IMMUNOLOGY

AbEpiTope-1.0: Improved antibody target prediction by use of AlphaFold and inverse folding

Joakim Nøddeskov Clifford[1], Eve Richardson[2], Bjoern Peters[2], Morten Nielsen[1]*

B cell epitope prediction tools are crucial for designing vaccines and disease diagnostics. However, predicting which antigens a specific antibody binds to and their exact binding sites (epitopes) remains challenging. Here, we present AbEpiTope-1.0, a tool for antibody-specific B cell epitope prediction, using AlphaFold for structural modeling and inverse folding for machine learning models. On a dataset of 1730 antibody-antigen complexes, AbEpiTope-1.0 outperforms AlphaFold in predicting modeled antibody-antigen interface accuracy. By creating swapped antibody-antigen complex structures for each antibody-antigen complex using incorrect antibodies, we show that predicted accuracies are sensitive to antibody input. Furthermore, a model variant optimized for antibody target prediction—differentiating true from swapped complexes—achieved an accuracy of 61.21% in correctly identifying antibody-antigen pairs. The tool evaluates hundreds of structures in minutes, providing researchers with a resource for screening antibodies targeting specific antigens. AbEpiTope-1.0 is freely available as a web server and software.

INTRODUCTION

B cells are vital for the adaptive immune system, providing long-term defense against pathogens and cancer. Their activation occurs when B cell receptors (BCRs)—membrane-bound antibodies—interact with specific antigens. The interaction sites on the antigen and antibody are known as the epitope and paratope, respectively.

Identifying epitopes is crucial for vaccine design (*1*), diagnostics (*2*), and therapeutic antibody development (*3*). However, experimental epitope identification is complex and costly, requiring extensive screening. In silico epitope prediction methods have, therefore, emerged as critical tools for predicting the most likely epitopes, thus reducing experimental workloads. Most approaches are antibody agnostic, focusing on antigen surface residues likely to interact with any antibody. Notable progress has been made with tools such as BepiPred (*4*–*6*), DiscoTope (*7*, *8*), and SEPPA (*9*), which provide predictions without needing antibody input.

However, antibody-antigen (AbAg) interactions are highly specific. A more precise task is, therefore, antibody-specific epitope prediction, incorporating the unique structure of an antibody to predict its epitope and paratope (*10*). This has far-reaching applications in understanding immune responses. For instance, in autoimmune diseases—where the immune system attacks the body's tissues—identifying antibody-specific epitopes can reveal targeted self-proteins. In cancer immunotherapy, it can help pinpoint antibodies against tumor-specific antigens for personalized treatment. In addition, in drug hypersensitivity, such predictions can identify drug components triggering adverse immune responses, guiding safer alternative treatments.

Experimental methods such as x-ray crystallography and cryo–electron microscopy offer precise insights into AbAg interactions but are resource intensive and limited by the need for crystallization (for x-ray crystallography) (*11*, *12*). Phage display is faster but lacks atomic-level precision (*13*). These challenges highlight the need for computational methods. Previously, the scarcity of paired antibody sequences has limited tool development and practical application. However, advancements in high-throughput single-cell sequencing and variable-diversity-joining (V(D)J) sequencing of the BCR repertoire have improved the availability of paired antibody sequence data (*14*).

Antibody-specific epitope prediction is challenging because of the flexible and diverse nature of the complementarity-determining regions (CDRs)—regions that mediate antigen contact—shaped by V(D)J recombination. Early work by Sela-Culang *et al.* (*15*) and Jespersen *et al.* (*16*) used machine learning to predict AbAg interactions based on sequential, physicochemical, and geometric features. AbAdapt advanced the field by developing models to score homology-modeled AbAg complexes docked using rigid-body protocols (*17*). Further improvements were demonstrated by modeling antibodies and antigens with the monomeric version of AlphaFold (*18*, *19*). Subsequent studies have shown that while the multimeric version of AlphaFold-2.3 performs worse at predicting AbAg interfaces compared to other protein-protein interfaces, it still outperforms contemporary tools (*20*).

In this work, we present computational methods for antibody-specific B cell epitope prediction using AlphaFold-2.3 multimer as the structural modeling backbone and inverse folding for evaluating modeled AbAg interfaces (*21*). We introduce two tools: AbEpiScore-1.0 for assessing the accuracy of modeled AbAg interfaces and AbEpiTarget-1.0 for selecting the antibody most likely to bind a given antigen.

RESULTS

AbEpiScore-1.0: Improved prediction of AbAg interface accuracy

Our first aim was to develop a method for predicting the interface accuracy of modeled AbAg structures. We first tested AbAg interface scores for AlphaFold-2.3 and inverse folding Geometric Vector Perceptron (GVP)-Transformers, Evolutionary Scale Modeling for Inverse Folding (ESMIF1) and AntiFold, on 1730 AbAgs without fine-tuning, creating 30 structures for each using AlphaFold-2.3 multimer, totaling 51,900 structures (*22*, *23*). Note that we here include structures that were released before AlphaFold-2.3's training cutoff date

[1]Department of Health Technology, Technical University of Denmark, Kgs. Lyngby 2800, Denmark. [2]Center for Infectious Disease and Vaccine Research, La Jolla Institute for Immunology, La Jolla, CA, USA.
*Corresponding author. Email: morni@dtu.dk

(30 September 2021). The impact of this is analyzed in detail in the "AbEpiTope-1.0 maintains performative edge on post-AlphaFold-2.3 training data" section.

Inverse folding aims to determine the amino acid sequence compatible with a given protein's three-dimensional (3D) structure. ESMIF1 and AntiFold can output per-residue probabilities or encodings that indicate antibody or antigen residue compatibility with the protein fold.

AbAg Intersection over Union (AbAgIoU) was used to measure the match between the predicted epitope and paratope residues and the corresponding ground truth crystal structures. For details on this performance metric, inverse folding, and the modeling pipeline, refer to Materials and Methods.

We illustrated the correspondence between AbAgIoU and ESMIF1 scores in a 2D hexagon plot. This plot indicated a large structure density of low-scoring structures in the bottom left corner and smaller densities in the top middle-right corner, contributing to a linear correspondence with a Pearson's correlation coefficient (PCC) of 0.6613. Furthermore, we colored bins containing a single structure in orange to indicate that there is little structure density in outlier hexagonal bins (Fig. 1A). In comparison, the weighted average of AlphaFold-2.3's confidence metrics [0.8 interface predicted Template Modeling score (ipTM) + 0.2 predicted Template Modeling score (pTM)] and the inverse folding model AntiFold had PCCs of 0.5602 and 0.5090, respectively. In text S4, we provide the same 2D hexagon plots for all models described in Table 1.

In Fig. 1B, we show the distribution and model scores for models binned on AbAgIoU, displaying for all three methods (AlphaFold-2.3, ESMIF1, and AntiFold) a positive trend between interface score and AbAgIoU, meaning that the structure accuracy is increased as a function of AlphaFold-2.3, ESMIF1, and AntiFold scores. However, despite these observations, we note that 17.6% of the AlphaFold-2.3 modeling attempts (306 AbAgs) yielded an AbAgIoU of 0 for all structures (leftmost bar in Fig. 1B). Furthermore, there are 38.3% (663 AbAgs) with a low epitope and paratope residue match with the ground truth (AbAgIoU, 0 < 0.1 bar; Fig. 1B).

Next, we investigated the ability of these models to classify accurately modeled AbAg interfaces. To avoid settling on an arbitrary

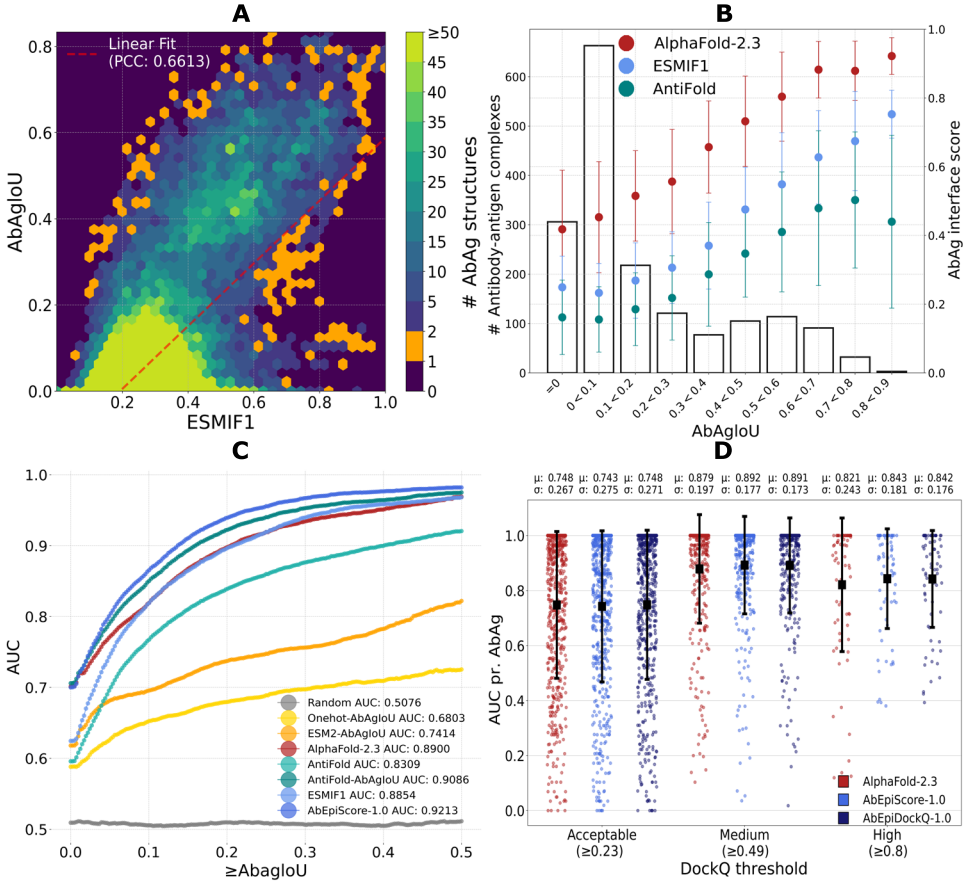

Fig. 1. We measure the ability of AbAg interface scoring models to predict and classify accurately modeled AbAg structures. (**A**) Min-max–scaled ESMIF1 scores (*x* axis) and AbAgIoU values (*y* axis) for 51,900 structures from 1730 AbAgs are placed into 2D hexagonal bins. A color scale capped at 50 structures shows the structure count per bin, and orange indicates bins containing a single structure. A red dashed line indicates a linear fit computed across all structures. (**B**) The best of 30 modeled structures (based on the highest AbAgIoU) for each AbAg is binned by AbAgIoU (*x* axis). The left *y* axis shows the number of AbAgs per bin, and the right *y* axis shows the average and SD of interface scores (AlphaFold-2.3, ESMIF1, and AntiFold) within each bin. (**C**) AUC values across 205 AbAgIoU accuracy thresholds (0.0 to 0.5) (*x* axis) compare the model performance of random, AlphaFold, AntiFold, ESMIF1, AbEpiScore, and AbAgIoU models (*y* axis). (**D**) Box plot of AUC scores per AbAg (*y* axis) compares AlphaFold, AbEpiScore-1.0, and AbEpiDockQ-1.0 at classifying CAPRI standard accuracy bin (defined by DockQ) structures (*x* axis): acceptable (≥0.23), medium (≥0.49), and high (≥0.8), containing 604, 359, and 72 AbAgs, respectively.

Table 1. A summary of evaluative metrics for random, AntiFold, ESMIF1, AlphaFold-2.3, and AbAgIoU accuracy models evaluated in nested cross-validation. The metrics included are PCCs between AbAgIoU and model scores, the Avg. AUC scores from Fig. 1C, and the AUC per (pr.) AbAg for different DockQ CAPRI accuracy categories (acceptable, medium, and high) from Fig. 1D. The best score within a metric category is marked in bold.

	PCC	Avg. AUC	Acceptable	Medium	High
Random	0.0108	0.5076	0.501 ± 0.192	0.512 ± 0.187	0.495 ± 0.140
Onehot-AbAgIoU	0.2168	0.6803	0.582 ± 0.293	0.606 ± 0.287	0.531 ± 0.245
ESM2-AbAgIoU	0.3765	0.7414	0.573 ± 0.299	0.608 ± 0.298	0.573 ± 0.264
AlphaFold-2.3	0.5602	0.8900	**0.748 ± 0.267**	0.879 ± 0.197	0.821 ± 0.243
AntiFold	0.5099	0.8309	0.699 ± 0.275	0.807 ± 0.237	0.760 ± 0.234
AntiFold-AbAgIoU	0.7483	0.9086	0.722 ± 0.273	0.868 ± 0.200	0.816 ± 0.212
ESMIF1	0.6613	0.8854	0.698 ± 0.282	0.834 ± 0.227	0.834 ± 0.176
AbEpiScore-1.0	**0.8036**	**0.9213**	0.743 ± 0.275	**0.892 ± 0.177**	**0.843 ± 0.181**

AbAgIoU threshold for labeling whether AbAg interfaces were accurate, we used 205 thresholds within the 0.0 to 0.5 range. At each threshold, structures with AbAgIoU above the value were labeled as accurate, and those below the value were labeled as inaccurate. These labels were used to compute an area under curve (AUC) score of a given prediction model, each represented by a point in the curves shown in Fig. 1C. An aggregated performance value was calculated as the average of the AUC scores, resulting in performance values of 0.8309 for AntiFold, 0.8854 for ESMIF1, and 0.8900 for AlphaFold-2.3. For comparison, random performance by drawing scores from a uniform distribution averaged 0.5076, indicating that these models significantly outperform random classification ($P < 0.001$, bootstrap; for details, see text S3) (Table 1).

To establish a baseline for sequence-based approaches, we trained and evaluated feedforward neural networks (FFNNs) using one-hot or evolutionary scale modeling (ESM2) encodings to predict AbAgIoU (*24*). For details on these models and their training, refer to Materials and Methods. These sequence-based models, Onehot-AbAgIoU and ESM2-AbAgIoU, achieved average AUCs of 0.6803 and 0.7414, respectively, falling short of AlphaFold, AntiFold, or ESMIF1. However, the FFNNs trained on AntiFold or ESMIF1 encodings both significantly outperformed AlphaFold-2.3, with average AUC scores of 0.9086 and 0.9213 ($P < 0001$, bootstrap; for details, see text S3) (Fig. 1C). We named the best-performing model that used the ESMIF1 encoding, AbEpiScore-1.0. This model's PCC value is also substantially improved compared to the performance using the raw ESMIF1 scores from 0.6613 to 0.8036 (Table 1). To further improve the model, we also tested adding AlphaFold confidence ranking as a feature. Although this feature-enhanced model showed slight gains in Avg. AUC (0.0028) and PCC (0.0131) over AbEpiScore-1.0, the improvements were minimal (see text S6).

AbAgIoU was found to correlate strongly with DockQ (see text S5), indicating that AbEpiScore-1.0 should generalize to the task of picking Critical Assessment of PRedicted Interactions (CAPRI) standard accurate structures (defined by DockQ) from each AbAg (*25*). To further quantify this, we labeled 30 individual structures for a given AbAg according to CAPRI standard DockQ accuracy thresholds: acceptable (≥0.23), medium (≥0.49), or high (0.8). AbAgs where all structures were labeled as either inaccurate or accurate at a given DockQ threshold were excluded in the analysis, as AUC cannot be computed for data with only one label. This evaluation revealed that all methods substantially outperformed random classification, with AbEpiScore-1.0 showing the highest performance in the medium and high categories and AlphaFold-2.3 having a slightly higher performance in the acceptable DockQ category. Furthermore, we found that the AbEpiScore-1.0 architecture can easily be adapted for the task of predicting the acceptable DockQ category only by replacing the AbEpiScore-1.0's output layer with two output neurons and training models to classify structures above or below a DockQ of 0.23 in the same nested cross-validation setup. We named this model AbEpiDockQ-1.0, which obtains identical performance to AlphaFold-2.3 in the acceptable DockQ category and near-identical performance to AbEpiScore-1.0 in the other categories (Table 1 and Fig. 1D).

Last, we showcase AbEpiScore-1.0 and AlphaFold-2.3 scores for 30 modeled structures of an antibody targeting insulin-like growth factor 2 [Protein Data Bank (PDB): 3KR3] (Fig. 2) (*26*, *27*). The ground truth antibody and antigen structures are highlighted in gray and black, while the modeled antibody structures are colored from blue to red based on predicted accuracy scores. Both methods assign higher scores to antibody structures aligning with the ground truth (gray area), but AbEpiScore-1.0 shows a much improved correspondence, achieving a PCC correlation of 0.8010 between AbAgIoU and predicted accuracy score, compared to AlphaFold's 0.2231. In addition, when assigning binary labels to 30 models (0 for DockQ < 0.23 and 1 for DockQ ≥ 0.23), AbEpiScore-1.0 classifies the eight modeled structures with DockQ ≥ 0.23, with an AUC of 0.903, whereas AlphaFold-2.3 does this with an AUC of 0.631.

To conclude, the results demonstrate that we can construct models to predict the accuracy of AbAg interfaces, with performance significantly improved beyond random. Among these, structure-based models demonstrate superiority, and our best model, AbEpiScore-1.0, which was fine-tuned on ESMIF1 encodings, significantly surpasses AlphaFold's intrinsic confidence ranking.

AbEpiTarget-1.0: Highly antibody-sensitive scoring for antibody target prediction

Our next aim was to develop a method to predict the antigen target of a given antibody, distinguishing modeled true AbAg complexes from those modeled with incorrect or "swapped" antibodies. We termed this task antibody target prediction. We hypothesized that interface scoring models (AlphaFold-2.3, ESMIF1, AntiFold, X-AbAgIoU, and AbEpiScore-1.0) could classify AbAgs modeled with the correct antibody based on the expectation that accurate interfaces are easier to model with the correct antibody.

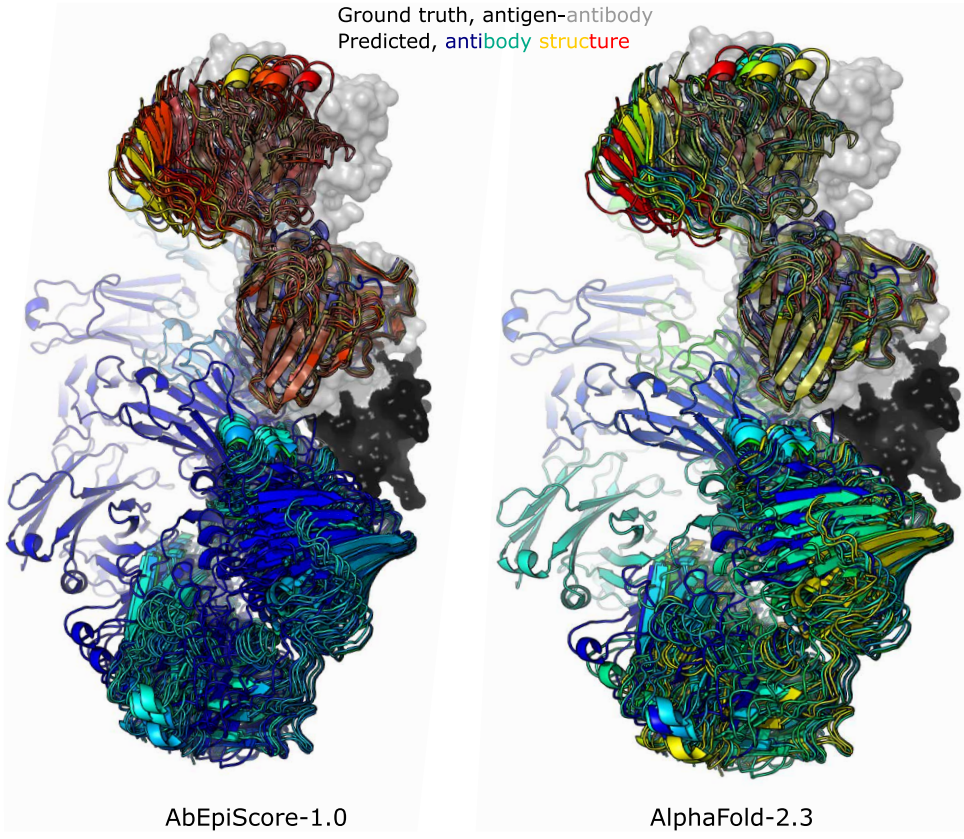

Fig. 2. We compare AbEpiScore-1.0 and AlphaFold-2.3 at scoring modeled structures of an antibody targeting insulin-like growth factor 2. Crystal of insulin-like growth factor 2 (black) bound to an antibody (gray) (PDB: 3KR3). Modeled antibody structures have been colored from low (blue) to high (red) according to AbEpiScore-1.0 (left; PCC, 0.8010) and AlphaFold-2.3 (right; PCC, 0.2231).

To test this, we created 1730 groups of AbAg complexes, each containing one true AbAg and three swapped AbAgs, all modeled with the same antigen. For each AbAg, AlphaFold-2.3 generated 30 structures, resulting in 51,900 true and 155,570 swapped AbAg structures (for details, see Materials and Methods). Using each interface score model, we selected the maximum scoring structure from the 30 structures per AbAg, yielding four scores (and structures) for each antigen group: one true AbAg score and three swapped AbAg scores. We then calculated the percentage of groups where the true AbAg structure ranked first, second, third, or fourth. The percentage where the true AbAg was ranked first was termed rank-1 accuracy (%). In addition, we computed the average ranking of true AbAg across all antigen groups ["Avg. True Rank (1–4)"] (Table 2 and Fig. 3A).

For comparison, random performance was simulated by assigning scores to structures from a uniform distribution, resulting in a baseline rank-1 accuracy of 23.12%. Models based on sequence-based interface scoring models, such as Onehot-AbAgIoU and ESM2-AbAgIoU, performed near-random with rank-1 accuracies of 25.84 and 25.26%, respectively. In contrast, structure-based models substantially outperformed random, achieving rank-1 accuracies between 31.33 and 42.08% and averaged true rank scores of 2.3283 to 2.1260, with AlphaFold-2.3 being the most accurate (Table 2 and Fig. 3A). To improve performance, we trained an FFNN classifier to discriminate between true and swapped AbAgs. This was done in a nested cross-validation setup (see Materials and Methods for details). Models using one-hot, ESM2, or ESMIF1 encodings outperformed their interface score model counterparts, achieving rank-1 accuracies of 35.90 to 61.21% and averaged true rank scores of 2.3075 to 1.7543. The best model, using ESMIF1 encodings, was named AbEpiTarget-1.0 and significantly outperformed AlphaFold-2.3 ($P < 0.001$; details in text S3). As both AlphaFold-2.3 and AbEpiScore-1.0 demonstrated performance in distinguishing true from swapped AbAgs, we explored adding these scores as features in AbEpiTarget-1.0 models. For one-hot and ESM2 encodings, this improved rank-1 accuracy by ~5 to 10% and boosted PCC. However, the AbEpiTarget-1.0 ESMIF1–based model only improved when AlphaFold-2.3 scores were added, and by less than 1%, suggesting it already captured an interface accuracy bias (Table 2 and Fig. 3A).

We expected that it is generally more difficult to place the antibody correctly for larger antigens. To examine this, we categorized AbAg groups by antigen size and recalculated rank-1 accuracy, and, as expected, AlphaFold-2.3's accuracy declined with increasing antigen size, from 50.26% for the smallest category (<250 residues) to 19.23% for the largest category (1000 to 1500 residues), which is lower than expected random performance. In contrast, however, AbEpiTarget-1.0 remained consistent, with accuracies from 50.79 to 69.64%, independent of antigen size (Fig. 3B). We next checked for performance biases by antigen type, where AbEpiTarget-1.0 consistently outperformed AlphaFold-2.3 across all antigen types. In addition, here,

Table 2. We evaluated 1730 and 5190 true and swapped AbAgs in a nested cross-validation using random, AlphaFold-2.3, ESMIF1, AntiFold, X-AbAgIoU, and AbEpiTarget-1.0 models. The evaluative metrics were modeled structures' AbAg interface score correspondence to AgIoU values (true PCC and swap PCC) and how well true AbAgs are scored compared to swapped AbAgs modeled with the same antigen [rank-1 accuracy (%) and Avg. True Rank (1 to 4)]. The best score within a metric category is marked in bold.

	True PCC	Swap PCC	Rank-1 (%)	Acc. Avg. True Rank (1–4)
Random	0.0131	0.0121	23.12	2.5705
AlphaFold-2.3	0.5407	**0.2541**	42.08	2.1260
Onehot-AbAgIoU	0.2100	0.0788	25.84	2.5121
AbEpiTarget-1.0 (Onehot)	0.1563	0.0422	35.90	2.3075
AbEpiTarget-1.0 (Onehot + AlphaFold-2.3)	0.4377	0.1513	46.71	2.0728
AbEpiTarget-1.0 (Onehot + AbEpiScore-1.0)	0.5273	0.0296	48.32	2.0711
ESM2-AbAgIoU	0.3527	0.1410	25.26	2.5387
AbEpiTarget-1.0 (ESM2)	0.0621	0.0060	49.08	2.0861
AbEpiTarget-1.0 (ESM2 + AlphaFold-2.3)	0.1689	0.0428	51.85	2.0064
AbEpiTarget-1.0 (ESM2 + AbEpiScore-1.0)	0.2703	0.0020	54.34	1.941
AntiFold	0.3871	0.0504	31.33	2.3283
AntiFold-AbAgIoU	0.5715	0.0844	38.50	2.2145
ESMIF1	0.5121	0.0687	38.03	2.2289
AbEpiScore-1.0	**0.6010**	0.0896	39.97	2.1740
AbEpiTarget-1.0	0.4533	−0.070	61.21	1.7543
AbEpiTarget-1.0 (+AlphaFold-2.3)	0.4569	−0.063	**61.73**	**1.7405**
AbEpiTarget-1.0 (+AbEpiScore-1.0)	0.4620	−0.079	60.58	1.7757

AbEpiTarget-1.0 displayed a consistent and high performance across antigen types in contrast to AlphaFold-2.3, where the performance variation was found to be more than 20 percentage points between the lowest [severe acute respiratory syndrome (SARS)] and highest (cancer) scoring antigen types (Fig. 3C).

When designing immunotherapies, antibody groups evaluated for potential antigen binding will typically consist of more than four antibodies. To evaluate antibody target prediction in larger antibody groups, we adapted a benchmark from Kilambi and Gray (*28*), who assessed 17 AbAgs, from antigen groups, each containing the true AbAg—featuring the correct antibody and antigen—and 16 swapped AbAgs, made by pairing the correct antigen with incorrect antibodies from each of the other AbAgs. Their method consisted of superimposing homology-modeled antibodies and unbound antigen structures onto the corresponding AbAg crystal complexes and then using Rosetta's docking interface score to distinguish true and swapped AbAg interfaces (*29*). As some true AbAgs in their dataset shared the same antigen, we selected 11 AbAgs structures with unique antigens supplemented with six randomly selected additional AbAgs targeting different unique antigens. Next, we evaluated the performance of AlphaFold-2.3 and AbEpiTarget-1.0 for identifying the true AbAgs within each antigen group, obtaining rank-1 accuracies of 35.29 and 47.06%, respectively, notably higher than an expected random accuracy of 5.88% (1 of 17) and the 11.76% accuracy for Rosetta in their study (for details, see text S8). Although AbEpiTarget-1.0's performance declined compared to our analysis with four antibody groups, this was expected, as antibody target prediction becomes more challenging as the number of antibodies increases.

In the "AbEpiScore-1.0: Improved prediction of AbAg interface accuracy" section, we showed that more accurately modeled true AbAgs receive higher AbAg interface scores. Here, we aimed to demonstrate that the same can generally not be said for swapped AbAgs. We assessed this by measuring the correspondence and distribution of model scores and Antigen Intersection over Union (AgIoU) values for all 51,900 true and 155,570 swapped AbAg structures. AgIoU measures the match between predicted epitope residues and their ground truth crystal structure epitopes. With the exception of AbEpiTarget-1.0 (ESM2), all models generally scored true AbAgs higher when predicted epitopes closely matched the crystal structure, with PCCs ranging from 0.1563 to 0.6010, where AbEpiScore-1.0 had the highest correlation. This affirms that predicted accuracy reflects actual epitope accuracy. In contrast, for swapped AbAgs, correlations were weak, with PCCs from −0.079 to 0.2541, meaning that few swapped interfaces are both predicted and actually accurate (Table 2 and Fig. 4).

In Fig. 4, we illustrate this analysis for AbEpiTarget-1.0 and provide the same 2D hexagon plots for all Table 2 models in text S9. This plot clearly illustrates that the highest scores and the most accurate placements of the antibody onto the antigen (AgIoU) require structural modeling with the true antibody and not a swapped antibody (Fig. 4, A and B, top right corner). We quantify this by displaying percent-wise score distributions across 25 bins using 2D histograms. While ~5% of true AbAgs had AgIoU and AbEpiTarget-1.0 scores above 0.6, only 0.1% of swapped AbAgs fell into this category (Fig. 4, C and D). Moreover, when analyzing which bins were overrepresented by true and swapped AbAg structures, we found that bins on the top right side of the diagonal, with high scores and AgIoU values, were overrepresented by true AbAg structures (Fig. 4E).

To conclude, these results demonstrate that structure-based AbAg interface scoring models, particularly AlphaFold-2.3 and AbEpiScore-1.0,

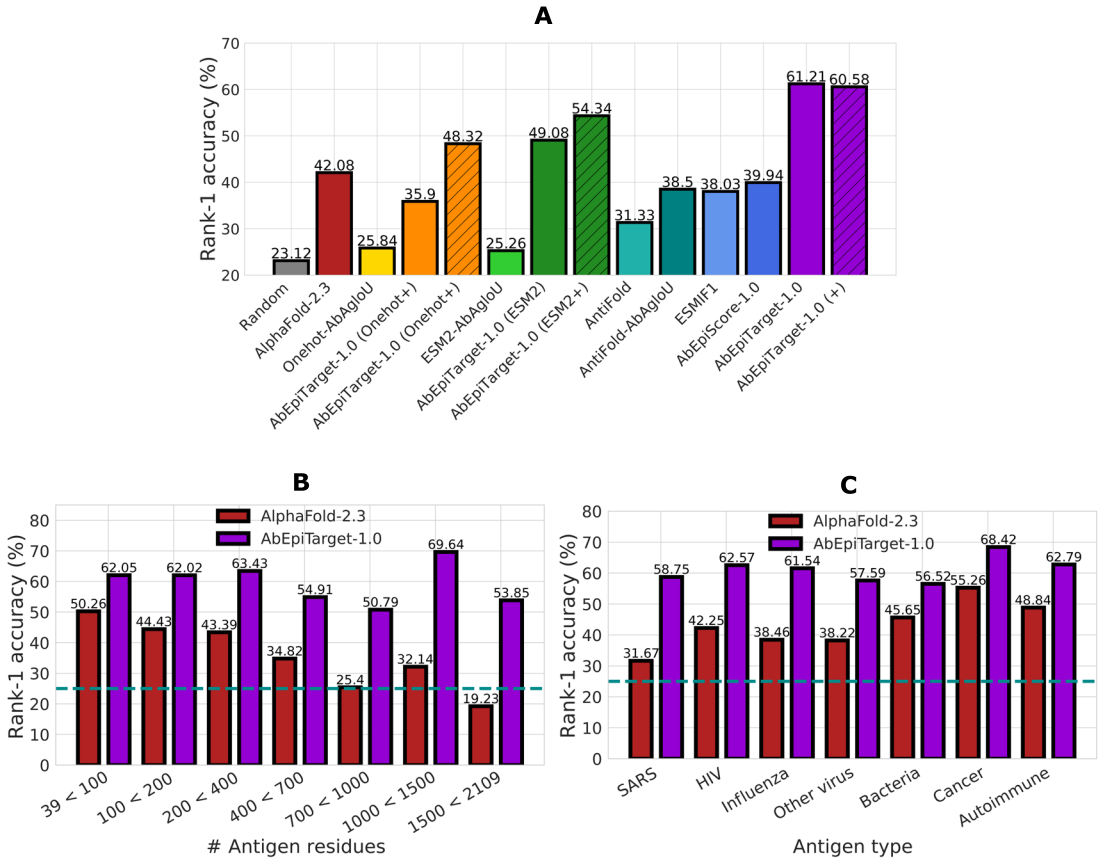

Fig. 3. We measure the model's ability to identify AbAg structures modeled with the correct antibody and antigen from those modeled with the incorrect antibody. (**A**) Rank-1 accuracy (%) for 1730 groups of true and swapped AbAgs modeled with the same antigen. The *x* axis shows the model score used, with (+) indicating models incorporating AbEpiScore-1.0 as an additional feature. (**B**) The groups were categorized by antigen sizes based on residue count (*x* axis), and the rank-1 accuracy for AlphaFold-2.3 and AbEpiTarget-1.0 computed (*y* axis). There were 195, 682, 484, 224, 63, 56, and 26 antigen groups in these categories from left to right. A dashed line indicates random performance (25%). (**C**) The groups were categorized by antigen type (*x* axis) and the rank-1 accuracy for AlphaFold-2.3 and AbEpiTarget-1.0 computed (*y* axis). The antigen types and number of groups in each category were SARS (240), HIV (187), influenza (117), other virus (191), bacteria (46), cancer (114), and autoimmune (43). "Other virus" includes malaria, dengue, Zika, hepatitis, and herpes viruses.

can effectively distinguish between true and swapped AbAgs. By training models specifically to classify true AbAgs from swapped AbAgs, AbEpiTarget-1.0 demonstrates major enhanced performance. Furthermore, while AlphaFold-2.3's confidence ranking appears sensitive to antigen size, AbEpiTarget-1.0 demonstrates robust performance across different antigen sizes.

AbEpiTope-1.0 maintains performative edge on post–AlphaFold-2.3 training data

AlphaFold-2.3 was trained on existing solved AbAg structures, improving modeling quality and, therefore, epitope or antibody target prediction for these AbAgs compared to complexes without known structures, which are more typical in real-world applications. To better estimate performance in such scenarios, we compared AlphaFold-2.3 and AbEpiTope-1.0 on AbAgs with structures released before and after AlphaFold-2.3's training cutoff of 30 September 2021. From our dataset of 1730 AbAgs, we created a subset called "Before," consisting of 1529 AbAgs released before this date. For the remaining 201 AbAgs, we excluded those with antigen or antibody sequences sharing ≥65% Ag or ≥95% Ab Many-against-Many sequence searching (MMseqs2) sequence identity with any AbAg in the Before set, reducing this subset to 109 AbAgs, termed "After."

Next, we evaluated AlphaFold-2.3 and AbEpiScore-1.0's on classifying acceptable structures (0.23 ≥ DockQ) per AbAg. We excluded AbAgs with all structures below this threshold, as AUC cannot be computed for data with one label. Although AlphaFold-2.3 obtained a slightly better average AUC over AbEpiScore-1.0 on the Before data, AbEpiScore-1.0 performs substantially better on the After data. Both models performed considerably worse on the After data, indicating that classification of acceptable quality AbAgs is more difficult for novel antibody targets (Fig. 5A).

Next, we compared AbEpiTarget-1.0 and AlphaFold-2.3 on classifying the true AbAg structure in antigen groups containing three other AbAg structures constructed with swapped antibodies, as described in the previous section. We note that swapped AbAgs in the After data may contain antibodies from the Before data, as no date constraint was applied when constructing swapped AbAgs. We found that the rank-1 accuracy, which is the frequency for which the true AbAg is ranked first across all antigen groups, was 42.38 and 44.04% for AlphaFold-2.3 on the Before and After data, respectively.

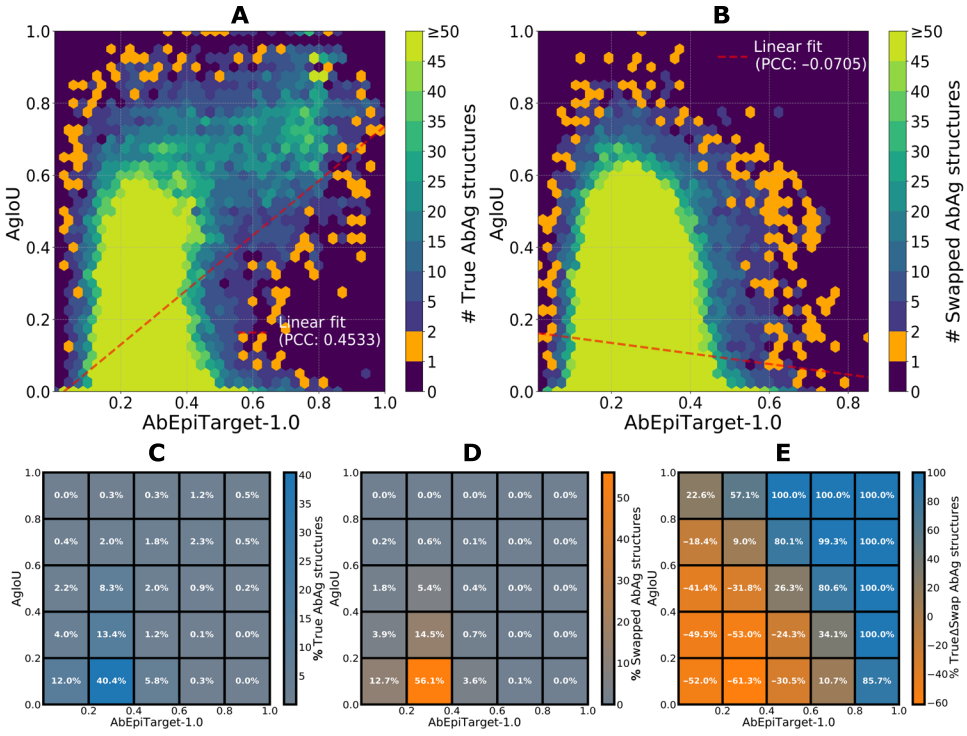

Fig. 4. We compare the AbEpiTarget-1.0 scores and ground truth epitope accuracy of AbAg structures modeled with the correct antibody and antigen (true AbAg) against those of structures modeled with an incorrect antibody (swapped AbAg). (**A**) Min-max–scaled AbEpiTarget-1.0 scores (*x* axis) for 51,900 true and (**B**) 155,570 swapped AbAg structures were plotted against corresponding AgIoU values (*y* axis) in hexagonal bins. Color scales capped at 50 structures show the structure count per bin, and orange indicates single structure bins. Red dashed lines indicate linear fits computed across all true or swapped AbAg structures. All true (**C**) and swapped (**D**) AbAg structures were placed into 25 square bins indicated by the black boundaries based on AbEpiTarget-1.0 scores and AgIoU values, with percentages and color scale indicating the distribution. (**E**) We compute a percentage score, TrueΔSwap (see Eq. 2 in Materials and Methods), indicating which bins the true or swapped AbAg structures are overrepresented. This score ranges from −100% (only swapped structures were counted) to 100% (only true structures were counted).

AbEpiTarget-1.0 obtains considerably higher scores of 62.38 and 55.96% on these data (Fig. 5B).

In the "AbEpiTarget-1.0: Highly antibody-sensitive scoring for antibody target prediction" section, we observed exceptionally high-scoring true AbAgs, with few swapped AbAgs reaching similar levels, indicating that these scores were only achievable with the correct antibody. Models such as AbEpiScore-1.0, AbEpiTarget-1.0, and AlphaFold-2.3 were able to make this distinction (Fig. 4, A and B, and text S9). This indicates that, for some antigen groups, true AbAgs can be easily ranked, as it is unlikely to construct a higher-scoring AbAg interface with a swapped antibody. To further test this observation and whether it was driven by overlaps with the AlphaFold training data, we recorded the max scores for each antigen group and ranked them from highest to lowest. We computed true rank scores (Eq. 3), measuring the rank of the true AbAg within each group, ranging from 0 (worst possible rank) to 1 (best possible rank). As seen in earlier evaluations, both AlphaFold-2.3 and AbEpiTarget-1.0 ranked true AbAgs above swapped ones far better than random scoring. The plots show a near-perfect prediction plateau (true rank score, ≥0.95), which is more stable for AbEpiTarget-1.0, maintaining ≥0.95 until 277 (18% of the antigen groups) and 13 (12% of the antigen groups) antigen groups in both datasets are reached. In contrast, AlphaFold-2.3 drops below this threshold after 44 and 4 antigen groups (Fig. 5, C and D). This suggests that when AbEpiTarget-1.0 assigns high confidence to an AbAg pair, its ranking is substantially more reliable than that of AlphaFold-2.3. To conclude, although these results show that AbEpiTarget-1.0's performance declines for targets released after AlphaFold-2.3's training cutoff date, it still substantially outperforms AlphaFold-2.3 on these more challenging, unseen data.

Web server and software package

For the computational tool, we used the AbEpiScore-1.0 and AbEpiTarget-1.0 models, which rely only on input from ESMIF1. Users can upload single or multiple files (in .zip format) to the web server (https://services.healthtech.dtu.dk/services/AbEpiTope-1.0/), with an adjustable angstrom distance to define AbAg interfaces (default, 4 Å) (Fig. 6). Each structure file must include at least one antibody chain (light, heavy, or both) and one or more antigen chains.

Because of server limits, users can only upload up to 100 files per submission. For larger batches, the local version, available via GitHub (https://github.com/mnielLab/AbEpiTope-1.0), accepts individual files or entire directories. The tool generates three output files: two CSV files and one FASTA file. The first CSV file (output.csv) contains AbEpiScore-1.0 and AbEpiTarget-1.0 scores for each structure. The second CSV file (interface.csv) lists the epitope and paratope residues used for scoring. The FASTA file contains sequences of the AbAg complexes, with chains separated by colons (":") and headers formatted as ">filename_chainids." In addition, a failed_files.csv file lists any input files that failed to process, detailing

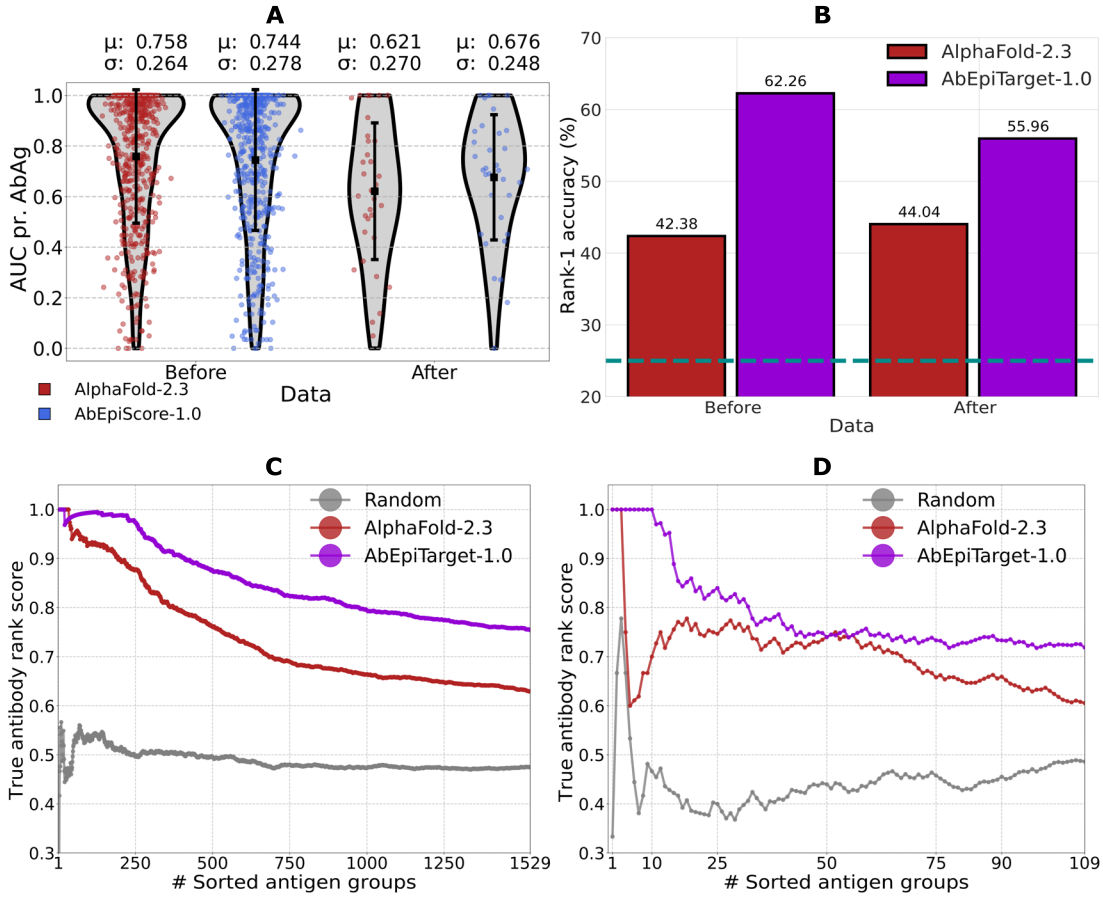

Fig. 5. Performance evaluation of AlphaFold-2.3 and AbEpiTarget-1.0 on AbAg subsets released before and after AlphaFold-2.3's training date (Before and After). (**A**) A violin scatter plot comparing model performance in classifying structures with DockQ ≥ 0.23 per AbAg. AUC scores were computed for 547 of 1529 (Before) and 52 of 109 (After) AbAgs (*y* axis). (**B**) Rank-1 accuracy (%) (*y* axis) for both subsets comparing AlphaFold-2.3 and AbEpiTarget-1.0, with a dashed line indicating random performance (25%). (**C**) Before antigen groups were ranked from highest to lowest based on their maximum score, whether from the true AbAg or one of three swapped AbAgs. True rank scores (0 = worst ranking, 1 = perfect ranking; see Eq. 3) were computed for all antigen groups. Then, average true rank scores were computed (*y* axis) as more antigen groups were included in this average along the *x* axis. (**D**) The same analysis as in (C) but for the After data.

issues such as invalid format, absence of an AbAg complex, or no interface detected within the set angstrom threshold. The web server also provides a sortable table of results, allowing users to sort by AbEpiScore-1.0 or AbEpiTarget-1.0 scores and download the sorted table as a CSV file. Last, in text S10, we provide a guidance section on how to use outputted AbEpiTope-1.0 scores for predicting modeled AbAg accuracy and antibody screening.

DISCUSSION

AlphaFold has changed the field of modeling proteins and their interactions, but since its initial release, it has consistently been challenging to model AbAg complexes. The challenge comes from the flexible and diverse CDR regions of antibodies, shaped by V(D)J recombination, which complicate modeling their binding conformation. In addition, because this diversity is not a product of protein evolution, the AbAg complex multiple sequence alignment (MSA) input is less informative for AlphaFold. Despite these challenges, releases of AlphaFold have shown progressive improvements for modeling AbAg interactions. With the recent emergence of AlphaFold-3, Boltz-1, and Chai-1 (*30*–*32*), which promise more accurate AbAg interface prediction, studies that evaluate their efficacy on downstream applications are crucial.

Here, we used AlphaFold-2.3 to generate complexes for AbAgs and evaluated various approaches to assess the accuracy of these structures, with the aim of developing a method for antibody-specific epitope prediction. In this goal, we first demonstrated that AlphaFold's confidence scoring metric (0.8 ipTM + 0.2 pTM) can be used to evaluate the accuracy of AbAg interfaces. Pretrained inverse folding models, particularly ESMIF1 and AntiFold, also performed well in this task without fine-tuning, with ESMIF1 showing superior performance. Further improvements were achieved by fine-tuning both models specifically for AbAg interface evaluation. Our best model, based on fine-tuned ESMIF1 and named AbEpiScore-1.0, consistently outperformed AlphaFold's confidence scoring in all evaluations.

We hypothesized that AbAg interfaces modeled with the correct antibody (true AbAg) would be more accurate than those modeled with incorrect or swapped antibodies (swapped AbAg), and, thus, our AbAg interface accuracy models should be able to distinguish

Input Page

Submission

1. Upload a single structure file (.pdb/.cif) or a .zip file.

Choose File Cancer.zip

A .zip file may contain a maximum of 120 structure files. Each structure must contain an antibody-antigen complex, that includes an antibody chain (light, heavy or both) along with any number of antigen chains. Structure files where this is not detected will not produce a score.

Example Files

We provide five example .zip files containing modeled structures of antibodies targeting antigens from a range of diseases: SARS, HIV, Pseudomonas aeruginosa (Bacteria), the PD-1 receptor (Cancer), and grass pollen (Autoimmune). Each file includes 30 structures in PDB format generated using AlphaFold-2.3. Additionally, a sixth example .zip file contains modeled structures for four different antibodies and a SARS antigen. Each antibody and the SARS antigen was modeled separately, not simultaneously. One of these antibodies has been experimentally confirmed to target the SARS antigen, while the others are known to target different antigens. The .zip file includes a total of 120 individually modeled structure files, all in PDB format.

SARS | HIV | Bacteria | Cancer | Autoimmune | Antibody Target Prediction

2. Set antibody-antigen interface distance (default: 4Å):

4

Submit Clear fields

Output Page

Antibody-Antigen Interface Scores

Click the 'AbEpiScore-1.0' or 'AbEpiTarget-1.0' header to sort rows.

Download Sorted Scores (.csv)

FileName	AbEpiScore-1.0	AbEpiTarget-1.0
7e9b_ag_C_ab_L_H_unrelaxed_rank_004_alphafold2_multimer_v3_model_4_seed_000_2024.pdb	0.252708	0.306679
7e9b_ag_C_ab_L_H_unrelaxed_rank_005_alphafold2_multimer_v3_model_4_seed_004_2024.pdb	0.235554	0.283597
7e9b_ag_C_ab_L_H_unrelaxed_rank_002_alphafold2_multimer_v3_model_3_seed_003_2024.pdb	0.2206	0.239715
7e9b_ag_C_ab_L_H_unrelaxed_rank_003_alphafold2_multimer_v3_model_3_seed_001_2024.pdb	0.211874	0.259501
7e9b_ag_C_ab_L_H_unrelaxed_rank_007_alphafold2_multimer_v3_model_4_seed_005_2024.pdb	0.196571	0.253529
7e9b_ag_C_ab_L_H_unrelaxed_rank_006_alphafold2_multimer_v3_model_4_seed_001_2024.pdb	0.187321	0.239073
7e9b_ag_C_ab_L_H_unrelaxed_rank_001_alphafold2_multimer_v3_model_3_seed_000_2024.pdb	0.173459	0.227502
7e9b_ag_C_ab_L_H_unrelaxed_rank_008_alphafold2_multimer_v3_model_4_seed_002_2024.pdb	0.143712	0.198975
7e9b_ag_C_ab_L_H_unrelaxed_rank_019_alphafold2_multimer_v3_model_5_seed_005_2024.pdb	0.036155	0.246668
7e9b_ag_C_ab_L_H_unrelaxed_rank_026_alphafold2_multimer_v3_model_2_seed_004_2024.pdb	0.028937	0.199654
7e9b_ag_C_ab_L_H_unrelaxed_rank_021_alphafold2_multimer_v3_model_2_seed_003_2024.pdb	0.027727	0.202305

Fig. 6. Screenshots of the input and output pages for AbEpiTope-1.0 web server. The input page allows users to upload a single AbAg complex in PDB or Crystallographic Information File (CIF) file format as well as multiple complexes in a .zip file, with example .zip files provided for SARS, HIV, *Pseudomonas aeruginosa*, PD-1 receptor, grass pollen, and a SARS antigen with four modeled antibodies (one experimentally confirmed to target SARS and three others targeting different antigens). Each .zip file contains 30 structures made with AlphaFold-2.3. Users can also set the angstrom distance to define AbAg interfaces (default, 4 Å). The output page provides a downloadable .zip file of all results and a table that can be sorted by AbEpiScore-1.0 or AbEpiTarget-1.0 scores and exported as a .csv file.

between them. By grouping the true and three swapped AbAgs modeled with the same antigen, we found that AlphaFold-2.3 and AbEpiScore-1.0 consistently scored true AbAgs higher than swapped AbAgs, outperforming random assignment. More broadly, there was a strong positive correlation between predicted accuracy and the match between predicted and actual epitopes (AgIoU), affirming that predicted accuracy reflects actual epitope accuracy. This correlation was substantially weaker for swapped AbAgs, indicating that few swapped interfaces are both predicted and actually accurate. In summary, if an interface is predicted to be highly accurate, then it is likely constructed with the correct antibody.

After this observation, we developed models specifically designed to differentiate true from swapped AbAg interfaces. These models outperform their AbAg interface accuracy counterparts in the task of ranking true AbAgs higher than swapped AbAgs modeled with the same antigen. Our best model, AbEpiTarget-1.0, using ESMIF1 as input, ranked true AbAgs higher than swapped AbAgs for 61.21% of cases, outperforming AlphaFold-2.3 ranking of 42.08%. In addition, we find that the AbAg structures with the highest scores and most accurate antibody placements on the antigen require structural modeling with the true antibody and not a swapped antibody. This suggests that further improvements on the structural modeling of true AbAg complexes would greatly benefit our findings. This could potentially be addressed using more refined variants of AlphaFold such as Chai-1, where explicit AbAg structural constraints can be included. We argue that this work marks a milestone in antibody-specific B cell epitope prediction tools. Current tools, although incorporating antibody input, generally do not evaluate antibody specificity by modeling swapped antibodies (*33*, *34*). To the best of our knowledge, no prior work has convincingly demonstrated that predicted AbAg interface accuracy can be used to identify antibodies best suited for a given antigen.

When designing immunotherapies, antibody groups evaluated for potential antigen binding will typically consist of more than four antibodies, as done in this work. A more ideal analysis for evaluating antibody target prediction should include at least 15 antibodies, all targeting different unique antigens. Unfortunately, our available compute resources prohibited us from doing this. To approximate this analysis, we created structures for a smaller expanded dataset of 17 AbAgs. Each antigen group contained the true AbAg—comprising the correct AbAg pair—along with 16 swapped AbAgs, where the correct antigen was paired with incorrect antibodies from the other AbAgs. As expected, performance declined in this setting because antibody target prediction naturally becomes more challenging when more antibodies are evaluated because of increased random variability. Despite this, AbEpiTarget-1.0 maintained strong predictive performance, ranking the true AbAg first in 47.06% of cases, compared to AlphaFold-2.3's 35.29%.

A cutoff date for AlphaFold-2.3 structure modeling was not used (specifically the AlphaFold-2.3 multimer training cutoff date of 30 September 2021), improving modeling quality compared to complexes without known structures, which are more typical in real-world applications. Applying date filtering and sequence identity filtering left us with too few AbAgs to build robust models, and we therefore chose not to use such a filter. To assess performance in cases where no known AbAg structures were available, we created an independent test dataset consisting of AbAgs with solved structures released after the training cutoff date and with no sequence identity to AbAgs published before the cutoff. As expected, both AlphaFold-2.3 and AbEpiTarget-1.0 showed performance declines, confirming that predicting novel antibody targets is indeed more challenging. Despite this decline, AbEpiTarget-1.0 remained substantially superior on the independent test data, ranking the true AbAg first in 55.96% of cases compared to AlphaFold-2.3's 44.04%. In addition, when AbEpiTarget-1.0 was restricted to making predictions only when highly confident, it achieved near-perfect accuracy in 11% of cases, outperforming AlphaFold-2.3's 3.6%.

To conclude, we present AbEpiTope-1.0, a tool for researchers who use structure prediction tools to do antibody-specific B cell epitope prediction, which is of primary medical and societal importance, such as vaccine development and personalized treatment strategies. The tool is available as a web server and a stand-alone package, making it easy to use by experts and nonexperts alike.

MATERIALS AND METHODS

Structural data

We extracted all crystal structures from the PDB deposited on the biological assembly FTP server before 2 February 2023. Structures were filtered to include at least one antibody containing both light and heavy chain variable domains and one nonantibody (antigen) protein chain with at least 40 residues. Antibody heavy and light chains were identified using BCR hidden Markov models developed by LYRA (*35*). After refining our search to include only structures with a resolution lower than 3.5 Å and an *R* factor below 0.26, 1735 PDB entries remained. From these, we identified the CDRs using AbRSA (*36*). AbAg complexes were defined as having one light and one heavy chain, with heavy chain CDR3 (HCDR3) residues within 4 Å of at least one antigen residue heavy atom (main chain or side chain). This HCDR3 filtering excluded complexes where the antibody constant region was targeted by another functional protein, such as antibody-binding Protein M (Protein M TD) (PDB: 4NZT) (*37*). Many PDBs contained multiple AbAg complexes, resulting in a comprehensive catalog of 2990 AbAg complexes.

Redundancy reduction, data partitioning, and picking antibody swaps

The dataset of 2990 AbAg complexes comprises 10,566 sequences, with 4628 from antigens and 5938 from antibodies. We used MMseqs2 easy search for an all-versus-all sequence alignment, identifying 1,939,885 significant matches with the parameters -e 0.1 (expected value), --min-seq-id 0.2 (sequence identity), --cov-mode 0 (coverage mode), and -c 0.85 (coverage) (*38*). The dataset was then reduced to 1932 AbAgs by randomly selecting representatives from "clusters" where both light and heavy chains, as well as all antigen chains, shared more than 99% MMseqs2 sequence identity.

Next, we partitioned the AbAgs into groups according to sequence identity using a graph-based clustering algorithm built on the Python package NetworkX (*39*), ensuring that no groups shared more than 65% antigen or 95% light and heavy chain identity. This process removed 277 AbAgs, resulting in a final dataset of 1733. These were distributed into five groups of approximately equal size (three groups of 347 and two groups of 346). A detailed description of this algorithm and the data processing pipeline is provided in text S1.

We term these 1733 AbAg complexes consisting of the true antigen and antibody "True AbAg." For each antigen in the true AbAg complexes, we constructed at least three antibody-swapped complexes. To avoid data leakage, swaps were generated within each of the five data partitions. Furthermore, to avoid swapping with similar antibodies, we randomly pooled antibodies and antigens together, where none of the sequences in their respective AbAg complexes was a significant match in the initial MMseqs2 easy search.

AbAg structure prediction

Structures for AbAg complexes were predicted using AlphaFold-2.3 ColabFold version (*40*). This tool takes three inputs: protein sequences, an MSA of evolutionarily related sequences, and an optional template structure. For our study, we excluded the template structures and modeled the 1733 true AbAg complexes and 5348 swapped AbAg complexes using only protein sequences and MSAs as inputs. MSAs were created using the MMseqs2 implementation found within ColabFold.

Initially, we modeled the true AbAg complexes with a single seed across all five AlphaFold-2.3 models, generating five structures per complex. We then extended the modeling by running each AlphaFold model with six different seeds, producing an additional 30 structures per complex. However, three complexes with large cytochrome C antigens (PDB: 3CXH, 3CX5, and 1KYO) were computationally intensive, with only a few structures placing the antibody within 4 Å of the antigen (*41*, *42*). Consequently, we were unable to complete their modeling, leading to the exclusion of these three AbAgs from the dataset. In addition, because the AbAg modeling was automated on a computer server, some AbAgs had more structures than planned due to multiple model runs when jobs did not finish before reaching the walltime limit. The distribution of generated structures and those successfully placing the antibody within 4 Å of the antigen is described in text S2. Ultimately, we modeled at least 35 structures for each of the 1730 AbAg complexes.

For each of these true AbAg complexes, we modeled antibody-swapped complexes, picking swaps as described above. Structures were predicted using AlphaFold-2.3 ColabFold with six seeds, generating 30 structures for each swapped AbAg. Three swaps were generated for 1677 antigens, six swaps for 52 antigens, and five swaps for 1 antigen. The extra swaps for some antigens resulted from a preliminary study conducted before scaling up to model the entire dataset.

Data encodings and features

The predicted AbAg structure interface residues, defined as residues on the antibody and antigen with heavy atoms within 4 Å of each other, were encoded using sequence-based and structure-based methods. For sequence encodings, we used both one-hot and numeric embeddings from the ESM2 protein language model. For one-hot encoding, each residue was represented by a 21D vector (20 amino acids plus a padding token for handling residues near the sequence start or end). One-hot encodings for each residue were created by concatenating the encodings of the residue and its eight neighboring residues (four on each side), resulting in a vector of size 189 (9×21) for each residue. ESM2 embeddings were generated by processing the antigen and antibody sequences through the pretrained ESM2 transformer, resulting in a 1280D vector representation for each residue.

Structure-based encodings were obtained using inverse folding GVP-Transformers, ESMIF1, and AntiFold. Inverse folding seeks to recover the amino acid sequence compatible with a given protein's 3D structure. ESMIF1 is a deep learning model developed by Meta's Fundamental AI Research and trained on a large dataset of 12 million AlphaFold and 16,000 CATH structures. AntiFold is a fine-tuned version of ESMIF1, specifically for antibody sequence recovery. Both models output a per-residue probability score for the likelihood of a residue given the protein fold and a 512D vector representation containing sequence and structural information. We collected these probabilities and encodings for all predicted AbAg structure interface residues. Last, we gathered AlphaFold-2.3's intrinsic confidence rankings (ipTM and pTM) for all AbAg structures.

Performance metrics and labels

Ground truth epitope residues in the crystal structures were labeled as any residue with at least one heavy atom (main chain or side chain) within 4 Å of any light or heavy chain. The corresponding residues on the light or heavy chain were labeled as paratope residues. Epitopes and paratopes predicted by AlphaFold-2.3 were defined in the same manner. To evaluate the correspondence of AbAg interfaces predicted by AlphaFold-2.3 to crystal AbAg interfaces, we used two metrics: AbAgIoU and AgIoU. AbAgIoU is calculated as the intersection divided by the union of ground truth epitope or paratope residues (True_{res}) and predicted epitope or paratope residues (Pred_{res}) (Eq. 1). AgIoU is computed similarly, focusing only on antigen residues. These metrics differ from the standard fraction of native contacts (Fnat) by penalizing both missing and extra predicted contacts, whereas Fnat only penalizes missing contacts. For comparison, we also computed DockQ scores for all true AbAg structure interfaces.

$$\text{AbAgIoU} = \frac{\text{True}_{\text{res}} \cap \text{Pred}_{\text{res}}}{\text{True}_{\text{res}} \cup \text{Pred}_{\text{res}}} \tag{1}$$

To compare the distribution of AbAgIoU and predicted AbAg interfaces scores between true and swapped AbAg structures, we used the percentage metric TrueΔSwap, quantifying the difference in their counts (Eq. 2). $\text{True}_{\text{count}}$ represents the number of true AbAg structures, while $\text{Swap}_{\text{count}}$ represents the number of swapped AbAg structures. The TrueΔSwap score ranges from −100% (only swapped structures counted) to 100% (only true structures counted).

$$\text{True}\,\Delta\,\text{Swap} = \frac{\text{True}_{\text{count}} - \text{Swap}_{\text{count}}}{\text{True}_{\text{count}} + \text{Swap}_{\text{count}}} \cdot 100 \tag{2}$$

Last, to evaluate the ability of AbAg interface scores to rank true AbAg structures—featuring the correct antibody and antigen—within antigen groups containing other AbAg structures constructed with swapped or incorrect antibodies, we used a ranking metric termed true rank score (Eq. 3). $\text{True}_{\text{Rank}}$ represents the ranking position of the true AbAg, which ranges from 1 to N, where N is the total number of AbAg structures in the antigen group, including the true AbAg. For antigen groups with three swapped AbAgs, N is 4. The score ranges from 0 to 1, corresponding to the worst and best possible rankings of the true AbAg, respectively.

$$\text{True rank score} = 1 - \frac{\text{True}_{\text{Rank}} - 1}{N - 1} \tag{3}$$

Table 3. AbAg interface scores used to predict AbAg interface accuracy measured by AbAgIoU. For AlphaFold-2.3, a weighted sum of intrinsic confidence metrics, ipTM and pTM, was used and set at 0.8 and 0.2, respectively. For ESMIF1 and AntiFold, average interface scores were calculated by summing residue probabilities (p_i) and dividing by the total number of residues, *N*. Machine learning models, X-AbAgIoU, were trained to map averaged One-hot, ESM2, ESMIF1, or AntiFold encodings (e_i) to AbAgIoU using an FFNN, *F*.

	AbAg interface score
AlphaFold-2.3	0.8 ipTM + 0.2 pTM
ESMIF1	$\frac{1}{N}\sum_{i=1}^{N} p_i$
AntiFold	$\frac{1}{N}\sum_{i=1}^{N} p_i$
X-AbAgIoU	$F\left(\frac{1}{n}\sum_{i=1}^{N} e_i\right)$
Random	$U(0, 1)$

Models

Model training and evaluation were done in nested fivefold cross-validation using the five data partitions and structures, as described above. In each outer loop, one partition was set aside as the blind test set, and the other four partitions were used for training in the inner loop. In the inner loop, three partitions were used for training, and one partition was used for validation, generating four models when rotating through all combinations. After each rotation, we evaluated models as an ensemble on the blind test set by averaging their outputs. Last, the model output on all data was obtained by concatenating model outputs from all blind test sets. Weights were initialized with a uniform distribution, $U\left(-\sqrt{k}, \sqrt{k}\right)$, where k is the number of neurons in a layer. We used a learning rate of 0.00005 and a weight decay of 0.001 for backpropagation, training models for up to 40 epochs.

Models for predicting AbAg interface accuracy

We created models to predict AbAg interface accuracy as measured by AbAgIoU. For AlphaFold, this accuracy was predicted using a weighted sum of its intrinsic confidence metrics (ipTM and pTM), as proposed by DeepMind researchers in Evans *et al.* (*21*). For the pretrained inverse folding GVP-Transformers (ESMIF1 and AntiFold), we averaged their probability outputs for epitope and paratope residues at the AbAg interface.

We also trained FFNNs to predict AbAgIoU using encodings of AbAg interface residues as input. One-hot, ESM2, ESMIF1, and AntiFold encodings of AbAg interface residues were averaged individually, resulting in input layer sizes of 189, 1280, 512, or 512 for four different FFNN models (Onehot-AbAgIoU, ESM2-AbAgIoU, AbEpiScore-1.0, and AntiFold-AbAgIoU). Although we experimented with a multiheaded attention architecture for encoding aggregation, it did not improve performance and was slower to train, leading to its exclusion. All models used a three-layer FFNN with 300, 150, and 100 neurons, mapping encodings to a single output neuron representing AbAg interface accuracy. Dropout rates of 0.6, 0.65, and 0.5 were applied between hidden layers. For generating random performance, we sampled scores from the uniform distribution $U(0,1)$ (Table 3).

During training, the data comprised all true AbAg structures where the antibody was positioned within 4 Å of the antigen. AbAgIoU was used as labels and a mean squared error loss function for weight updates and early stopping model selection. A custom batch function was implemented to group modeled structures of the same AbAg together. After training to avoid bias, we limited the evaluation to 30 structures per AbAg (across all five AlphaFold models and six seeds). In addition, structures where the antibody was not positioned within 4 Å of the antigen and could, therefore, not be scored were assigned the same score as the lowest-scoring structure of the 30 structures for that specific AbAg.

Models for classifying true from swapped AbAg complexes

We developed classifiers to distinguish between true AbAg complexes modeled with the correct antibody and those modeled with incorrect or swapped antibodies. Similar to the models for predicting AbAg interface accuracy, we used input layers sized 189, 1280, or 512 for averaged one-hot, ESM2, or ESMIF1 encodings of residues at the AbAg interface. We also tested adding AlphaFold-2.3 or AbEpiScore-1.0 scores as additional features, adjusting input layer sizes accordingly. A three-layer dense network with 450, 250, and 50 neurons was used to map encodings to two output neurons for logit scores of true and swapped AbAg. A softmax function was used to convert logits to probability outputs, with dropout rates of 0.65, 0.65, and 0.5 between hidden layers.

For training and evaluation, we first selected a structural model with the highest score (as evaluated by AbEpiScore-1.0) for each AbAg (true or swapped). In cases where more than 30 structural models were available for a given AbAg, the selection was limited to a random subset of 30. We used a custom batch function to make 1730 groups, each containing true and swapped AbAg structures modeled with the same antigen. Binary labels were assigned (0 for swapped AbAg and 1 for true AbAg), and a cross-entropy loss function was used to update weights. Early stopping based on maximal validation AUC was used to select the model.

We evaluated the classifiers by grouping true and swapped AbAgs modeled with the same antigen, totaling 1730 groups. To reduce data bias, we randomly selected three swapped AbAgs from the 53 groups with more than three swaps. This resulted in groups of four for all AbAgs (one true AbAg and three swapped AbAg). Statistical significance for some model comparisons was assessed using binomial tests (detailed in text S3). Later, when evaluating all true and swapped AbAg structures, we found that not all structures placed the antibody within 4 Å of the antigen and could, therefore, not be scored. In such cases, we assigned these structures the same score as the lowest-scoring structure of the 30 structures for that specific AbAg.

Supplementary Materials

This PDF file includes:

Supplementary Text
Figs. S1 to S28
Table S1

REFERENCES AND NOTES

1. S. K. Roe, B. Felter, B. Zheng, S. Ram, L. M. Wetzler, E. Garges, T. Zhu, C. A. Genco, P. Massari, In vitro pre-clinical evaluation of a gonococcal trivalent candidate vaccine identified by transcriptomics. *Vaccines* **11**, 1846 (2023).

2. T. Liu, C. Gao, J. Wang, J. Song, X. Chen, H. Chen, X. Zhao, H. Tang, D. Gu, Peptide aptamer-based time-resolved fluoroimmunoassay for CHIKV diagnosis. *Virol. J.* **20**, 166 (2023).
3. E. E. G. Kozlova, L. Cerf, F. S. Schneider, B. T. Viart, C. Nguyen, B. T. Steiner, S. de Almeida Lima, F. Molina, C. G. Duarte, L. Felicori, C. Chávez-Olórtegui, R. A. Machado-de-Ávila, Computational B-cell epitope identification and production of neutralizing murine antibodies against Atroxlysin-I. *Sci. Rep.* **8**, 14904 (2018).
4. J. E. Larsen, O. Lund, M. Nielsen, Improved method for predicting linear B-cell epitopes. *Immunome. Res.* **2**, 2 (2006).
5. M. C. Jespersen, B. Peters, M. Nielsen, P. Marcatili, BepiPred-2.0: Improving sequence-based B-cell epitope prediction using conformational epitopes. *Nucleic Acids Res.* **45**, W24–W29 (2017).
6. J. N. Clifford, M. H. Høie, S. Deleuran, B. Peters, M. Nielsen, P. Marcatili, BepiPred-3.0: Improved B-cell epitope prediction using protein language models. *Protein Sci.* **31**, e4497 (2022).
7. J. V. Kringelum, C. Lundegaard, O. Lund, M. Nielsen, Reliable B cell epitope predictions: Impacts of method development and improved benchmarking. *PLoS Comput. Biol.* **8**, e1002829 (2012).
8. M. H. Høie, F. S. Gade, J. M. Johansen, C. Würtzen, O. Winther, M. Nielsen, P. Marcatili, DiscoTope-3.0: Improved B-cell epitope prediction using inverse folding latent representations. *Front. Immunol.* **15**, 1322712 (2024).
9. C. Zhou, Z. Chen, L. Zhang, D. Yan, T. Mao, K. Tang, T. Qiu, Z. Cao, SEPPA 3.0—Enhanced spatial epitope prediction enabling glycoprotein antigens. *Nucleic Acids Res.* **47**, W388–W394 (2019).
10. I. Sela-Culang, Y. Ofran, B. Peters, Antibody specific epitope prediction-emergence of a new paradigm. *Curr. Opin. Virol.* **11**, 98–102 (2015).
11. P. Weber, C. Pissis, R. Navaza, A. E. Mechaly, F. Saul, P. M. Alzari, A. Haouz, High-throughput crystallization pipeline at the crystallography core facility of the institut pasteur. *Molecules* **24**, 4451 (2019).
12. L. A. Earl, S. Subramaniam, Cryo-EM of viruses and vaccine design. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 8903–8905 (2016).
13. L. Ledsgaard, A. Ljungars, C. Rimbault, C. V. Sørensen, T. Tulika, J. Wade, Y. Wouters, J. McCafferty, A. H. Laustsen, Advances in antibody phage display technology. *Drug Discov. Today* **27**, 2151–2169 (2022).
14. L. D. Goldstein, Y. J. Chen, J. Wu, S. Chaudhuri, Y. C. Hsiao, K. Schneider, K. H. Hoi, Z. Lin, S. Guerrero, B. S. Jaiswal, J. Stinson, A. Antony, K. B. Pahuja, D. Seshasayee, Z. Modrusan, I. Hötzel, S. Seshagiri, Massively parallel single-cell B-cell receptor sequencing enables rapid discovery of diverse antigen-reactive antibodies. *Commun. Biol.* **2**, 304 (2019).
15. I. Sela-Culang, S. Ashkenazi, B. Peters, Y. Ofran, PEASE: Predicting B-cell epitopes utilizing antibody sequence. *Bioinformatics* **31**, 1313–1315 (2015).
16. M. C. Jespersen, S. Mahajan, B. Peters, M. Nielsen, P. Marcatili, Antibody specific B-cell epitope predictions: Leveraging information from antibody-antigen protein complexes. *Front. Immunol.* **10**, 298 (2019).
17. A. Davila, Z. Xu, S. Li, J. Rozewicki, J. Wilamowski, S. Kotelnikov, D. Kozakov, S. Teraguchi, D. M. Standley, AbAdapt: An adaptive approach to predicting antibody-antigen complex structures from sequence. *Bioinform. Adv.* **2**, vbac015 (2022).
18. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
19. Z. Xu, A. Davila, J. Wilamowski, S. Teraguchi, D. M. Standley, Improved antibody-specific epitope prediction using AlphaFold and AbAdapt. *Chembiochem* **23**, e202200303 (2022).
20. R. Yin, B. G. Pierce, Evaluation of AlphaFold antibody–antigen modeling with implications for improving predictive accuracy. *Protein Sci.* **33**, e4865 (2024).
21. R. Evans, M. O'Neill, A. Pritzel, N. Antropova, A. Senior, T. Green, Augustin Žídek, R. Bates, S. Blackwell, J. Yim, O. Ronneberger, S. Bodenstein, M. Zielinski, A. Bridgland, A. Potapenko, A. Cowie, K. Tunyasuvunakool, R. Jain, E. Clancy, P. Kohli, J. Jumper, D. Hassabis, Protein complex prediction with AlphaFold-Multimer. bioRxiv 463034 [Preprint] (2022). https://doi.org/10.1101/2021.10.04.463034.
22. C. Hsu, R. Verkuil, J. Liu, Z. Lin, B. Hie, T. Sercu, A. Lerer, A. Rives, Learning inverse folding from millions of predicted structures. bioRxiv 487779 [Preprint] (2022). https://doi.org/10.1101/2022.04.10.487779.
23. M. H. Høie, A. M. Hummer, T. H. Olsen, B. Aguilar-Sanjuan, M. Nielsen, C. M. Deane, AntiFold: Improved structure-based antibody design using inverse folding. *Bioinform. Adv.* **5**, vbae202 (2025).
24. Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, C. A. Dos Santos, M. Fazel-Zarandi, T. Sercu, S. Candido, A. Rives, Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
25. S. Basu, B. Wallner, DockQ: A quality measure for protein-protein docking models. *PLOS ONE* **11**, 0161879 (2016).
26. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne, The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
27. D. T. Dransfield, E. H. Cohen, Q. Chang, L. G. Sparrow, J. D. Bentley, O. Dolezal, X. Xiao, T. S. Peat, J. Newman, P. A. Pilling, T. Phan, I. Priebe, G. V. Brierley, N. Kastrapeli, K. Kopacz, D. Martik, D. Wassaf, D. Rank, G. Conley, Y. Huang, T. E. Adams, L. Cosgrove, A human monoclonal antibody against insulin-like growth factor-II blocks the growth of human hepatocellular carcinoma cell lines in vitro and in vivo. *Mol. Cancer Ther.* **9**, 1809–1819 (2010).
28. K. P. Kilambi, J. J. Gray, Structure-based cross-docking analysis of antibody-antigen interactions. *Sci. Rep.* **7**, 8145 (2017).
29. S. Chaudhury, J. J. Gray, Conformer selection and induced fit in flexible backbone protein-protein docking using computational and NMR ensembles. *J. Mol. Biol.* **381**, 1068–1087 (2008).
30. J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A. J. Ballard, J. Bambrick, S. W. Bodenstein, D. A. Evans, C. C. Hung, M. O'Neill, D. Reiman, K. Tunyasuvunakool, Z. Wu, A. Žemgulytė, E. Arvaniti, C. Beattie, O. Bertolli, A. Bridgland, A. Cherepanov, M. Congreve, A. I. Cowen-Rivers, A. Cowie, M. Figurnov, F. B. Fuchs, H. Gladman, R. Jain, Y. A. Khan, C. M. R. Low, K. Perlin, A. Potapenko, P. Savy, S. Singh, A. Stecula, A. Thillaisundaram, C. Tong, S. Yakneen, E. D. Zhong, M. Zielinski, A. Žídek, V. Bapst, P. Kohli, M. Jaderberg, D. Hassabis, J. M. Jumper, Addendum: Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **636**, E4 (2024).
31. J. Wohlwend, G. Corso, S. Passaro, M. Reveiz, K. Leidal, W. Swiderski, T. Portnoi, I. Chinn, J. Silterra, T. Jaakkola, R. Barzilay, Boltz-1: Democratizing biomolecular interaction modeling. bioRxiv 624167 [Preprint] (2024). https://doi.org/10.1101/2024.11.19.624167.
32. Chai Discovery team, J. Boitreaud, J. Dent, M. McPartlon, J. Meier, V. Reis, A. Rogozhonikov, K. Wu, Chai-1: Decoding the molecular interactions of life. bioRxiv 615955 [Preprint] (2024). https://doi.org/10.1101/2024.10.10.615955.
33. T. Qiu, L. Zhang, Z. Chen, Y. Wang, T. Mao, C. Wang, Y. Cun, G. Zheng, D. Yan, M. Zhou, K. Tang, Z. Cao, SEPPA-mAb: Spatial epitope prediction of protein antigens for mAbs. *Nucleic Acids Res.* **51**, W528–W534 (2023).
34. C. Wang, J. Wang, W. Song, G. Luo, T. Jiang, EpiScan: Accurate high-throughput mapping of antibody-specific epitopes using sequence information. *NPJ Syst. Biol. Appl.* **10**, 101 (2024).
35. M. S. Klausen, M. V. Anderson, M. C. Jespersen, M. Nielsen, P. Marcatili, LYRA, a webserver for lymphocyte receptor structural modeling. *Nucleic Acids Res.* **43**, W349–W355 (2015).
36. L. Li, S. Chen, Z. Miao, Y. Liu, X. Liu, Z. X. Xiao, Y. Cao, AbRSA: A robust tool for antibody numbering. *Protein Sci.* **28**, 1524–1531 (2019).
37. R. K. Grover, X. Zhu, T. Nieusma, T. Jones, I. Boreo, A. S. MacLeod, A. Mark, S. Niessen, H. J. Kim, L. Kong, N. Assad-Garcia, K. Kwon, M. Chesi, V. V. Smider, D. R. Salomon, D. F. Jelinek, R. A. Kyle, R. B. Pyles, J. I. Glass, A. B. Ward, I. A. Wilson, R. A. Lerner, A structurally distinct human mycoplasma protein that generically blocks antigen-antibody union. *Science* **343**, 656–661 (2014).
38. M. Steinegger, J. Söding, MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
39. A. A. Hagberg, D. A. Schult, P. J. Swart, "Exploring network structure, dynamics, and function using NetworkX," in *Proceedings of the 7th Python in Science Conference (SciPy2008)*, G. Varoquaux, T. Vaught, J. Millman, Eds. (Pasadena, CA, 2008), pp. 11–15.
40. M. Mirdita, K. Schütze, Y. Moriwaki, L. Heo, S. Ovchinnikov, M. Steinegger, ColabFold: Making protein folding accessible to all. *Nat. Methods* **19**, 679–682 (2022).
41. C. Lange, C. Hunte, Crystal structure of the yeast cytochrome bc_1 complex with its bound substrate cytochrome *c*. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 2800–2805 (2002).
42. S. R. Solmaz, C. Hunte, Structure of complex III with bound cytochrome *c* in reduced state and definition of a minimal core interface for electron transfer. *J. Biol. Chem.* **283**, 17542–17549 (2008).

Acknowledgments

Funding: This work was funded by the National Cancer Institute (NCI), with award number U24CA248138 (B.P. and M.N.), and the National Institute of Allergy and Infectious Diseases (NIAID), with award number 75N93019C00001 (B.P. and M.N.). **Author contributions:** Conceptualization: J.N.C., M.N., and B.P. Methodology: J.N.C. and M.N. Investigation: J.N.C. and M.N. Visualization: J.N.C. Funding acquisition: M.N. and B.P. Project administration: M.N. Supervision: M.N. and B.P. Writing—original draft: J.N.C. and M.N. Writing—review and editing: J.N.C., M.N., E.R., and B.P. Data curation: J.N.C. and M.N. Validation: J.N.C. and M.N. Formal analysis: J.N.C. and M.N. Software: J.N.C. and M.N. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. We also provide the software as a web server at https://services.healthtech.dtu.dk/services/AbEpiTope-1.0/ and a stand-alone version for local installation at https://github.com/mnielLab/

AbEpiTope-1.0/. In addition, we provide all AlphaFold-2.3 fasta input files for reproducing both true and swapped AbAg complexes structures at https://services.healthtech.dtu.dk/services/AbEpiTope-1.0/ under the "Data" tab. We have also provided the data needed to recreate the graphs present in this paper, such as all model scores and labels (AbAgIoU, AgIoU, and DockQ).

Submitted 27 October 2024
Accepted 8 May 2025
Published 13 June 2025
10.1126/sciadv.adu1823