Evaluating Zero-Shot Scoring for *In Vitro* Antibody Binding Prediction with Experimental Validation

Divya Nori*
Massachusetts Institute of Technology
Department of Electrical Engineering and Computer Science
77 Massachusetts Ave, Cambridge, MA 02139
divnor80@mit.edu

Simon V. Mathis*
University of Cambridge
Department of Computer Science and Technology
simon.mathis@cl.cam.ac.uk

Amir Shanehsazzadeh
Absci Corporation
ashanehsazzadeh@absci.com

Abstract

The success of therapeutic antibodies relies on their ability to selectively bind antigens. AI-based antibody design protocols have shown promise in generating epitope-specific designs. Many of these protocols use an *inverse folding* step to generate diverse sequences given a backbone structure. Due to prohibitive screening costs, it is key to identify candidate sequences likely to bind *in vitro*. Here, we compare the efficacy of 8 common scoring paradigms based on open-source models to classify antibody designs as binders or non-binders. We evaluate these approaches on a novel surface plasmon resonance (SPR) dataset, spanning 5 antigens. Our results show that existing methods struggle to detect binders, and performance is highly variable across antigens. We find that metrics computed on flexibly docked antibody-antigen complexes are more robust, and ensembles scores are more consistent than individual metrics. We provide experimental insight to analyze current scoring techniques, highlighting that the development of robust, zero-shot filters is an important research gap.

1 Introduction

Antibodies are a prominent class of therapeutic agents, primarily because they can selectively bind to a wide array of disease-causing target antigens. This binding specificity is governed by interactions between the antibody's paratope region and the antigen's epitope region [1]. Epitope specificity is primarily facilitated by the structure of hypervariable loops in the paratope called complimentarity determining regions (CDRs).

Computationally designing the CDRs such that they bind to specific epitopes with high affinity has become a problem of great interest. Researchers often start with a known or predicted antibody backbone structure, and the task is to design sequences that assume this backbone conformation [2]. Recently, generative machine learning methods have proven effective for this *inverse folding* problem on protein design [3]. Given the hypervariability of CDRs, there are many sequences that can fold into the same 3D backbone structure. Filtering the diverse outputs of an inverse folding model *in silico* to select for sequences with greater chance of *in vitro* success remains an important challenge, since experimental resources are often limited.

*Work done during internship at Absci.

NeurIPS 2023 Generative AI and Biology Workshop.

In the protein space, many *in silico* metrics are commonly used to evaluate the quality of designs [4], [5]. Here, we investigate how these metrics perform for antibodies. We benchmark several scoring paradigms on their ability to filter antibody designs by predicted success in a surface plasmon resonance (SPR) assay, an industry standard assay used to measure binding affinity.

2 Background

Problem set-up As shown in Figure 1, inverse folding models design candidate sequences given a desired backbone structure. A subset of these sequences will fold into a 3D structure *in vitro*, and a smaller subset then assume paratope residue conformations that interact strongly with the antigen [1].

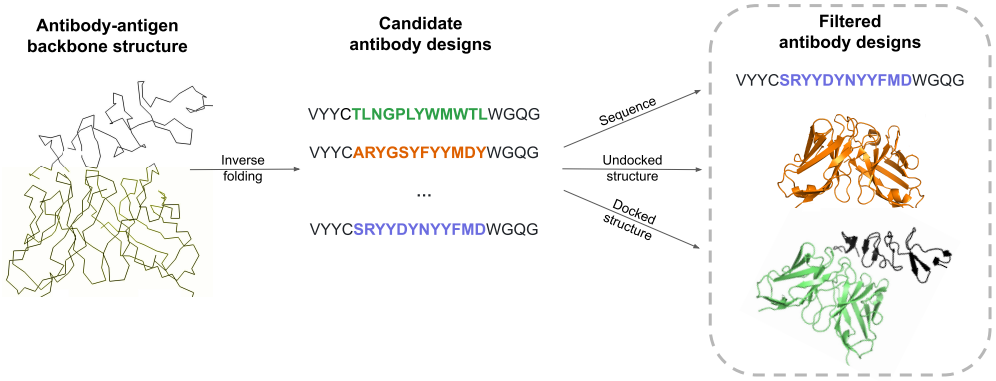

Figure 1: Antibody-inverse folding model with antibody-antigen complex used as input predicts a set of candidate CDR sequences. These sequences are then filtered at the sequence, undocked structure, or docked structure level, which is the focus of this study. The HCDR3 sequences are from [6].

Here, we assess whether filtering at the sequence level, undocked structure level, or docked structure level is effective. Particularly, we compare 8 different scoring paradigms on the binder vs non-binder classification task. The candidate scoring paradigms we evaluate are all zero-shot methods – meaning they are not trained on any binding affinity data which is often available in only limited quantities.

Related work In general protein binder design, *in silico* metrics have been used to filter libraries and improve binder success rates. Most recently, it was shown that monomer metrics are less effective in predicting binders than complex metrics [7]. On antibodies, prior work assesses Rosetta's predicted free energy of binding ($\Delta\Delta G$) as a binder scoring function and achieves an AUROC of 0.55 on a library designed for HER2 [8]. Recently, a diffusion model was trained to assess the quality of antibody designs by generating and scoring docked poses [9]. Language models have also been used to perform efficient evolution on antibodies, indicating that their likelihoods can be used to select predicted binders [10]. We assess whether such approaches are effective, given novel wet-lab insight.

3 Methods

Data To generate the antibody libraries scored in this work, an antibody-specific inverse folding model was given an antibody-antigen complex backbone structure, antibody framework sequences, and antigen sequence as input. The model was then used to design several candidate antibody sequences, specifically all three HCDRs of a reference antibody to the antigen of interest. Libraries were designed for 5 different antigens, and wet-lab data was generated by SPR. A future publication will describe additional details on data generation. See Table 2 for details on library size and % binders.

Candidate scoring functions Table 1 describes the 8 candidate scoring functions we evaluate on the binder classification task. We hypothesize that antibody candidate sequences that contain common or predictable motifs are more likely to be binders. Therefore, we benchmark pseudo-perplexity predicted by the ESM-2 650M parameter model as a sequence scoring metric [11]. Pseudo-perplexity

was computed for each sequence, scaled by min-max normalization, flipped to range [-1,0], and translated to range [0,1]. All scoring paradigms going forward are normalized in a similar manner. Subsequently, we hypothesize that antibody candidates with biologically plausible structural qualities are more likely to be successful binders. As an undocked structure scoring metric, we benchmark the residue-level confidence scores given by ABodyBuilder2 (ABB2), an antibody-specific folding model [12].

Other discriminative scoring functions used in-practice are self-consistency metrics, such as similarity between the folded antibody candidate and the input backbone structure [5]. Here, we benchmark root mean square deviation (RMSD) as a similarity measure, under three Kabsch alignments [13].

Given that flexible interactions between antibody and antigen is critical for binding, we benchmark metrics on flexibly docked complexes. To generate these complexes, we apply a previously proposed architecture, dyMEAN [14], and make several modifications to the training procedure which are described in the Appendix. We evaluate three scoring paradigms on the predicted complexes, selected because they capture established heuristics on antibody-antigen interactions. The scores we evaluate measure interface proximity, contact proximity, and interface interactions (details in Appendix).

	Model/Method	**Metric**
Sequence Only	ESM	Pseudo-Perplexity
Undocked Structure	ABodyBuilder2	Residue-Level Model Confidence
	Antibody-Aligned Antibody	RMSD
	All HCDR-Aligned Antibody	RMSD
	HCDR3-Aligned Antibody	RMSD
Docked Complex	Modified dyMEAN	Interface Proximity
	Modified dyMEAN	Contact Proximity
	Modified dyMEAN	Interface Interactions

Table 1: Evaluation of 8 candidate scoring functions on the binder classification task across 3 classes of metrics: sequence only, undocked structure, and docked complex.

4 Results

We assess the performance of each scoring paradigm in the classification setting, evaluating across multiple antigens. Figure 2 shows each scoring function's distribution of AUROC values. We observe that antibody-aligned RMSD performs the best (mean of 0.65 and standard deviation of 0.15), indicating that undocked structure filtering may be effective. However, all three of the RMSD-based undocked structure metrics display high variance in performance. Interface proximity is the most consistent metric (mean 0.51 and standard deviation 0.028), followed by contact proximity and interface interactions. This reveals that docked complex metrics exhibit less variability.

After evaluating the performance of each metric individually, we assess whether we can achieve better discrimination with a combined scoring paradigm. Since the scoring functions are not strongly correlated, as shown in the Appendix, we hypothesize that a combined method may boost performance.

For any linear combination of metrics, we fit coefficients on 4 out of 5 antigens, where our objective is mean AUROC. We do not consider combinations whose standard deviation exceeds 0.1. To avoid leakage, we evaluate the best-performing ensemble score on the held-out antigen. We call this approach "Best Combination on Hold-Out," and report retrieval precision on the task of filtering to 10% of the original antibody design library's size in Figure 3. Retrieval precision, or Precision @ K, measures the proportion of true binders in the top K highest-scoring candidates. Retrieval precision and recall curves for the individual scoring metrics are given in the Appendix. We also evaluate performance on an oracle approach where we find the best ensemble score on all 5 antigens, which we refer to "Best Combination on All." Our baseline is given by the binder percentage in each of the libraries, and performance of the best individual metric (antibody-aligned RMSD) is also shown for comparison.

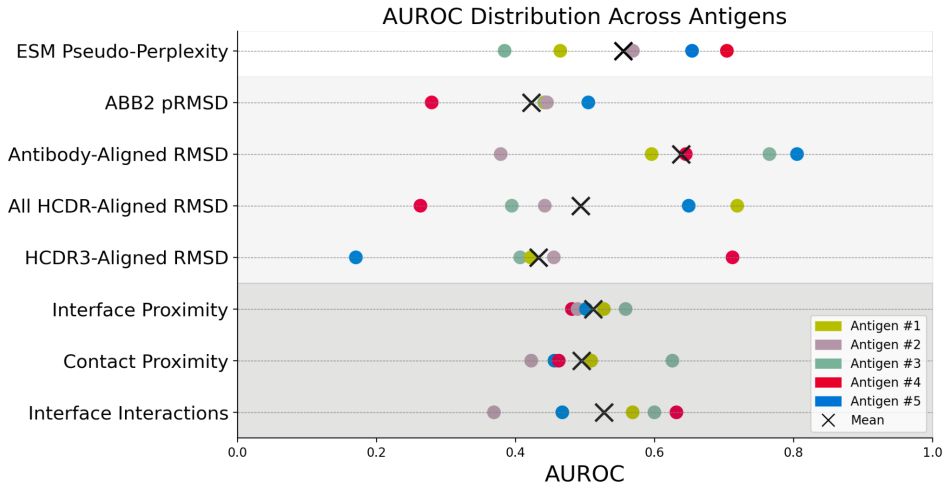

Figure 2: Distribution of AUROC values for various scoring functions across 5 antigens.

We observe that ensembling does not improve performance. However, the standard deviation of the hold-out ensemble scores is 0.074, compared to 0.22 for antibody-aligned RMSD. This indicates that ensembling is more robust to antigen variation. Additionally, the hold-out ensemble scores are just 2.9 percentage points below the oracle for retrieval precision, revealing that we cannot perform better on this classification task using these scoring criteria. This underscores the need for better prediction methods beyond those evaluated here. We find that in almost all folds, interface proximity has the highest weight. The specific weights are given in the Appendix. This reveals that flexible docking information is critical for this task, and better antibody-antigen docking models can push the needle.

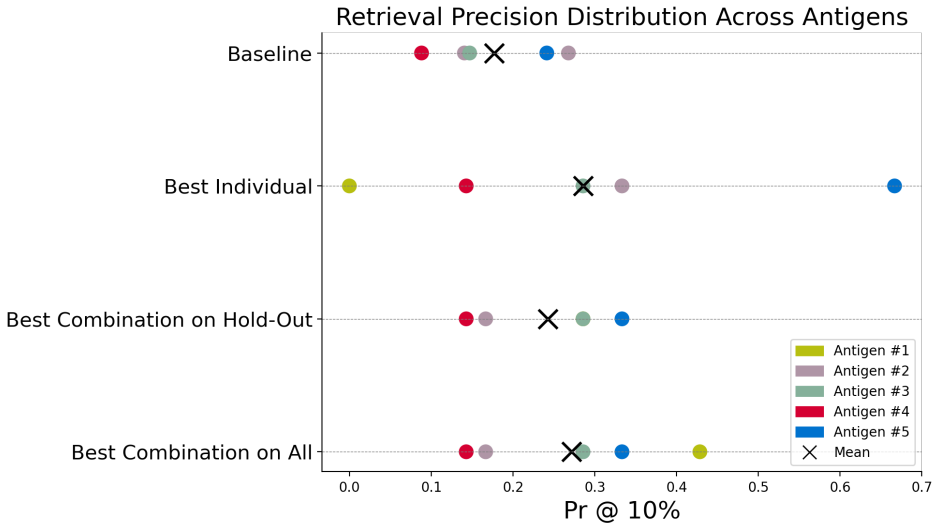

Figure 3: Distribution of Precision @ 10% for ensemble scores and baselines across 5 antigens.

5 Conclusions

This study provides experimental insight to show that there is no one-size-fits-all approach for antibody assessment, and current open-source tools are insufficient to reliably filter diverse antibody libraries. We learn that antibody-aligned RMSD outperforms other individual scoring paradigms. However, sequence and undocked structure-only scoring paradigms struggle to handle antigen

variation, and metrics that use antigen information are more robust. We find that linear combinations of the individual scoring paradigms exhibit lower variance in performance but leave much to be desired. Improved flexible docking models and further experimental benchmark efforts are critical to identify promising antibody candidates *in silico* and accelerate their clinical development.

Acknowledgments

The authors wish to thank Kieran Didi, Amaro Taylor-Weiner, Christine Lemke, Byron Olsen, and Kristin Iannacone for critical review of this manuscript.

Competing interest statement

The authors are current or former employees, contractors, interns, or executives of Absci Corporation and may hold shares in Absci Corporation.

References

[1] Hung-Pin Peng, Kuo Hao Lee, Jhih-Wei Jian, and An-Suei Yang. Origins of specificity and affinity in antibody–protein interactions. *Proceedings of the National Academy of Sciences*, 111(26):E2656–E2665, 2014.

[2] Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, and Alexander Rives. Learning inverse folding from millions of predicted structures. pages 8946–8970, 2022.

[3] J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, A. Courbet, R. J. de Haas, N. Bethel, P. J. Y. Leung, T. F. Huddy, S. Pellock, D. Tischer, F. Chan, B. Koepnick, H. Nguyen, A. Kang, B. Sankaran, A. K. Bera, N. P. King, and D. Baker. Robust deep learning–based protein sequence design using ProteinMPNN. *Science*, 378(6615):49–56, October 2022. doi: 10.1126/science.add2187. URL `https://doi.org/10.1126/science.add2187`.

[4] Sean R Johnson, Xiaozhi Fu, Sandra Viknander, Clara Goldin, Sarah Monaco, Aleksej Zelezniak, and Kevin K Yang. Computational scoring and experimental evaluation of enzymes generated by neural networks. *bioRxiv*, pages 2023–03, 2023.

[5] Kevin Eric Wu, Kevin K Yang, Rianne van den Berg, James Zou, Alex Xijie Lu, and Ava P Amini. Protein structure generation via folding diffusion. 2022.

[6] Amir Shanehsazzadeh, Sharrol Bachas, Matt McPartlon, George Kasun, John M Sutton, Andrea K Steiger, Richard Shuai, Christa Kohnert, Goran Rakocevic, Jahir M Gutierrez, et al. Unlocking de novo antibody design with generative artificial intelligence. *bioRxiv*, pages 2023–01, 2023.

[7] Nathaniel R Bennett, Brian Coventry, Inna Goreshnik, Buwei Huang, Aza Allen, Dionne Vafeados, Ying Po Peng, Justas Dauparas, Minkyung Baek, Lance Stewart, et al. Improving de novo protein binder design with deep learning. *Nature Communications*, 14(1):2625, 2023.

[8] Derek M Mason, Simon Friedensohn, Cédric R Weber, Christian Jordi, Bastian Wagner, Simon Meng, Pablo Gainza, Bruno E Correia, and Sai T Reddy. Deep learning enables therapeutic antibody optimization in mammalian cells by deciphering high-dimensional protein sequence space. *BioRxiv*, page 617860, 2019.

[9] Zhangzhi Peng, Chenchen Han, Xiaohan Wang, Dapeng Li, and Fajiie Yuan. Generative diffusion models for antibody design, docking, and optimization. *bioRxiv*, 2023. doi: 10.1101/2023.09.25.559190.

[10] Brian L Hie, Varun R Shanker, Duo Xu, Theodora UJ Bruun, Payton A Weidenbacher, Shaogeng Tang, Wesley Wu, John E Pak, and Peter S Kim. Efficient evolution of human antibodies from general protein language models. *Nature Biotechnology*, 2023.

[11] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *BioRxiv*, 2022:500902, 2022.

[12] Brennan Abanades, Wing Ki Wong, Fergus Boyles, Guy Georges, Alexander Bujotzek, and Charlotte M Deane. Immunebuilder: Deep-learning models for predicting the structures of immune proteins. *Communications Biology*, 6(1):575, 2023.

[13] Wolfgang Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, 32(5):922–923, 1976.

[14] Xiangzhe Kong, Wenbing Huang, and Yang Liu. End-to-end full-atom antibody design. *ICML Conference Poster*, 2023.

[15] James Dunbar, Konrad Krawczyk, Jinwoo Leem, Terry Baker, Angelika Fuchs, Guy Georges, Jiye Shi, and Charlotte M Deane. Sabdab: the structural antibody database. *Nucleic acids research*, 42(D1): D1140–D1146, 2014.

[16] Peter Eastman, Jason Swails, John D Chodera, Robert T McGibbon, Yutong Zhao, Kyle A Beauchamp, Lee-Ping Wang, Andrew C Simmonett, Matthew P Harrigan, Chaya D Stern, et al. Openmm 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS computational biology*, 13(7): e1005659, 2017.

A Dataset details

Table 2 shows the number of screened antibodies and the percentage of binding antibodies for each library.

	Antigen #1	**Antigen #2**	**Antigen #3**	**Antigen #4**	**Antigen #5**
# of Sequences	64	56	68	68	58
# of Binders	9	15	10	6	14
% Binders	14.1	26.8	14.7	8.8	24.1

Table 2: Five libraries, each designed and screened against one of five antigens using surface plasmon resonance (SPR), are used as experimental data in this study. Each library consists of between 56 and 68 unique sequences with between 8.8% and 24.1% of the sequences binding to their target antigen.

B dyMEAN training

We make the following modifications to the training procedure of dyMEAN. While we train on the Structural Antibody Database [15] as in the original dyMEAN paper, we split SAbDab into training and validation sets according to clusters formed by complexes sharing $\geq 40\%$ antigen sequence similarity. The clusters containing any of the 5 evaluation antigens were dropped from both training and validation. Additionally, we use OpenMM [16] to relax the bound antigen structures before model input to ensure that we are not leaking information about docked position. Finally, we provide the model with relaxed ABB2-predicted antibody structures to focus on learning flexible docking rather than folding.

C Docked complex scoring paradigms

We benchmark three scoring functions computed on docked complexes predicted by dyMEAN. The first is an interface proximity score which gives the average distance between the residues on the antibody interface and their nearest corresponding antigen residues. Here, we define interface as the contiguous region between the first residue of HCDR1 to the last residue of HCDR3, including HFWR2 and HFWR3. Residue-residue distances are computed as minimum distance between the side chain atoms of both residues. We also compute a contact proximity score which computes the same score on just the antibody contact residues which are defined as having at least one side chain atom within 5 Angstroms of an antigen atom.

Lastly, we compute an interface interactions score which evaluates the types of interactions at the antibody-antigen interface. We begin by identifying all antibody-antigen residue pairs where an antibody residue atom is within at most 3 Angstroms of an antigen residue atom. We then classify each residue-residue pair, awarding higher scores to polar-polar interactions, nonpolar-nonpolar Van der Waals interactions, positive-negative ionic interactions, and aromatic stacking interactions.

D Correlation analysis

Figure 4 shows correlations between the 8 scoring paradigms. Correlation coefficients were computed on all pairs of each antigen's set of scores, and these coefficients were averaged across all 5 targets. As expected, the RMSD-based undocked structure metrics exhibit some correlation, and the metrics computed on flexibly docked complexes are also correlated. With the exception of ESM pseudo-perplexity and ABB2 pRMSD, which are weakly correlated, we do not observe much correlation across the sequence, undocked structure, and docked complex metric groups. Therefore, we hypothesize that combining individual metrics may reveal new information useful for binder classification.

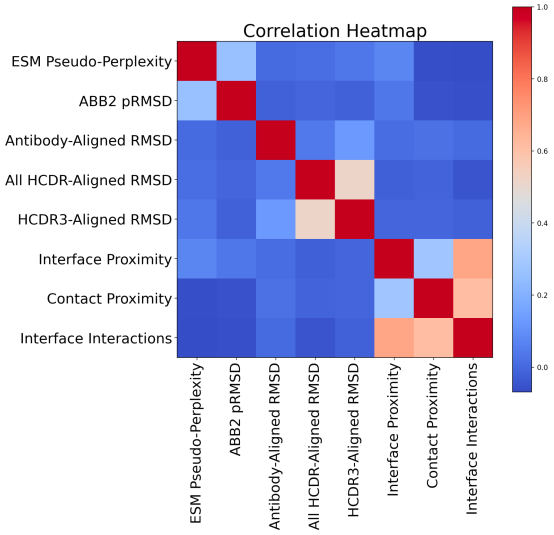

Figure 4: Correlations between the 8 scoring paradigms, averaged across antigens, show that each metric is capturing fairly different signal. Combining metrics could yield favorable performance.

E Retrieval precision and recall

Retrieval precision and recall metrics are particularly relevant because we can measure the number of true binders amongst the K we may select to test in the lab, while ensuring that a high proportion of binders are not missed. Precision @ K measures the proportion of true binders in the top K highest-scoring candidates, and Recall @ K measures the proportion of relevant items in the top K. We evaluate retrieval precision and recall values for each scoring function averaged across the 5 antigens, as shown in Figure 5. We focus on the task of filtering to 10% of the original antibody design library's size. We observe that antibody-aligned RMSD has the highest Precision @ 10% with a value of 0.26, averaged across antigens, compared to a baseline mean binder rate of 0.13. Antibody-aligned RMSD also has the highest retrieval recall of 0.16 at this threshold.

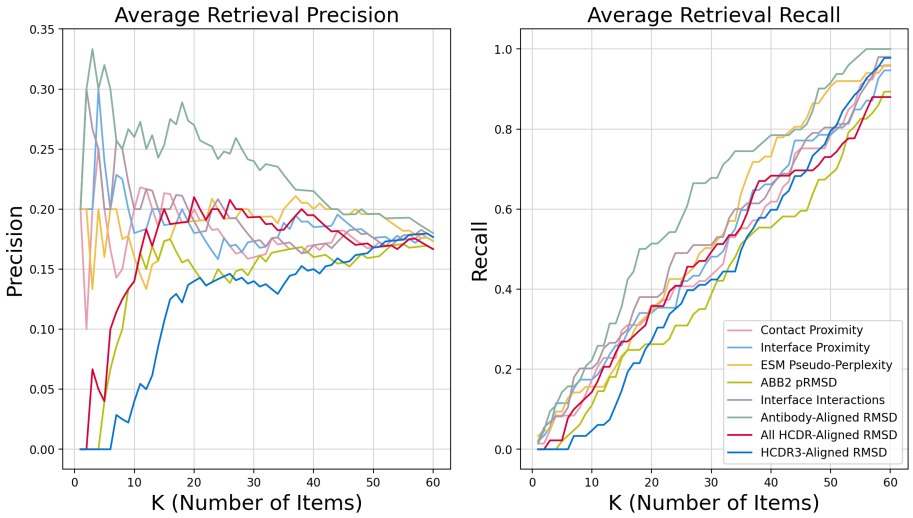

Figure 5: The retrieval precision and recall curves, averaged across antigens, show how effectively the different scoring paradigms can select K designs. At a threshold of 10%, which is approximately $K = 6$, antibody-aligned RMSD displays the best retrieval precision and recall.

F Combined scoring function weights

In Table 3, we report the coefficients found when each group of 4 antigens was used for optimization.

Held-out Antigen	Metric 1	Metric 2	Metric 3
Antigen #1	Interface Proximity, 0.7	Contact Proximity, 0.2	Interface Interactions, 0.1
Antigen #2	Antibody-Aligned RMSD, 0.5	Interface Interactions, 0.4	ESM Pseudo-Perplexity, 0.1
Antigen #3	Interface Proximity, 0.7	Contact Proximity, 0.2	Interface Interactions, 0.1
Antigen #4	Interface Proximity, 0.7	Contact Proximity, 0.2	Interface Interactions, 0.1
Antigen #5	Interface Proximity, 0.7	ESM Pseudo-Perplexity, 0.2	Contact Proximity, 0.1
Oracle	Interface Proximity, 0.6	Contact Proximity, 0.3	HCDR3-Aligned RMSD, 0.1

Table 3: Best weighted ensemble scores for each of the 5 folds, as measured by mean AUROC.