Figure 2 and Figure 3, **MochiBind** consistently outperforms the structure-based Graphinity model and proxy models (GeoDock and Boltz-2), showcasing the capability of prioritizing the strongest binders from large candidate pools. The Graphinity model is evaluated after finetuning on the training dataset, since the original model was trained using single-point mutation data and did not perform well on the AlphaBind dataset (more details in Section 4.5).

While MochiBind leads in both ranking accuracy and retrieval precision overall, we observe that retrieval performance varies considerably across different antigens, particularly in the precision-based evaluations shown in Figure 2. This variability underscores the importance of assessing model performance across a diverse set of antigens and large antibody pools that include both strong binders and non-binders. As illustrated in Figure 5, the binding affinity distributions differ markedly among antigens. For example, some exhibit bimodal patterns, such as PEM and TRA, whereas others, such as VHH72, follow a right-skewed normal distribution. Notably, the TRA and PEM datasets contain a high proportion of non-binders, closely mirroring real-world *in silico* discovery scenarios, whereas VHH72 includes nearly 40% strong binders.

The retrieval-precision metric effectively captures a model's ability to identify true binding signals within these heterogeneous candidate pools. For antigens such as PEM and TRA, which follow bimodal distributions, precision tends to increase with K, as additional true positives are gradually recovered when expanding the candidate set. In contrast, for antigens with a higher proportion of binding-capable variants (e.g., VHH72), precision remains relatively stable or may decrease slightly, as strong binders are already well represented among the top-ranked candidates and additional candidates introduce more variability. The performance curve of random guess shows the

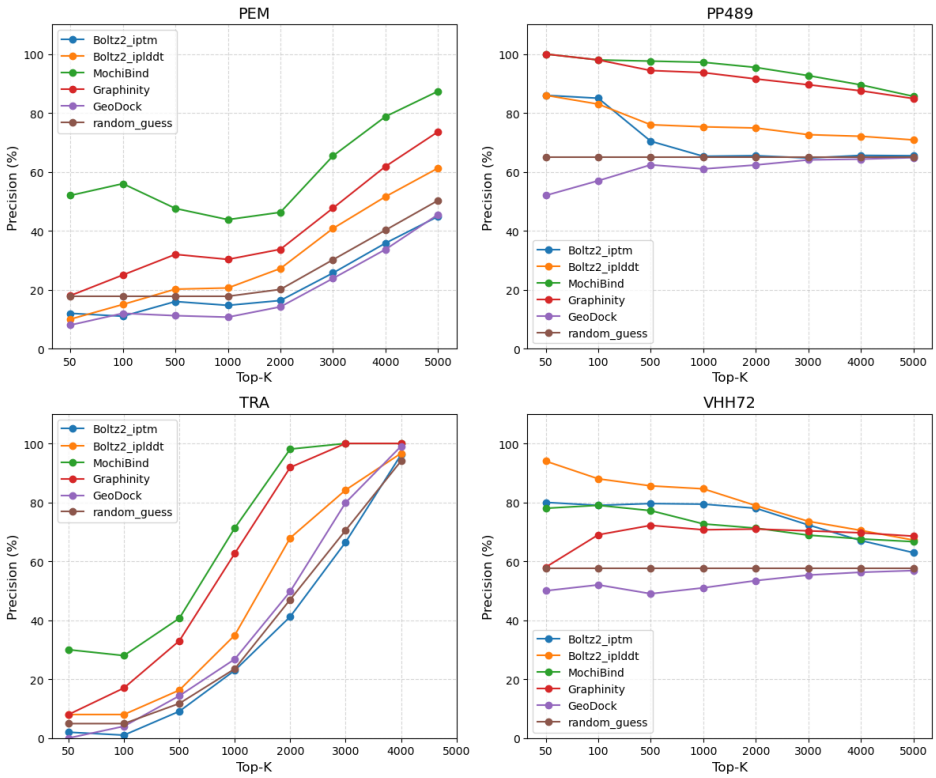


Fig. 2: Retrieval precision comparison for different antigens. Retrieval Precision evaluates the purity of the retrieved candidates rather than their completeness. It is defined as the proportion of correctly identified strong binders among the top-K predictions, given a pre-determined threshold. In this plot, an antibody with $log(Kd) < 2$, with Kd in nanomolars (nM), is deemed a good binder. For most of the antigens, our proposed MochiBind model (green line) achieves the highest retrieval precision, followed by the finetuned Graphinity model (red line). The brown line indicates the random guess precision, and is closely followed by the structure confidence score produced by Boltz-2 and GeoDock models. In the dataset of VHH72, a camelid single-chain antibody targeting the RBD of SARS-CoV1 with cross-reactivity to SARS-CoV2, the majority of the antibody mutations are good binders (see Section 4.1 for more details), thus, it is a less predominant case in the real-world antibody-antigen design task.

null hypothesis under such distributions, and the performance of MochiBind follows the same trend and is consistently better. These differences highlight that retrieval precision captures not only ranking quality but also the difficulty and composition of each antigen-specific dataset. As such, Precision@K provides a meaningful and practical evaluation metric for candidate selection, particularly in settings where the distribution of strong binders varies across tasks.

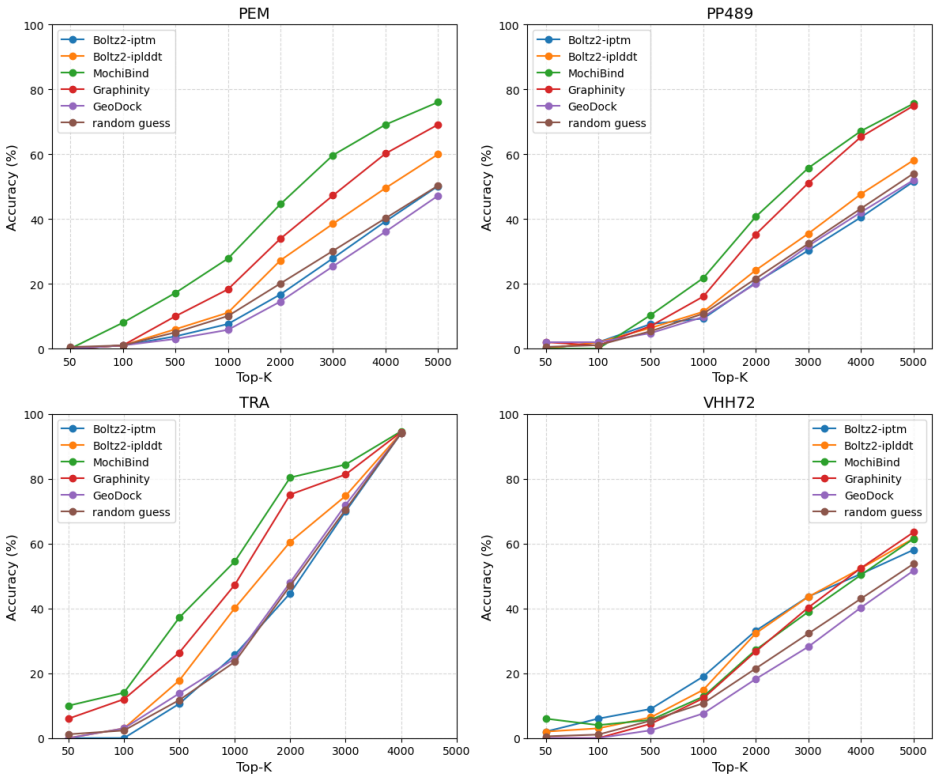


Fig. 3: Retrieval accuracy comparison for different antigens. Retrieval accuracy measures how well a model ranks antibody candidates according to their true binding affinities. For all antigens (see Section 4 for details), our proposed MochiBind model (green line) achieves the highest retrieval accuracy, followed by the finetuned Graphinity model (red line). The brown line indicates the random guess accuracy, and is closely followed by the structure confidence score produced by Boltz-2 and GeoDock models. This indicates that the structural confidence score may not be suitable as a binding affinity proxy for new antigens, or when weak binders exist.

3.2 Pair-wise Accuracy

To complement retrieval-based evaluation, we also assess pairwise prediction accuracy, which quantifies a model's ability to correctly identify the stronger binder within sampled pairs. As shown in Table 2, MochiBind achieves higher accuracy than all baseline models, across all datasets, confirming the potential of a simple sequence-based approach, such as MochiBind, can reliably capture affinity ordering, which provides a promising direction for binding affinity modeling for out-of-sample antigen systems. To validate the significance of model performance differences, we used one-sided McNemar's chi-squared test to compare the MochiBind model with the baseline models, across the four antigen datasets. Across all four datasets, McNemar's chi-squared test rejects the null hypothesis that MochiBind and each baseline model (GeoDock,

Boltz-2, and Graphinity) have equal marginal probabilities of accuracy, based on the 200,000 pairs from the test set (all p-values < 0.0001). This indicates that the observed improvement achieved by MochiBind is statistically significant. The detailed McNemar's statistical test results with chi-squared values are reported in Appendix A. The Graphinity model result is reported using the finetuned model. The accuracy performance before and after finetuning is shown in Appendix B.

Because experimental measurements of binding affinity are often subject to noise and variability, some antibody pairs may exhibit only marginal differences in their recorded affinity values. In such cases, relative rankings between two antibodies can vary across experiments: for example, one experiment may report antibody A as the stronger binder, whereas another may favor antibody B. To better understand the effect of such measurement uncertainty on MochiBind's pairwise prediction accuracy, we stratified the 200,000 test pairs into groups based on the relative percent difference in their measured binding affinities, as shown in Figure 4. We observe that the model's accuracy increases consistently as the affinity gap between pairs widens, indicating that MochiBind can confidently and reliably distinguish between antibodies with large true affinity differences. This trend suggests that prediction errors primarily occur in ambiguous regions where the relative binding strengths are within the margin of error. Despite this general trend, the relative accuracy across antigens differ, potentially due to intrinsic antigen properties. For example, TRA maintains higher accuracy across most intervals, even though it contains the largest proportion of pairs with small affinity differences (0–10%). This is because TRA contains more weak binders (shown in Figure 5) with relatively large $log(Kd)$ values. Meanwhile, datasets such as VHH72 and PEM exhibit broader distributions of pairwise affinity gaps. These results suggest that both the composition of affinity differences and other dataset-specific characteristics jointly influence the observed performance patterns.

While pairwise accuracy provides a clear measure of local ranking consistency, retrieval precision offers a more realistic assessment of model performance in discovery settings, as it captures whether truly strong binders can be effectively prioritized from among thousands of candidates. Taken together, these analyses demonstrate that MochiBind is able to capture binding signals in out-of-distribution antigen systems, and provides a robust and practically useful framework for antibody–antigen binding prediction and candidate recommendation.

4 Methods

4.1 Data and Data Preparation

We use the AlphaBind dataset[28] for training and evaluation due to its diversity of antigens and the large number of antibody variants associated with each antigen. This dataset contains four unique antigens, each with thousands of experimentally characterized antibody candidates covering a wide affinity range: (1) AAB-PP489 (PP489), a previously undescribed humanoid scFv targeting human TIGIT; (2) Pembrolizumab-scFv (PEM), a humanized mouse antibody formatted as an scFv targeting PD-1; (3) VHH72, a camelid single-chain antibody targeting the RBD of SARS-CoV-1 with cross-reactivity to SARS-CoV-2; and (4) Trastuzumab-scFv (TRA), a humanized

Table 2: Pairwise accuracy (%) comparison of baselines and our proposed MochiBind approach on AlphaBind datasets. p indicates significance levels for one-sided McNemar's χ^2 tests comparing the MochiBind model against other baseline methods (Graphinity, Boltz-2, and GeoDock). Values denote χ^2 statistics with significance levels in parentheses. For all datasets, our model outperforms all baselines.

Antigen	Graphinity (%)	Boltz-2 - ipIddt (%)	GeoDock (%)	MochiBind (%)
PEM	63.20%	57.93%	47.11%	72.35% ($p < 0.0001$)
PP489	67.01%	55.52%	50.17%	70.29% ($p < 0.0001$)
TRA	68.55%	60.09%	51.21%	74.27% ($p < 0.0001$)
VHH72	58.63%	57.90%	49.58%	59.89% ($p < 0.0001$)

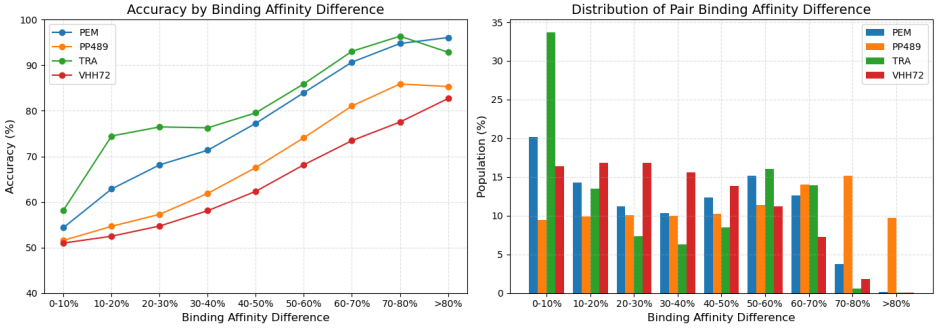


Fig. 4: MochiBind model accuracy (left) and data distribution (right) by threshold. Left: While the proposed MochiBind model achieves the highest accuracy in the pairwise binding affinity discrimination task, the accuracy continues to improve as the binding affinity difference becomes wider among the two complexes. Right: In the test dataset, a significant amount of the pairs are within 0-10% difference in binding affinity, a category that does not exist in the training and validation dataset due to data filtering (see Section 4.1 for more details). This signifies the challenges in antibody design and recommendation, as it is often difficult to discriminate binders with similar binding affinity, especially between weak binders with large $log(Kd)$ values.

mouse antibody targeting human HER2. Each antigen system contains approximately 30 000 antibody variants generated and validated using AlphaSeq.

Beyond providing high-quality experimental data, the AlphaBind study also introduced a sequence-based pre-training and fine-tuning framework that enables rapid adaptation of a general antibody–antigen model to new datasets. In the original work, PP489, PEM, and VHH72 were used for both fine-tuning and evaluation, whereas TRA served exclusively as an alternative fine-tuning dataset to test model generalization. For our benchmarking experiments, we compare MochiBind with structure-based baselines such as Graphinity, GeoDock, and Boltz, which require structural coordinates. To reduce computational cost while maintaining data diversity, we applied stratified

sampling to select 5,000 antibodies per antigen following the same binding-affinity distribution as the full AlphaBind dataset. This sampling strategy follows the experimental design of the original AlphaBind study, in which PP489, PEM, and VHH72 are used for both finetuning and evaluation, while TRA is treated as a held-out antigen system for assessing cross-antigen generalization. As a result, for PP489, PEM, and VHH72, we sample 5,000 variants from the finetuning set and 5,000 from the evaluation set, yielding 10,000 antibody variants per antigen. In contrast, 5,000 variants are sampled from the TRA dataset. Because our model employs a pairwise prediction setup, the subset is able to generate more pairs than needed (see Section 4.1.1 for details), and this sampling does not compromise data diversity.

The binding affinity distributions across these antigens differ substantially, as shown in Figure 5. The PEM and TRA datasets display bimodal distributions, likely reflecting clear distinctions between strong and weak binders, while VHH72 follows a right-skewed unimodal distribution, and PP489 exhibits an intermediate pattern. The number of strong binders and non-binders also varies across antigens: PP489 contains the highest proportion of very strong and strong binders, whereas TRA consists predominantly of weak or non-binding variants. These distinct affinity profiles emphasize the importance of evaluating models across diverse antigens and large candidate pools to assess generalization under realistic experimental variability.

4.1.1 Data Pair Sampling Strategy

Because MochiBind operates on pairwise comparisons, naive construction of all possible antibody pairs would be computationally prohibitive. For example, 5,000 antibody variants yield approximately 25 million pairs per antigen. In addition, many antibody pairs exhibit marginal affinity differences that may vary within experimental uncertainty, potentially biasing the model toward trivial comparisons. To address both challenges, we perform stratified random sampling to generate a representative yet tractable training set. Specifically, we sample 200,000 pairs per antigen for training, restricting selected pairs to those whose relative affinity difference exceeds 10%. This threshold ensures that the selected pairs reflect meaningful distinctions in binding strength, promoting efficient learning of robust and interpretable signals.

For the test datasets, no such thresholding is applied, allowing evaluation under realistic experimental conditions that include small and noisy affinity differences. This distinction between training and testing distributions ensures that the model is exposed to meaningful supervision during training while being evaluated under authentic, noise-prone conditions encountered in practice.

4.1.2 Cross-Antigen Evaluation Setup

To rigorously assess the models' performance on out-of-distribution systems, we employ a cross-antigen evaluation scheme. In each experimental configuration, two antigens are used for training (400,000 pairs), one for validation (200,000 pairs), and one for testing (200,000 pairs without threshold). For example, when the model is trained on PEM and PP489, the VHH72 dataset is used for validation to select the best-performing checkpoint, and the model is subsequently evaluated on TRA. This process is repeated so that each antigen serves as the held-out test set once. Such

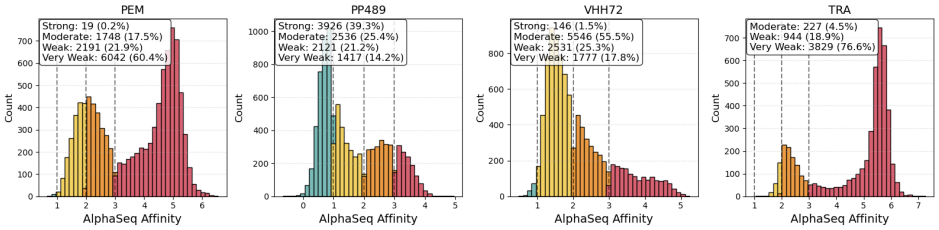


Fig. 5: Binding affinity (measured in $log_{10}(K_d)$) distribution across different antigens. The distributions of experimentally measured binding affinities are shown for each antigen, with strong, moderate, weak, and very weak binders colored in blue, yellow, orange, and red, respectively. The results reveal substantial heterogeneity among antigens: TRA contains no strong binders, PEM has only a few, while PP489 includes nearly 40% strong binders, and VHH72 is dominated by moderate binders. Most antigens (PEM, PP489, TRA) exhibit bimodal distributions, whereas VHH72 shows a right-skewed profile. These pronounced differences highlight the challenge of developing a universal affinity predictor that can robustly identify good binders despite varying underlying distributions across antigens.

a rotation-based setup ensures that the model's generalization capability is measured on antigens that were completely unseen during training, providing a fair and comprehensive evaluation of universal antibody–antigen binding prediction.

4.2 Modeling Framework

Computational approaches for antibody–antigen binding prediction can be broadly categorized into two modeling paradigms: pairwise prediction and complex-wise prediction. In the pairwise formulation, the model receives two antibody candidates targeting the same antigen and predicts which one exhibits higher binding affinity. This setting is particularly useful for ranking-based tasks, where relative comparison is more informative than absolute affinity estimation. Both Graphinity[27] and our proposed MochiBind adopt this formulation.

In contrast, complex-wise models directly estimate the binding affinity or interface stability of a single antibody–antigen complex. Such methods typically rely on 3D structural inputs, either experimentally determined or predicted using protein structure prediction tools. Representative examples include Boltz-2, which produces interface confidence metrics (*ipTM*, *ipLDDT*) as proxies for binding strength, and GeoDock, a geometric deep-learning docking framework that estimates interface stability using *ipLDDT*-derived features. These complex-wise models are widely used in structure-based affinity estimation but often show limited generalization when applied to novel antigens without reliable structural templates.

Experimental measurements of binding affinity are inherently noisy, and even replicate assays can yield slightly different quantitative values for the same antibody–antigen complex. Such variability makes the absolute affinity labels of individual complexes uncertain, especially when differences between binders are small. In contrast, pairwise modeling mitigates the impact of this label noise by focusing on the

relative strength between two antibodies targeting the same antigen. Since relative ordering tends to remain more consistent across experiments than absolute affinity measurements, pairwise formulations provide a more robust and noise-tolerant learning objective. This makes them well suited for integrating heterogeneous datasets collected under varying experimental conditions.

4.3 Design of MochiBind

To develop a scalable and generalizable alternative to structure-dependent approaches, we designed MochiBind (see Figure 1(B)), a purely sequence-based pairwise model for antibody–antigen binding prediction. The model takes as input the amino acid sequences of the antibody and antigen, which are embedded using pretrained protein language models. The result displayed uses ESM-2[44] (esm2_t36_3B_UR50D), which used an embedding dimension of 2560 to represent each residue of antigen and antibody sequences. The antigen and antibody embeddings are subsequently concatenated, resulting in a combined representation of dimension 5120. For each sequence, mean-pooling is applied over residue embeddings to obtain fixed-length representations ($d = 5120$) that capture global sequence-level context.

These representations are projected into a shared latent space through a learnable projection layer, after which a multilayer perceptron (MLP) head predicts the relative binding strength between antibody pairs. We set the dimensions of the projection layer to 1024 and the MLP layers to $[256, 128, 64]$. During training, MochiBind is optimized using a Mean Squared Error (MSE) loss, encouraging the model to not only assign correct signs to antibodies with experimentally measured stronger binding affinities, but also measure the degree of difference accurately. Learning rate is set to 0.001, and batch size is set to 32. Dropout is set to 0.3 to prevent overfitting, the Adam optimizer is used, and the training stops if the holdout loss does not decrease after 5 epochs. This sequence-only formulation eliminates dependence on structural models and allows evaluation across large and diverse antigen–antibody pools. At inference time, for 200,000 pairs derived from 5,000 antibodies, the MochiBind model requires approximately 13 seconds for scoring on CPU, while the TrueSkill algorithm takes about 76 seconds to produce the ranking.

4.4 Retrieval Strategy

4.4.1 Ranking Algorithms

Since the model is trained in a pairwise setting, its raw outputs correspond to predicted preferences between two antibody candidates for the same antigen. To convert these pairwise comparisons into a global ranking over all candidates, we adopt TrueSkill Bayesian ranking algorithm [43] as it integrates uncertainty and also support live ranking.

The TrueSkill[43] method was originally developed for multiplayer game rating. Each antibody candidate a_i is assigned a latent "skill" parameter μ_i with uncertainty σ_i. Given pairwise outcomes (a_i, a_j) where the model predicts that a_i binds stronger

than a_j, TrueSkill updates posterior distributions as:

$$\mu_i' = \mu_i + \frac{\sigma_i^2}{c} v\left(\frac{\mu_i - \mu_j}{c}\right), \quad \mu_j' = \mu_j - \frac{\sigma_j^2}{c} v\left(\frac{\mu_i - \mu_j}{c}\right),$$

where $c = \sqrt{2(\sigma_i^2 + \sigma_j^2)}$ and $v(\cdot)$ is a scaling function derived from the standard normal PDF/CDF ratio. After iterating over all observed pairwise comparisons, candidates are ranked according to their posterior means μ_i. This method integrates uncertainty and transitive consistency, producing a probabilistically coherent ranking across the full candidate pool. Our implementation follows the default configuration of the TrueSkill package: we initialize $\mu = 25$ and $\sigma = 25/3$. Pairwise comparison outcomes are then processed sequentially using the `rate_1vs1` update rule, which updates the posterior (μ, σ) for each candidate.

4.4.2 Top-K Retrieval Accuracy

The Top-K retrieval accuracy quantifies the fraction of experimentally verified high-affinity binders that appear among the top-K predictions made by the model. Formally, let $\mathcal{A} = \{a_1, a_2, \ldots, a_N\}$ denote the set of antibody candidates targeting a specific antigen. Each antibody a_i has an associated global ranking score $s(a_i)$, derived from the TrueSkill algorithm, and an experimentally measured binding strength $y(a_i)$ (e.g., $\log(K_D)$, where lower values indicate stronger binding). The predicted ranking list $\mathcal{R}$ is obtained by sorting all antibodies in descending order of $s(a_i)$, while the ground-truth ranking $\mathcal{R}_{\text{true}}$ is obtained by sorting in ascending order of $y(a_i)$. The Top-K retrieval accuracy is then defined as:

$$\text{Accuracy@}K = \frac{|\mathcal{R}_{1:K} \cap \mathcal{R}_{\text{true},1:K}|}{K}.$$

This metric measures the overlap between the model's top-K predicted candidates and the true top-K experimental binders. A higher value indicates stronger agreement between model predictions and experimental results, implying that the model effectively ranks the most promising antibodies at the top of its recommendation list.

4.4.3 Top-K Retrieval Precision

Complementary to retrieval accuracy, the Top-K retrieval precision quantifies the reliability of a model's top-ranked predictions by measuring how many of the predicted top-K antibodies are experimentally confirmed as strong binders. This metric reflects the model's ability to minimize false positives, which is an important property during the early stages of antibody discovery, when only a limited number of candidates can be experimentally tested. In practice, strong-binding antibodies often share similar sequence motifs or local structural features; therefore, a model's capacity to identify these binding signals from a large and diverse candidate pool can provide valuable insights into effective antibody design. Such predictive capability enables researchers to iteratively refine antibody libraries toward stronger binders based on computational screening outcomes.

Formally, let $\mathcal{B} = \{a_i \mid y(a_i) < \tau\}$ represent the set of experimentally verified strong binders, where $y(a_i)$ denotes the measured binding strength (e.g., $\log(K_D)$) and τ is a threshold distinguishing strong from weak or non-binders (for MochiBind evaluation results, we used $\tau = 2$ for $\log(K_D) < 2$). Given the predicted ranking list $\mathcal{R}$ obtained from the ranking scores, the Top-K retrieval precision is defined as:

$$\text{Precision@}K = \frac{|\mathcal{R}_{1:K} \cap \mathcal{B}|}{K}.$$

This formulation evaluates the fraction of the model's top-K predictions that truly correspond to experimentally strong binders. A higher precision value indicates that the model's top-ranked antibodies are more likely to be valid binders, providing a direct measure of prediction quality under practical screening constraints. Together with Accuracy@K, which emphasizes correct coverage of the true best candidates, this metric offers a complementary perspective by focusing on the purity and reliability of the model's recommendations on binding signals.

4.5 Graphinity Model Finetuning

Although the Graphinity model[27] is designed to address generalization by training on a large-scale, high-quality simulated mutation dataset (∼1M single-point variants), we observed that its pretrained checkpoint does not transfer effectively to our AlphaBind pairwise affinity prediction setup. When evaluated directly on the 200,000 test pairs, the pretrained model's accuracy is close to random guessing. This mismatch can be attributed to several structural differences between the simulated training distribution and the real-world AlphaBind data.

First, the Graphinity training data consists exclusively of single-point mutations sampled from computationally generated complexes. This leads to relatively small and localized graph perturbations, whereas the AlphaBind dataset contains multi-point mutations, producing substantially larger and more diverse structural changes. Consequently, the pretrained model has limited exposure to the broader mutational landscape present in experimental data and struggles to generalize to these larger structural deviations. Second, in the simulated dataset, all antigen–antibody pairs share identical backbone coordinates; only the mutated side-chain atoms differ. In contrast, AlphaBind complexes are independently generated by the Boltz-2 structure prediction model, resulting in non-trivial translational and rotational differences even for the same antigen across different predictions. These variations introduce additional noise in the spatial representation and may violate the pretrained model's implicit assumptions about the invariance and alignment of complex coordinates. As a result, the model can misinterpret geometric discrepancies as meaningful structural signals, degrading predictive accuracy.

To ensure a fair comparison and provide a realistic baseline, we finetune Graphinity on the AlphaBind training data. Finetuning allows the model to adjust to the broader mutation spectrum, structural variability, and coordinate inconsistencies present in experimentally inspired datasets. During finetuning, we initialize from the pretrained Graphinity weights and optimize the model on the AlphaBind training split using

the same pairwise ranking objective as the MochiBind model. This adaptation step improves performance and ensures fair comparison, as shown in Appendix B. The finetuned Graphinity model is used as our baseline throughout all evaluations.

5 Discussion

By developing a unified framework for systematically evaluating existing state-of-the-art models and introducing the MochiBind approach, we highlight several key findings. First, pretrained protein language model (PLM) embeddings, such as those from ESM-2, capture substantial binding-relevant information even without explicit structural inputs. Sequence-based models, therefore, offer a scalable and easily adaptable solution for new antigen–antibody design tasks. In addition, the pairwise learning formulation, which focuses on relative rather than absolute binding strength, enables straightforward adaptation to new datasets and experimental assays that may differ in measurement units or noise characteristics. Furthermore, the proposed retrieval-based evaluation framework provides a more robust and practically meaningful measure of model performance that is less sensitive to noisy experimental binding data by assessing how effectively a model prioritizes high-affinity candidates, which is the primary goal in early-stage antibody screening. Consequently, model evaluation should emphasize identifying strong binders rather than predicting absolute affinity values.

Despite these promising results, progress is limited by the scarcity of large-scale, high-quality antibody–antigen affinity datasets suitable for commercial or open use. Expanding such datasets and designing more effective retrieval and ranking strategies will be crucial for improving the reliability and generalization power of computational antibody discovery pipelines. The unified framework developed in this work will serve as a comprehensive evaluation tool for such platforms.

6 Conclusion

In this study, we established a comprehensive benchmark for evaluating universal antibody–antigen binding affinity prediction and proposed a sequence-based model, MochiBind, designed for scalable antibody discovery for out-of-distribution antigen systems. Our benchmark provides standardized data splits, evaluation protocols, and retrieval-based metrics that together enable rigorous assessment of both sequence- and structure-based approaches. By focusing on ranking and retrieval rather than absolute regression, this framework provides a practical approach for candidate prioritization in *in silico* design workflows.

The proposed MochiBind model demonstrates promising out-of-sample performance, accurately predicting binding affinity for large pools of antibodies targeting unseen antigens. Across both pairwise accuracy and retrieval-based evaluation, MochiBind consistently outperforms structure-based models such as Graphinity and complex-wise approaches, including Boltz-2 and GeoDock, when tested on real-world experimental datasets. These results highlight the potential of sequence-only modeling as a simple, scalable, and computationally efficient approach for prioritizing candidates in out-of-distribution antibody discovery..

Resource Availability. All code used and model weights obtained in this study are available on https://osf.io/56azs/overview?view_only=df049360b3684b7db76f567502b164ea.

Acknowledgments. The authors declare no funding and no additional contributions to acknowledge.

Author Contributions. Conceptualization, Y.L., M.Y., Y.Z., K.S., and L.G.; methodology, Y.L. and M.Y.; software, Y.L.; analysis, Y.L.; investigation, Y.L.; data curation, Y.L. and Y.Z.; visualization, Y.L.; writing – original draft, Y.L.; writing – review & editing, all authors; supervision, M.Y.

Appendix A McNemar's Chi-squared Test

Detailed chi-squared values of MochiBind against all baseline models are shown in the table below.

Table A1: One-sided McNemar's χ^2 tests comparing the proposed model against baseline methods across datasets. Values denote χ^2 statistics with significance levels in parentheses.

Dataset	MochiBind vs GeoDock	MochiBind vs Boltz	MochiBind vs Graphinity
PEM	24 400.93 ($p < 0.0001$)	8 993.02 ($p < 0.0001$)	4 482.58 ($p < 0.0001$)
PP489	16 284.77 ($p < 0.0001$)	9 978.19 ($p < 0.0001$)	696.58 ($p < 0.0001$)
TRA	21 424.34 ($p < 0.0001$)	9 061.44 ($p < 0.0001$)	2 269.25 ($p < 0.0001$)
VHH72	4 275.65 ($p < 0.0001$)	169.26 ($p < 0.0001$)	94.21 ($p < 0.0001$)

Appendix B Graphinity Model Finetuning

Table A1 reports the performance of the pretrained Graphinity model on the AlphaBind pairwise dataset. Without any adaptation, the model performs only similar to a random guess across all antigen systems, indicating that the pretrained checkpoint does not transfer effectively to real experimental complexes.

After finetuning on the AlphaBind training split, the model exhibits substantial improvements. For example, accuracy on the TRA antigen system increases from 47.4% to 68.55%, with similar gains across PEM, VHH, and PP489. These results confirm that finetuning is necessary for Graphinity to function as a meaningful baseline in our setting.

Accordingly, all comparisons in the main text use the finetuned Graphinity model.

Appendix C Comparison with Win-Rate Ranking

To assess whether our results depend on the specific ranking aggregation method, we compare the TrueSkill-based ranking with a simpler win-rate ranking approach. The

Table B2: Pairwise accuracy (%) of Graphinity before and after fine-tuning on AlphaBind pairs.

Antigen	Pre-trained (%)	Finetuned (%)
PEM	48.80%	63.20%
PP489	49.80%	67.01%
TRA	47.40%	68.55%
VHH	50.90%	58.63%

goal of this comparison is to verify that the observed retrieval performance is primarily driven by the quality of the learned pairwise predictions, rather than the choice of ranking algorithm.

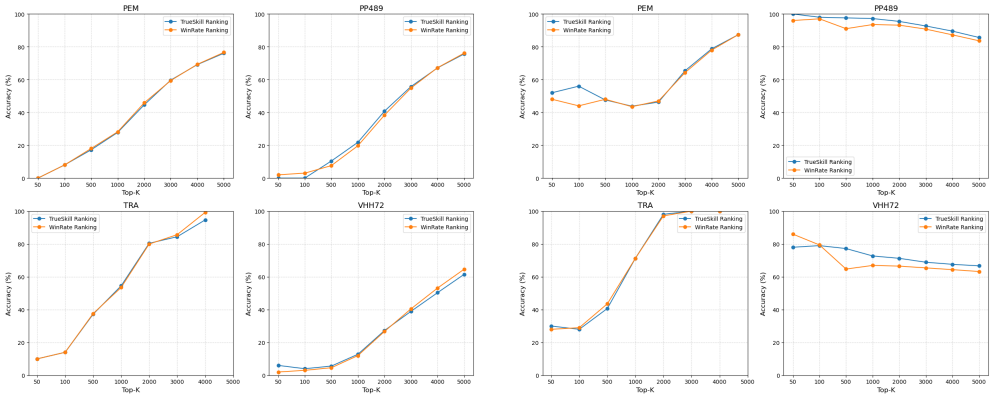
In the win-rate approach, each antibody is assigned a score based on the fraction of pairwise comparisons it wins. Formally, for an antibody i, the win-rate is defined as:

$$\text{WinRate}(i) = \frac{W_i}{N_i}, \tag{C1}$$

where W_i is the number of pairwise comparisons in which antibody i is predicted to outperform its opponent, and N_i is the total number of comparisons involving antibody i. Antibodies are then ranked in descending order of WinRate(i).

We evaluate retrieval accuracy and precision using the win-rate ranking and compare the results with those obtained using TrueSkill. As shown in Figure C1, the win-rate approach exhibits similar trends to TrueSkill across both metrics, indicating that the relative performance of the models is consistent under different aggregation strategies.

However, win-rate has several limitations in this setting. It does not account for the strength of opponents and can be unstable for antibodies that participate in only a small number of comparisons. In addition, win-rate requires recomputing the full ranking when new comparisons are added, making it less suitable for iterative screening workflows. In contrast, TrueSkill provides a more robust and scalable framework by modeling uncertainty and supporting incremental updates. For these reasons, we adopt TrueSkill as the primary ranking method in this work.



(a) Retrieval accuracy comparison between WinRate ranking and TrueSkill ranking approach.

(b) Retrieval precision comparison between WinRate ranking and TrueSkill ranking approach.

Fig. C1: Comparison between ranking algorithms. Both metrics show similar performance between the WinRate and TrueSkill ranking methods.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work, the author(s) used ChatGPT in order to polish the writing and correct grammar mistakes. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

References

[1] Köhler, G. & Milstein, C. Continuous cultures of fused cells secreting antibody of predefined specificity. *nature* **256**, 495–497 (1975).

[2] Mitra, S. & Tomar, P. C. Hybridoma technology; advancements, clinical significance, and future aspects. *Journal of Genetic Engineering and Biotechnology* **19**, 159 (2021).

[3] Smith, G. P. Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science* **228**, 1315–1317 (1985).

[4] McCafferty, J., Griffiths, A. D., Winter, G. & Chiswell, D. J. Phage antibodies: filamentous phage displaying antibody variable domains. *nature* **348**, 552–554 (1990).

[5] Marks, J. D. *et al.* By–passing immunization: building high affinity human antibodies by chain shuffling. *Bio/technology* **10**, 779–783 (1992).

[6] Hoogenboom, H. R. *et al.* Antibody phage display technology and its applications. *Immunotechnology* **4**, 1–20 (1998).

[7] Hammers, C. M. & Stanley, J. R. Antibody phage display: technique and applications. *Journal of Investigative Dermatology* **134**, 1–5 (2014).

[8] Poosarla, V. G. *et al.* Computational de novo design of antibodies binding to a peptide with high affinity. *Biotechnology and bioengineering* **114**, 1331–1342 (2017).

[9] Bennett, N. R. *et al.* Atomically accurate de novo design of antibodies with rfdiffusion. *Nature* 1–11 (2025).

[10] Shanehsazzadeh, A. *et al.* Unlocking de novo antibody design with generative artificial intelligence. *BioRxiv* 2023–01 (2023).

[11] Chidyausiku, T. M. *et al.* De novo design of immunoglobulin-like domains. *Nature communications* **13**, 5661 (2022).

[12] Kortemme, T. De novo protein design—from new structures to programmable functions. *Cell* **187**, 526–544 (2024).

[13] Chames, P., Van Regenmortel, M., Weiss, E. & Baty, D. Therapeutic antibodies: successes, limitations and hopes for the future. *British journal of pharmacology* **157**, 220–233 (2009).

[14] Wang, B., Gallolu Kankanamalage, S., Dong, J. & Liu, Y. Optimization of therapeutic antibodies. *Antibody therapeutics* **4**, 45–54 (2021).

[15] Hummer, A. M., Abanades, B. & Deane, C. M. Advances in computational structure-based antibody design. *Current opinion in structural biology* **74**, 102379 (2022).

[16] Xie, X., Valiente, P. A., Lee, J. S., Kim, J. & Kim, P. M. Antibody-sgm, a score-based generative model for antibody heavy-chain design. *Journal of Chemical Information and Modeling* **64**, 6745–6757 (2024).

[17] Rabia, L. A., Desai, A. A., Jhajj, H. S. & Tessier, P. M. Understanding and overcoming trade-offs between antibody affinity, specificity, stability and solubility. *Biochemical engineering journal* **137**, 365–374 (2018).

[18] Kairys, V., Baranauskiene, L., Kazlauskiene, M., Matulis, D. & Kazlauskas, E. Binding affinity in drug design: experimental and computational techniques. *Expert opinion on drug discovery* **14**, 755–768 (2019).

[19] Ross, G. A. *et al.* The maximal and current accuracy of rigorous protein-ligand binding free energy calculations. *Communications Chemistry* **6**, 222 (2023).

[20] Jarmoskaite, I., AlSadhan, I., Vaidyanathan, P. P. & Herschlag, D. How to measure and evaluate binding affinities. *Elife* **9**, e57264 (2020).

[21] Guest, J. D. *et al.* An expanded benchmark for antibody-antigen docking and affinity prediction reveals insights into antibody recognition determinants. *Structure* **29**, 606–621 (2021).

[22] Li, L. *et al.* Machine learning optimization of candidate antibody yields highly diverse sub-nanomolar affinity antibody libraries. *Nature communications* **14**, 3454 (2023).

[23] Mason, D. M. *et al.* Optimization of therapeutic antibodies by predicting antigen specificity from antibody sequence via deep learning. *Nature biomedical engineering* **5**, 600–612 (2021).

[24] Frey, N. C. *et al.* Lab-in-the-loop therapeutic antibody design with deep learning. *bioRxiv* 2025–02 (2025).

[25] Passaro, S. *et al.* Boltz-2: Towards accurate and efficient binding affinity prediction. *BioRxiv* 2025–06 (2025).

[26] Chu, L.-S., Ruffolo, J. A., Harmalkar, A. & Gray, J. J. Flexible protein–protein docking with a multitrack iterative transformer. *Protein Science* **33**, e4862 (2024).

[27] Hummer, A. M., Schneider, C., Chinery, L. & Deane, C. M. Investigating the volume and diversity of data needed for generalizable antibody–antigen $\delta\delta$ g prediction. *Nature Computational Science* 1–13 (2025).

[28] Agarwal, A. A. *et al.* Alphabind, a domain-specific model to predict and optimize antibody–antigen binding affinity. *Mabs* **17**, 2534626 (2025).

[29] Yuan, Y., Chen, Q., Mao, J., Li, G. & Pan, X. Dg-affinity: predicting antigen–antibody affinity with language models from sequences. *BMC bioinformatics* **24**, 430 (2023).

[30] Shrestha, P. *et al.* Nanobinder: a machine learning assisted nanobody binding prediction tool using rosetta energy scores. *Journal of Cheminformatics* **17**, 96 (2025).

[31] Chen, C. *et al.* Affinityflow: Guided flows for antibody affinity maturation. *arXiv preprint arXiv:2502.10365* (2025).

[32] Chong, L. T., Duan, Y., Wang, L., Massova, I. & Kollman, P. A. Molecular dynamics and free-energy calculations applied to affinity maturation in antibody 48g7. *Proceedings of the National Academy of Sciences* **96**, 14330–14335 (1999).

[33] Laitinen, T., Kankare, J. A. & Peräkylä, M. Free energy simulations and mm–pbsa analyses on the affinity and specificity of steroid binding to antiestradiol antibody.

PROTEINS: Structure, Function, and Bioinformatics **55**, 34–43 (2004).

[34] Xia, Z., Huynh, T., Kang, S.-g. & Zhou, R. Free-energy simulations reveal that both hydrophobic and polar interactions are important for influenza hemagglutinin antibody binding. *Biophysical Journal* **102**, 1453–1461 (2012).

[35] Chong, W. L., Saparpakorn, P., Sangma, C., Lee, V. S. & Hannongbua, S. Insight into free energy and dynamic cross-correlations of residue for binding affinity of antibody and receptor binding domain sars-cov-2. *Heliyon* **9** (2023).

[36] Zhu, F. *et al.* Large-scale application of free energy perturbation calculations for antibody design. *Scientific Reports* **12**, 12489 (2022).

[37] Voets, P. J. On the antigen-antibody interaction: a thermodynamic consideration. *Human Antibodies* **26**, 39–41 (2018).

[38] Kelley, R. F. & O'Connell, M. P. Thermodynamic analysis of an antibody functional epitope. *Biochemistry* **32**, 6828–6835 (1993).

[39] de Ruiter, A., Boresch, S. & Oostenbrink, C. Comparison of thermodynamic integration and bennett acceptance ratio for calculating relative protein-ligand binding free energies. *Journal of computational chemistry* **34**, 1024–1034 (2013).

[40] Gutiérrez, M., Vallejos, G. A., Cortés, M. P. & Bustos, C. Bennett acceptance ratio method to calculate the binding free energy of bace1 inhibitors: Theoretical model and design of new ligands of the enzyme. *Chemical Biology & Drug Design* **93**, 1117–1128 (2019).

[41] Gromiha, M. M., Yugandhar, K. & Jemimah, S. Protein–protein interactions: scoring schemes and binding affinity. *Current opinion in structural biology* **44**, 31–38 (2017).

[42] Barlow, K. A. *et al.* Flex ddg: Rosetta ensemble-based estimation of changes in protein–protein binding affinity upon mutation. *The Journal of Physical Chemistry B* **122**, 5389–5399 (2018).

[43] Herbrich, R., Minka, T. & Graepel, T. Trueskill™: a bayesian skill rating system. *Advances in neural information processing systems* **19** (2006).

[44] Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).