

Supplementary information

Contents

1	Methods	1
1.1	The core module of AbOptimizer	1
1.2	Datasets	1
1.3	Surrogate Models of Property Scoring	3
	Surrogate models of emibetuzumab affinity and specificity • Surrogate model of trastuzumab affinity • Surrogate model of siltuximab fitness	
	• Surrogate models of common biophysical properties	
1.4	Baselines	5
1.5	Molecular docking	6
2	More results about AbOptimizer	6
2.1	Statistical analysis.	6
2.2	Analysis of top-optimized variants generated by different methods	6
2.3	Discussion on the reward function of AbOptimizer	7
2.4	Originality and germline bias analyses of optimized variants	9
2.5	Discussion of generic surrogates for AbOptimizer-guided optimization	11
2.6	Multi-site mutagenesis achieved by AbOptimizer	11
3	Optimized candidates for bioassay experiments	13
3.1	Details of selection rules for experiment validation	13
3.2	Detailed results of molecular docking	13
3.3	Details of SDS-PAGE and SEC characterization of purified antibody variants	16
3.4	Details of EC50 fitting parameters for antigen-binding ELISA curves	17
3.5	Details of thermal stability analysis	17
	References	17

1 Methods

1.1 The core module of AbOptimizer

AbOptimizer is a policy gradient-based reinforcement learning (RL) framework consisting of 3 core modules (Supplementary Fig. 1): the Editor, the Interpreter, and the Evaluator. The specific functions of each module are described as follows:

Editor: The Editor's policy is represented by neural networks, where the one-hot encoded sequence is input into a transformer encoder. This encoder generates representations that are subsequently decoded into the action space.

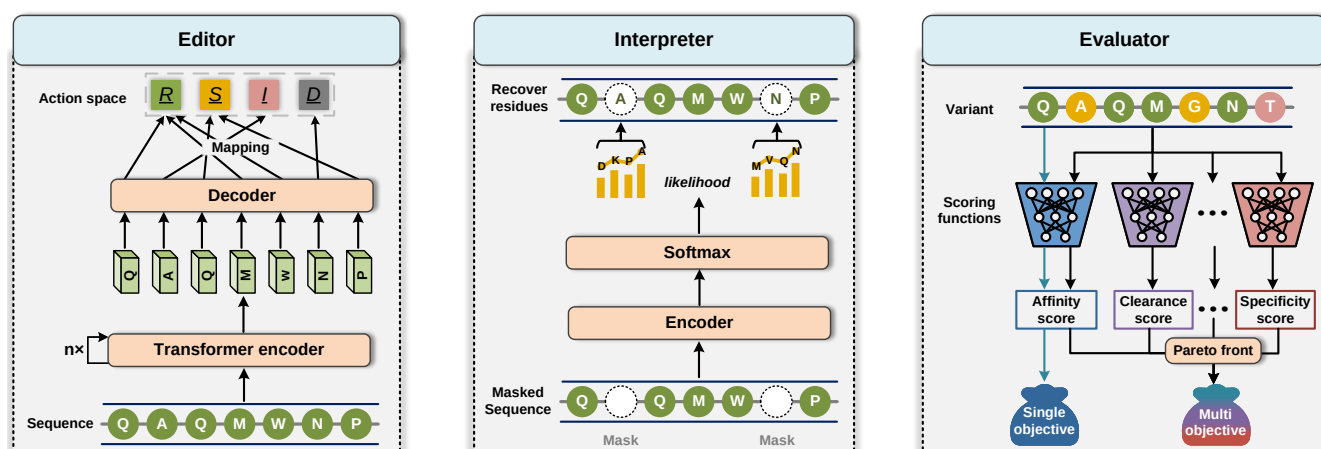
Interpreter: The Interpreter is built upon the pre-trained antibody LLM model AbLang, and predicts plausible mutations for residues.

Evaluator: The Evaluator's reward function consists of single- or multi-property scoring surrogates, integrated using a Pareto-based fusion scheme.

1.2 Datasets

In this work, we considered 3 distinct therapeutic antibodies (emibetuzumab, trastuzumab, and siltuximab) for optimization. Emibetuzumab targets the hepatocyte growth factor receptor (HGFR) and is currently in Phase II clinical trials for non-small cell lung cancer¹. Despite its high binding affinity, it exhibits elevated non-specific binding. Trastuzumab targets the human epidermal growth factor receptor 2 (HER2) and is widely used in HER2-positive breast cancer treatment². Siltuximab is a chimeric monoclonal antibody that neutralizes interleukin-6 (IL-6), a soluble cytokine implicated in inflammation and tumorigenesis^{3,4}. The mutation datasets we collected for emibetuzumab⁵ and trastuzumab^{6,7} are summarized in Supplementary Table 1.

For emibetuzumab, we directly adopted the dataset from Makowski et al.⁵, which includes 4,000 equal-length variants, each annotated with binary labels for affinity (binding strength to HGFR) and non-specific binding (binding to polyspecificity reagents). Among these, 1,516 and 2,484 variants are labeled as 'binding' and 'non-binding', respectively, and 2,000 and 2,000 variants are labeled as low non-specific binders ("specific") and high non-specific binders ("non-specific"). It is important to note that high specificity is defined as low non-specific binding, but this does not necessarily imply a direct correlation with high affinity. Additionally, Makowski et al.⁵ provided an independent dataset with 125 Fab variants with K_D measurements.



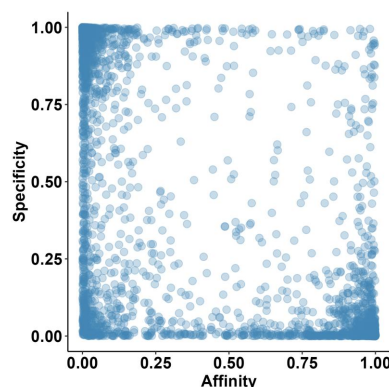
Supplementary Fig. 1. The core module of AbOptimizer.

For trastuzumab, we compiled the dataset from 2 sources. Firstly, we adopted the classification dataset from Mason et al.⁶, comprising 38,839 equal-length variants, of which 11,300 are labeled as binding and 27,539 as non-binding⁶. Notably, in the source study⁶, “specificity” refers to antigen binding. To maintain terminological consistency, we uniformly use the term affinity to denote this property. Then, we considered another classification dataset with 1,434 variable-length variants, of which 337 are labeled as binding and 1,097 as non-binding⁷, to train a binding predictor better suited for handling variable-length sequences. By integrating the 2 sources, we created a training set of 11,637 binding and 28,636 non-binding variants. An independent dataset of 421 variable-length variants with K_D measurements was used to evaluate the predictor’s performance⁷.

Supplementary Table 1. Dataset statistics for emibetuzumab and trastuzumab.

Antibody	Subset	Type	Length	Binding	Non-binding	High Non-specific	Low Non-specific	K_D	Total
emibetuzumab	Train	Binary	Equal	1,516	2,484	2,000	2,000	–	4,000*
	Test	Numeric	Equal	–	–	–	–	125	125
trastuzumab	Train	Binary	Equal	11,300	27,539	–	–	–	38,839
	Train	Binary	Variable	337	1,097	–	–	–	1,434
	Test	Numeric	Variable	–	–	–	–	421	421

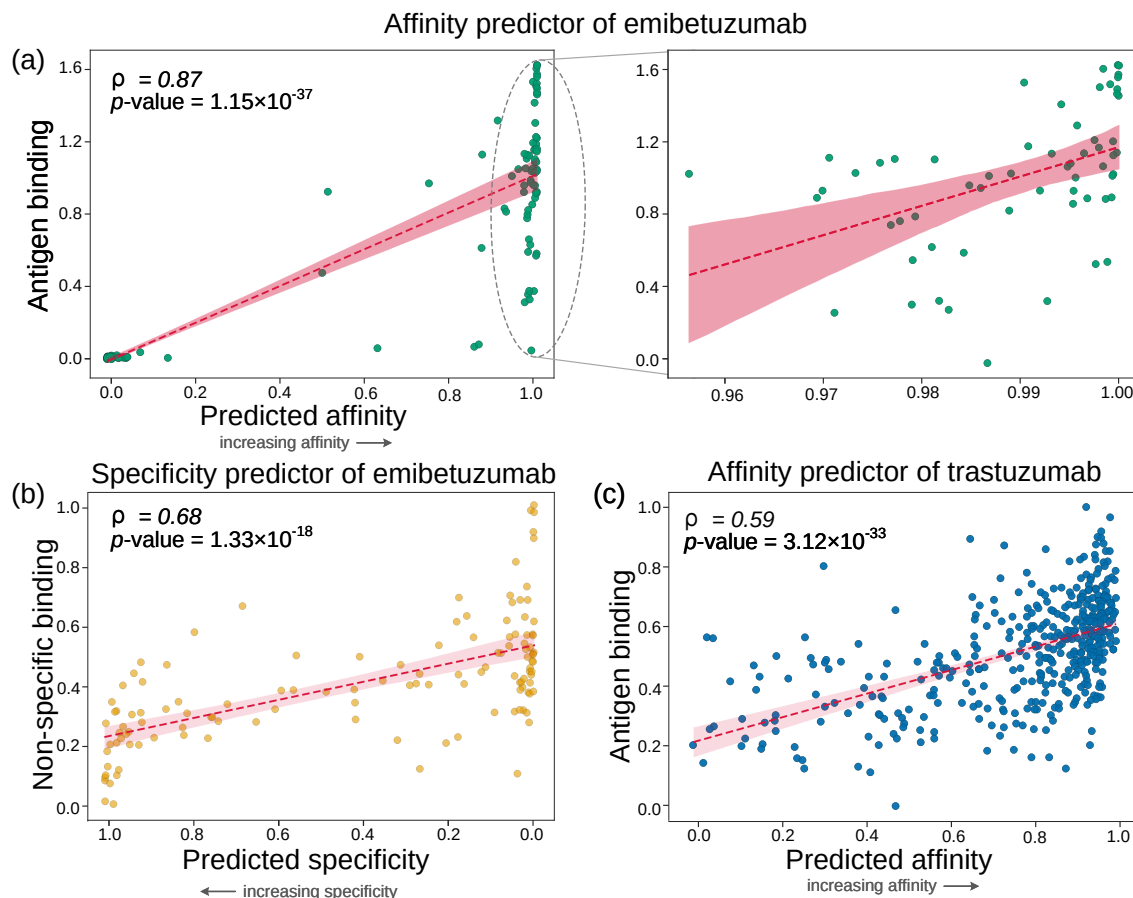
*The emibetuzumab training set contains 4,000 sequences in total, each annotated as either binding or non-specific binding.



Supplementary Fig. 2. Affinity and specificity distributions of the emibetuzumab dataset as estimated by the 2 predictors.

1.3 Surrogate Models of Property Scoring

Using the datasets described above, we trained an affinity predictor and a specificity predictor for emibetuzumab, and an affinity predictor for trastuzumab. For siltuximab, we did not construct any antibody-specific predictors. Instead, we employed a general-purpose fitness predictor within AbOptimizer as a surrogate. This allowed us to assess the generalizability of our method in guiding affinity maturation beyond the antibody-specific setting. The predictors were trained on the training sets, and their hyperparameters were empirically determined based on hold-out validation results (80%: 20% training: test random split, evaluated in terms of AUC) within the training data. Final evaluations were performed on independent test sets. Detailed parameter settings and implementation details are available in our GitHub repository (<https://anonymous.4open.science/r/AbOptimizer>).



Supplementary Fig. 3. Correlation between the predicted affinity and specificity scores and experimental measurements for emibetuzumab and trastuzumab dataset. **a**, Correlation between emibetuzumab predicted affinity and experimentally measured antigen binding scores. **b**, Correlation between emibetuzumab predicted specificity and experimentally measured non-specific binding scores. **c**, Correlation between trastuzumab predicted affinity and experimentally measured antigen binding scores.

1.3.1 Surrogate models of emibetuzumab affinity and specificity

For emibetuzumab, we followed a recent work⁵ to train the affinity predictor and specificity predictor. Specifically, using the training set of emibetuzumab, we constructed 2 lightweight convolutional neural network (CNN) classifiers as surrogates within AbOptimizer, one for affinity and one for specificity. Each predictor comprises 2 stacked 1D convolutional layers, including a dilated convolution to capture long-range dependencies, followed by global max pooling and a fully connected layer. A sigmoid activation function is applied to produce the final score. Following the evaluation protocol of Makowski et al.⁵, we assessed the correlation between the continuous outputs of the classifiers and experimentally measured K_D values (Supplementary Fig. 3a, b). These classifiers show strong Spearman correlation with experimentally measured K_D values for binding affinity ($r = 0.87$, $p = 1.15 \times 10^{-37}$, Supplementary Fig. 3a) and non-specific binding ($r = 0.68$, $p = 1.33 \times 10^{-18}$, Supplementary Fig. 3b), indicating the reliable performances of the trained predictors for affinity and specificity.

Subsequently, we applied the affinity and specificity predictors to estimate the binding and specificity probabilities for the 4,000 antibody variants of emibetuzumab, and then analyzed the distribution of these predicted affinity and specificity scores (Supplementary Fig. 2). Data points are predominantly clustered in the upper-left, lower-left, and lower-right quadrants, while the upper-right quadrant, representing high affinity and high specificity, remains notably sparse. This distribution suggests a significant trade-off between affinity and specificity among emibetuzumab variants, indicating that high-affinity antibodies typically exhibit low specificity, and vice versa. Only a limited number of variants achieve both desirable properties simultaneously, highlighting the inherent challenge of co-optimizing affinity and specificity.

1.3.2 Surrogate model of trastuzumab affinity

For trastuzumab, we trained an affinity predictor using a collected dataset comprising both equal⁶- and variable⁷-length sequences, employing the same CNN architecture as for emibetuzumab. We used CNNs to encode sequences rather than position-sensitive one-hot vectors directly as features. CNNs apply shared convolutional filters across positions, making them inherently translation-invariant and well-suited for encoding variable-length sequences. On an independent test set of variable-length sequences with measured K_D values, the predictor achieved a strong Spearman correlation coefficient with binding affinity ($r = 0.59$, $p = 3.12 \times 10^{-33}$, Supplementary Fig. 3c).

1.3.3 Surrogate model of siltuximab fitness

We adopted the fitness predictor developed by Shanker et al.⁸ as the surrogate model within AbOptimizer for optimizing siltuximab. This fitness predictor is a general-purpose protein language model, ESM-IF1 (GVP-Transformer, April 10, 2022 release), which incorporates backbone structural coordinates and assigns a likelihood score to each input variant. Higher likelihood scores indicate a greater probability that the variant will fold reliably into the reference backbone structure.

Shanker et al.⁸ demonstrated that constraining sequence search within high-likelihood regions of this model's latent space serves as an effective prior for identifying high-affinity variants. They further validated that the log-likelihoods assigned by the model show significant Spearman correlations (ranging from 0.50 to 0.65) with experimentally measured binding affinities (K_D) across multiple datasets. This predictive power is attributed to the model's ability to implicitly capture structural properties relevant to binding, such as interface packing and steric complementarity, even without explicitly modeling chemical interactions. In addition, the fitness scores, estimated from sequence likelihood or perplexity, capture sequence-level constraints that are informative for antibody developability^{9,10}. These scores may provide a generic sequence-level prior that is indirectly informative for binding-relevant structural compatibility and some developability-associated constraints.

In our study, this fitness predictor was integrated into AbOptimizer as a surrogate for guiding the affinity maturation of siltuximab. The surrogate takes as input the variant sequences along with the wild-type complex structure and outputs the likelihood of each variant, which is regarded as its fitness score. Given that the model was primarily validated for single-point mutations⁸, we extended its use to multi-point variants by averaging the predicted fitness scores of the corresponding single-residue substitutions. The 3D structure of the IL-6–siltuximab complex used for fitness estimation was predicted using ColabFold¹¹. The resulting complex structure is consistent with previous reports, indicating that the antibody binds to the C-terminal region of helix D of IL-6¹².

1.3.4 Surrogate models of common biophysical properties

In addition to affinity and specificity, we consider 3 common biophysical properties: immunogenicity, viscosity and clearance.

To evaluate immunogenicity, we employed NetMHCIIpan¹³ to estimate the propensity of antibody-derived peptides to bind MHC class II molecules. Specifically, for each antibody sequence, we constructed a local sequence centered on the CDRH3 by padding the CDRH3 region with 10 flanking amino acids on both sides. All overlapping 15-mer peptides derived from this padded sequence were then evaluated against a panel of 34 reference Human Leukocyte Antigen (HLA) alleles using NetMHCIIpan. For each peptide–HLA pair, NetMHCIIpan outputs a percentile rank, which reflects the predicted binding strength relative to a reference set of random peptides. A higher percentile rank indicates weaker predicted binding to MHC-II and, consequently, lower immunogenic potential. Following previous study^{14,6}, we used the minimum percentile rank (minPR) across all 15-mer peptides derived from the padded sequence as an indicator of immunogenicity. Intuitively, a higher minPR indicates that even the most strongly binding peptide is predicted to bind relatively weakly to MHC-II, suggesting lower immunogenic potential. The resulting minPR was subsequently transformed into the range of [0, 1] using a sigmoid function.

To evaluate viscosity and clearance, we utilized three metrics¹⁵: antibody variable fragment net charge (FvNetCharge), antibody variable fragment charge symmetry parameter (FvCSP), and hydrophobicity index sum (HISum). These properties are closely associated with the viscosity and clearance rates of antibodies. FvNetCharge and FvCSP primarily reflect the charge characteristics of the antibody's variable region, with increasing values of both indices reported to correlate with decreased viscosity¹⁵. Specifically, FvNetCharge represents the sum of the net charges of the heavy chain (VH) and light chain (VL), while FvCSP denotes the product of these 2 charges. The net charge for a given sequence was calculated by summing the contributions from all charged amino acids, utilizing the known pKa values of the side chains and the Henderson–Hasselbalch

equation. HISum, on the other hand, quantifies the hydrophobicity of the antibody's variable region and can be calculated as follows:

$$HI = - \left(\frac{\sum n_i E_i}{\sum n_j E_j} \right) \quad (1)$$

where i represents hydrophobic amino acids (i.e., A, C, F, I, L, P, V, W, and Y), and j denotes hydrophilic amino acids (i.e., D, E, G, H, K, M, N, Q, R, S, and T). In this context, n refers to the quantity of each amino acid, while E signifies the Eisenberg scale value of each amino acid.

Optimal clearance rates were associated with FvNetCharge values ranging from 0 to 6.2¹⁵. Furthermore, the HISum values for CDR-L1, CDR-L3, and CDR-H3 should be maintained below 4¹⁵. In alignment with previous work by Xu et al.¹⁴, we normalized the scores of the 5 properties to a range of 0 to 1, where higher scores indicate superior performance in each respective property.

1.4 Baselines

Virtual Screening (VS): Previous studies have leveraged VS as a computational strategy for antibody optimization^{5,6}. In their studies, an antibody variant library was generated through single-point or multi-point random mutation strategies. The library was subsequently evaluated using antibody property predictors, from which the variant with the highest property score was selected as the optimized antibody. In this study, to optimize a given antibody sequence, we mutated all selected positions by replacing the corresponding amino acids with alternative residues. This process generated an antibody variant library consisting of 128 variants, which we then screened using the antibody property predictors outlined in the main text. The variant with the highest property score was ultimately chosen as the optimized antibody.

Genetic Algorithm (GA): Bachas et al.¹⁶ demonstrated the utility of GA in antibody optimization. GA emulates the principles of natural selection, serving as optimization techniques that identify optimal solutions through processes of inheritance, mutation, and selection. In this study, a total of 128 variants were generated using a GA, which began with a population of 16 variants and evolved over 8 generations. Specifically, an initial population of 16 variants was generated using multi-point random mutation. The antibody property predictors were utilized as fitness functions to assess these variants. Offspring individuals were produced through single-point crossover and subsequently diversified by single-point random mutation. After iterating through the entire population for 8 generations, the variant with the highest property score was selected as the optimized antibody.

Bayesian Optimization (BO): Khan et al.¹⁷ employed a BO algorithm for antibody optimization. This approach is an effective method for optimizing black-box functions, based on the principle of utilizing a surrogate function to approximate the unknown black-box function. By optimizing the surrogate function, researchers can indirectly improve the performance of the black-box function itself. In this study, an initial set of 64 mutated variants was generated using multi-point random mutation. These initial variants were evaluated using the antibody property predictors, which formed the black-box function to be optimized. BO was then applied to explore an additional 64 variants through a random forest-based surrogate model, resulting in a total of 128 searched sequences. The surrogate model guided the search toward promising regions of the sequence space by balancing exploration and exploitation. The antibody sequence with the highest predicted property score from this process was selected as the optimized result.

Gradient Ascent (GAsc): The potential of gradient-based optimization in antibody engineering was exemplified by Liu et al.¹⁸, who introduced a GAsc strategy for sequence refinement. This GAsc relies on gradient-based methods to improve performance. In this study, the GAsc was employed to optimize antibody sequences by computing gradients from the antibody property predictors and backpropagating them to the input. For each input sequence, GAsc was performed for 128 steps, progressively modifying the sequence toward higher predicted property scores. The final output after 128 updates was taken as the optimized antibody sequence. This approach requires that the antibody property predictors be differentiable, with inputs represented as one-hot encoded sequences. Among the property predictors integrated into AbOptimizer, only the affinity predictor fulfilled the conditions necessary for GAsc optimization.

Protein LLM (ProtLLM): Hie et al.¹⁹ introduced a language model for antibody optimization. It calculates the likelihood of each position by ESM1 in the antibody sequence, and a greedy strategy is employed to select the substitution with the highest likelihood compared to the wild-type for amino acid replacement. In this study, we used the AbLang, which was introduced by Olsen et al.²⁰, a model trained on large-scale natural antibody sequences, to calculate the likelihood.

Protein LLM*: The ProtLLM* is an extension of the ProtLLM. Unlike its predecessor, ProtLLM* samples amino acid substitutions from the language model's likelihood distribution, rather than deterministically selecting the highest-likelihood residues.

To ensure a fair comparison, the number of optimized sequences was constrained to 128 for each method. Given the limited set of potential mutational sites, searching an excessively large number of sequences could obscure performance differences

among baselines. Therefore, we adopted a smaller search space to more clearly assess the ability of different approaches to rapidly identify high-scoring variants. All methods were restricted to a maximum of 8 mutations for emibetuzumab, 10 mutations for trastuzumab and 23 mutations for siltuximab. Detailed parameter settings and implementation details are available in our GitHub repository (<https://anonymous.4open.science/r/AbOptimizer>).

1.5 Molecular docking

Antibody structure modeling. For emibetuzumab, we employed trRosetta²¹ to predict the structures of the heavy chain variable region (VH) and the light chain variable region (VL). The structures of the VH and VL from the Protein Data Bank (PDB) entry 6WVZ [<https://doi.org/10.2210/pdb6WVZ/pdb>] were utilized as a framework to establish the initial orientations of these regions. We subsequently combined the predicted structures of VH and VL by aligning their residues with the corresponding residues in the framework, thereby constructing the fragment variable region (Fv). The final Fv structures of emibetuzumab variants were visualized using the open-source software PyMOL.

Docking Software. ZDOCK 3.0.2 is a rigid-body molecular docking algorithm that employs fast Fourier transforms to enhance performance during translational searches²². ZRANK 2.0 is a scoring function that refines initial docking predictions by incorporating detailed electrostatics, van der Waals forces, repulsive energies, and desolvation effects²³. The crystal structure of the HGFR–amivantamab complex (PDB ID: 6WVZ) was utilized to extract the receptor structure, which includes the SEMA and PSI domains. For the docking process, the paratope and epitope were identified, and residues not involved in these sites were excluded using the ZDOCK pipeline. The paratope comprises 6 complementarity-determining regions (CDR), while the epitope consists of specific regions of HGFR located within the SEMA domain, which are targeted by emibetuzumab. These regions include residues at positions 123 to 128, 144 to 156, 192 to 195, and 220 to 227¹.

Hydrogen Bond Analysis. Based on the docking conformations obtained from ZDOCK, interface residues between the antigen and antibody were identified by analyzing the changes in solvent-accessible surface areas of the complex relative to each binding partner. Residues exhibiting a difference in surface area greater than 1 Å² were classified as interface residues. Subsequently, hydrogen bonds among these interface residues were detected using a cutoff distance of 3.6 Å²⁴ and a donor–H–acceptor angle >110°²⁵.

2 More results about AbOptimizer

2.1 Statistical analysis.

Statistical tests were selected according to the experimental design, with two-sided paired Student's t-test applied to matched *in silico* comparisons and ANOVA-based multiple comparisons used for independent *in vitro* experiments.

For *in silico* analyses, two-sided paired Student's t-test were performed on results matched by the same starting sequence. The Wilcoxon signed-rank test was used for LD analysis in affinity optimization. Raw *P*-values were adjusted using the Holm correction. For emibetuzumab affinity maturation, co-optimization of affinity and specificity, and multi-property optimization, AbOptimizer was compared pairwise with baseline methods using RI scores (Supplementary Table 2a), with additional performance metrics provided in Supplementary Table 2b. For trastuzumab, affinity scores and RI scores under the two-action and four-action settings were compared (Supplementary Table 2c). For siltuximab, optimization was performed from a single starting sequence using three independent random seeds to account for stochasticity. Given the limited replicates (*n* = 3), results are reported as mean ± SD without formal hypothesis testing.

For *in vitro* experiments involving multiple groups, statistical significance was assessed using one-way ANOVA followed by Dunnett's multiple comparisons test. The corresponding adjusted *P*-values are summarized in Supplementary Table 3.

2.2 Analysis of top-optimized variants generated by different methods

To further identify the best-performing variants, we pooled all optimized variants from all methods, ranked them for each property, and then aggregated the ranks by averaging across properties (Borda count²⁶). To prevent the ranking from masking deficiencies in any single property, variants with a score below 0.5 in any property were excluded from the aggregation. Then, we recorded the methods corresponding to the Top-5, Top-10, and Top-15. Methods that generate a greater number of optimal variants tend to achieve a higher performance in identifying the best-optimized variant.

In co-optimization of affinity and specificity, we recorded the methods corresponding to the Top-5, 10 and 15 within each group using the average of affinity and specificity scores. As summarized in Supplementary Table 4, the majority of high-ranking optimized variants were obtained from AbOptimizer, which consistently outperformed other methods across all groups. Specifically, among the top 5 variants in each of the 4 interval groups (20 variants in total), 19 were optimized by AbOptimizer (95%) and 1 by ProteinLLM* (5%), with no variants contributed by other methods. Expanding to the top 10 variants in each group (40 variants in total), 33 originated from AbOptimizer (82%), 7 from ProteinLLM* (18%), while no other methods were represented. Further, among the top 15 variants in each group (60 variants in total), 45 were optimized by AbOptimizer (75%) and 15 by ProtLLM* (25%), again with no variants from the remaining methods.

Supplementary Table 2. Adjusted P values for emibetuzumab and trastuzumab optimization tasks. LD comparisons in affinity optimization used the Wilcoxon signed-rank test; other comparisons used two-sided paired Student's t -test. All P values were Holm-adjusted.

a. RI-related comparisons between AbOptimizer and baseline methods in emibetuzumab.

Baseline	RI in affinity optimization	RI in affinity–specificity optimization	RI in multi-property optimization
VS	< 0.0001	0.0005	< 0.0001
BO	< 0.0001	0.0002	< 0.0001
GAsc	0.0044	N.A.	N.A.
GA	< 0.0001	< 0.0001	< 0.0001
ProtLLM	< 0.0001	< 0.0001	0.0002
ProtLLM*	0.1230	0.0257	0.0473

b. Other metric comparisons between AbOptimizer and baseline methods in emibetuzumab.

Baseline	Affinity score	Humanness in affinity optimization	LD in affinity optimization
VS	< 0.0001	< 0.0001	< 0.0001
BO	< 0.0001	< 0.0001	< 0.0001
GAsc	0.0052	< 0.0001	< 0.0001
GA	< 0.0001	< 0.0001	< 0.0001
ProtLLM	< 0.0001	< 0.0001	< 0.0001
ProtLLM*	0.0057	< 0.0001	< 0.0001

c. Comparisons between two-action and four-action settings in trastuzumab.

Parameter	P value
Affinity score	< 0.0001
RI	0.0003

N.A., not applicable; LD, Levenshtein distance. Raw P -values were adjusted using the Holm correction. All values smaller than 0.0001 are reported as < 0.0001.

In multiple optimization, we recorded the methods corresponding to the Top-5, Top-10, and Top-15 using the average of the 6 property scores. As summarized in Supplementary Table 5, the majority of high-ranking optimized variants were obtained from AbOptimizer, which consistently outperformed other methods across all groups. AbOptimizer accounted for 5 (100%), 10 (100%), and 14 (93%) of the variants in the Top 5, 10, and 15 rankings, respectively, substantially outperforming other baselines.

2.3 Discussion on the reward function of AbOptimizer

We implemented 3 distinct reward functions in AbOptimizer for the co-optimization of affinity and specificity: weighted sum (WS), dynamic weighted sum (DWS), and Pareto front (PF) optimization. In the WS scheme, affinity and specificity are equally weighted (1:1). The DWS method emphasizes balancing trade-offs among multiple properties by dynamically adjusting the weight assigned to each property. Properties with lower optimization scores at any given time are assigned higher weights, whereas properties that are closer to their optimal targets receive lower weights. In contrast, the PF method concentrates on the performance of individual optimized antibodies within a given set. Here, a non-dominated sorting algorithm ranks each antibody variant and assigns rewards based on its rank. Antibodies that lie on the Pareto front and exhibit balanced performance across multiple properties are rewarded with the highest weights.

As illustrated in Supplementary Fig. 4, the Pareto-based function effectively balances trade-offs between multiple properties, specifically in the low-affinity/high-specificity and high-affinity/low-specificity regions, yielding optimal results. Notably, the Pareto-based method demonstrated significantly superior performance in affinity optimization, while also achieving commendable results in specificity across the 3 property groups. In contrast, other approaches, such as the weighted sum and

Supplementary Table 3. Adjusted *P*-values for LogAUC, MFI, EC50, and Tm comparisons between optimized variants and parental antibodies. Statistical significance was assessed by one-way ANOVA with Dunnett's two-sided multiple comparisons test.

Sample	Control	LogAUC	MFI	EC50	Tm
Ab1-2	Ab1	0.0003	0.0084	< 0.0001	N.A.
Ab1-3	Ab1	0.0005	0.0164	< 0.0001	N.A.
Ab2-2	Ab2	< 0.0001	0.0003	< 0.0001	N.A.
Ab2-3	Ab2	< 0.0001	0.0011	< 0.0001	N.A.
Ab2-4	Ab2	< 0.0001	0.0005	0.0008	N.A.
Ab3-1	Ab3	< 0.0001	< 0.0001	0.0013	< 0.0001
Ab3-2	Ab3	0.2700	< 0.0001	< 0.0001	0.0134
Ab3-3	Ab3	< 0.0001	< 0.0001	< 0.0001	0.2950
Ab3-4	Ab3	0.0035	< 0.0001	< 0.0001	< 0.0001
Ab4-2	Ab4	N.A.	N.A.	0.0004	N.A.
Ab4-3	Ab4	N.A.	N.A.	0.0007	N.A.
Ab4-4	Ab4	N.A.	N.A.	< 0.0001	N.A.

All values are adjusted *P* values. Values smaller than 0.0001 are reported as < 0.0001. N.A., not applicable.

Supplementary Table 4. Ranking analysis based on 4 starting groups in co-optimization of affinity and specificity.

Low-Low				Low-High			
Model	Top 5	Top 10	Top 15	Model	Top 5	Top 10	Top 15
AbOptimizer	5	8	9	AbOptimizer	4	8	12
VS	0	0	0	VS	0	0	0
BO	0	0	0	BO	0	0	0
GA	0	0	0	GA	0	0	0
ProtLLM	0	0	0	ProtLLM	0	0	0
ProtLLM*	0	2	6	ProtLLM*	1	2	3
Median-Median				High-Low			
Model	Top 5	Top 10	Top 15	Model	Top 5	Top 10	Top 15
AbOptimizer	5	8	10	AbOptimizer	5	9	14
VS	0	0	0	VS	0	0	0
BO	0	0	0	BO	0	0	0
GA	0	0	0	GA	0	0	0
ProtLLM	0	0	0	ProtLLM	0	0	0
ProtLLM*	0	2	5	ProtLLM*	0	1	1

Low-Low, Low-High, Median-Median, and High-Low correspond to groups with low affinity and low specificity, low affinity and high specificity, median affinity and median specificity, and high affinity and low specificity, respectively.

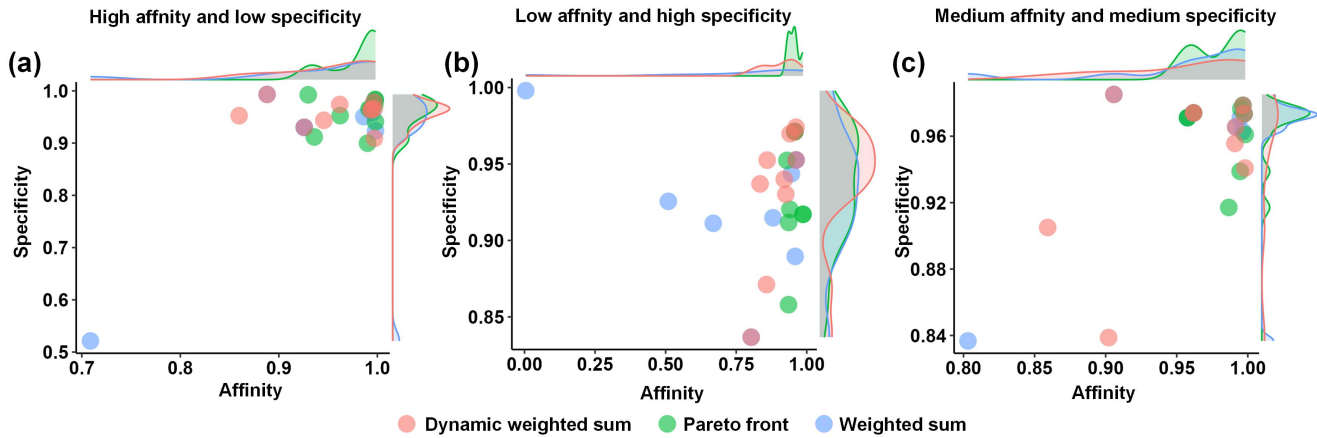
Supplementary Table 5. Ranking analysis of optimized variants from multi-property optimization

Model	Top 5	Top 10	Top 15
AbOptimizer	5	10	14
VS	0	0	0
BO	0	0	0
GA	0	0	0
ProtLLM	0	0	0
ProtLLM*	0	0	1

dynamic weighted sum methods, struggled to effectively manage the trade-offs between affinity and specificity, resulting in a predominance of variants optimized for only one property.

Supplementary Table 6. Performance of different methods across multiple optimization objectives in siltuximab. Values for optimized variants are reported as mean (SD) across three random seeds. The wild-type siltuximab value is reported as a single baseline score. SD values smaller than 0.001 are shown as < 0.001.

Method	Fitness	MHC	FvNetCharge	FvCSP	HISum
siltuximab (WT)	0.492	0.999	0.999	0.143	0.066
AbOptimizer	0.526 (0.005)	0.999 (< 0.001)	0.953 (0.017)	0.870 (0.015)	0.994 (0.001)
BO	0.482 (0.008)	0.999 (< 0.001)	0.962 (0.016)	0.863 (0.013)	0.995 (< 0.001)
GA	0.490 (0.009)	0.999 (< 0.001)	0.964 (0.017)	0.860 (0.015)	0.994 (0.002)
ProtLLM*	0.464 (0.002)	0.999 (< 0.001)	0.966 (0.033)	0.812 (0.105)	0.993 (0.004)
ProtLLM	0.458 (< 0.001)	0.999 (< 0.001)	0.978 (< 0.001)	0.062 (< 0.001)	0.995 (< 0.001)
VS	0.480 (0.006)	0.999 (< 0.001)	0.968 (0.007)	0.858 (0.008)	0.995 (< 0.001)



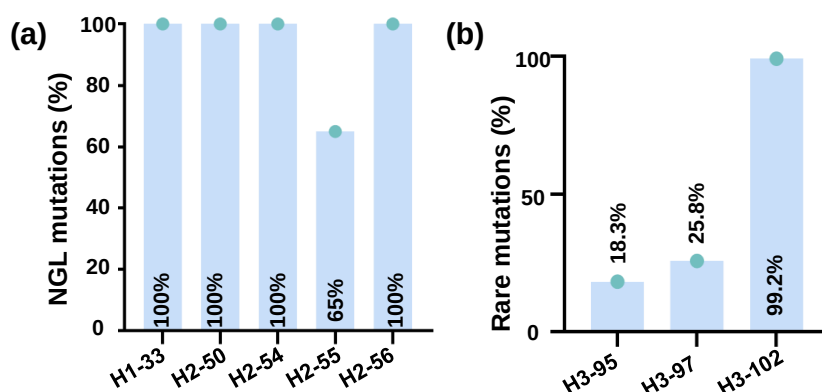
Supplementary Fig. 4. Comparison of different reward mechanisms for affinity and specificity optimization under three groups: **a**, high affinity and low specificity, **b**, low affinity and high specificity, and **c**, medium affinity and medium specificity.

2.4 Originality and germline bias analyses of optimized variants

In the co-optimization task, we assessed the originality of mutations by analyzing position-specific residue frequencies from MSAs built on OAS²⁷ and UniRef90²⁸. Following Hie et al.¹⁹, originality was quantified as the observed frequency of a given residue at a specific position in reference sequence databases, including a V/J-matched OAS sequence set and UniRef90, where lower frequencies correspond to higher originality. The V/J-matched OAS sequence set was processed as previously described²⁹. OAS antibody sequences downloaded on March 31, 2026, were filtered to exclude duplicates, sequences lacking conserved cysteines, heavily fragmented sequences with >16 N-terminal or >7 C-terminal residues missing, and singletons. Sequences with the same V and J genes as emibetuzumab were retained, yielding 10,429 background sequences for novelty analysis. A mutation is considered ‘rare’ if the wild-type residue frequency is > 5% and the mutant residue frequency is ≤ 5% (Yellow highlights in Supplementary Table 7). Supplementary Table 7 displays five representative sequences containing rare substitutions in the V/J-matched OAS sequence set and UniRef90. Across all 120 variants, approximately 20% contained at least one mutation in which a wild-type residue with a frequency greater than 5% was replaced by a rare residue with a frequency below 5% in the V/J-matched OAS sequence set. A similar trend was observed using UniRef90, where this proportion increased to 31%.

To further assess whether AbOptimizer is influenced by germline bias, we divided the analysis into two parts, as non-CDR-H3 positions are largely constrained by germline-encoded V genes, whereas CDR-H3 positions are strongly affected by loop length and junctional diversity. Outside CDRH3, we compared each mutation in the designed variants against the closest inferred human germline residue to assess whether AbOptimizer proposes germline or non-germline mutations (NGLs). Consistent with previous work²⁹, we first used ANARCI³⁰ to infer the closest human germline assignment for the heavy chain of emibetuzumab. Based on the inferred germline sequence (IGHV1-2*06 and IGHJ4*01 for the V and J genes, respectively), we annotated mutations outside CDRH3 in the 120 designed variants as either germline or non-germline. We then calculated the proportion of NGL mutations among all mutations to assess whether AbOptimizer tends to propose mutations biased toward the germline. We found that NGL mutations were highly enriched among the 120 optimized sequences (Supplementary Fig. 5a).

Except for position H2-55, where the proportion of NGL mutations was 65%, all other analyzed positions showed an NGL proportion of 100%. This analysis indicates that AbOptimizer does not simply favor germline-reverting substitutions. Within CDRH3, we assessed the frequency of amino acid residues at each position using OAS sequences matched for V gene, J gene, and CDRH3 length (22 sequences in total). Amino acids with a background frequency of $\leq 5\%$ were defined as rare. We further quantified, across 120 variants, the proportion of sequences containing rare amino acids at positions H3-95, H3-97, and H3-102. As shown in Supplementary Fig. 5b, the proportion of rare mutations was relatively low at H3-95 and H3-97, but reached as high as 99.2% at H3-102. However, the presence of rare mutations across the three positions suggests that AbOptimizer is not strictly biased toward germline or high-frequency residues, but can explore less common sequence space.



Supplementary Fig. 5. Germline-bias analysis of optimized antibody variants. **a**, Mean percentage of non-germline-like (NGL) mutations outside CDRH3 across 120 optimized sequences. **b**, Mean mutation frequency in CDRH3 across 120 optimized sequences.

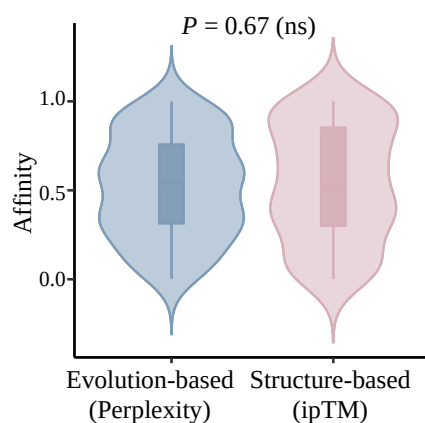
Supplementary Table 7. Originality analysis of five representative emibetuzumab variants generated by co-optimization

Source sequences	Optimized sequences	Optimized affinity	Optimized specificity	Substitution	V/J-matched OAS subset		UniRef90	
					WT freq	MT freq	WT freq	MT freq
DAGGGYSV	YRGGGDAV	0.96	0.95	D33Y	<1%	34%	6%	3%
				A50R	<1%	89%	5%	5%
				Y95D	<1%	17%	4%	5%
				S97A	6%	4%	8%	4%
YRKGGDLY	YRTGGDLV	0.91	0.99	K54T	<1%	1%	5%	7%
				Y102V	31%	1%	4%	12%
YRKGTAGD	YRTGGDAV	0.96	0.97	K54T	<1%	1%	5%	7%
				T56G	<1%	17%	7%	6%
				A95D	5%	16%	4%	5%
				G97A	5%	5%	9%	4%
				D102V	1%	31%	6%	12%
YKTKSDGV	YRTGGDPV	0.94	0.97	K50R	<1%	89%	5%	5%
				K55G	<1%	93%	5%	7%
				S56G	1%	17%	9%	6%
				G97P	5%	5%	9%	3%
SRGGGVGY	YRTGGDSY	0.92	0.97	S33Y	3%	33%	6%	3%
				G54T	2%	11%	7%	7%
				V95D	2%	15%	11%	5%
				G97S	5%	5%	9%	8%

'WT freq' denotes the Wild-type residue frequency, and 'MT freq' denotes the Mutant residue frequency. Yellow highlights indicate rare mutations.

2.5 Discussion of generic surrogates for AbOptimizer-guided optimization

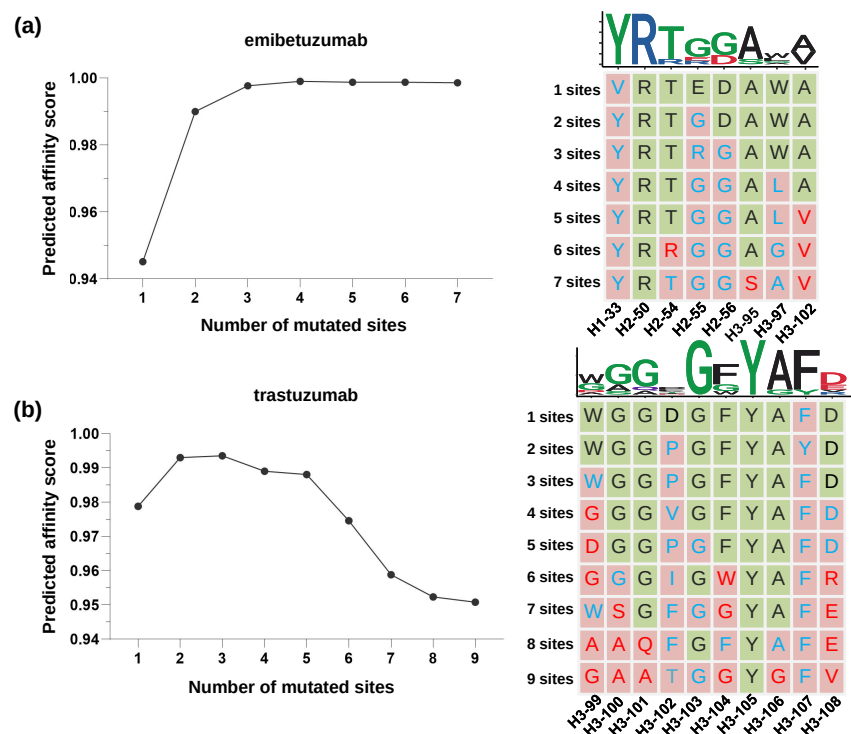
We are interested in exploring whether incorporating Boltz or additional AlphaFold-family models as a surrogate in AbOptimizer could improve the performance. To evaluate how different generic property predictors influence the optimization outcomes of AbOptimizer, **we compared two guidance signals: a structure-based score derived from the interface predicted TM-score (ipTM) generated by an antibody–antigen complex structure prediction model, and an evolution-based likelihood score (or perplexity) generated by a structure-informed sequence model.** These two signals were selected because prior studies have shown that both are informative of experimentally measured properties, including binding affinity and other biophysical characteristics^{8,19,31,9}. We chose IntelliFold-2 as the structure prediction model because it is an open-source, state-of-the-art model for antibody–antigen complex prediction, with reported performance comparable to or exceeding that of AlphaFold 3³². We chose a structure-informed language model based on an inverse folding framework (ESM-IF1 GVP-Transformer) because, among existing language models, it explicitly incorporates structural information and has previously been demonstrated to be effective for antibody optimization⁸. We used the same experimental setup as described in the manuscript for affinity maturation experiments, where 90 different starting sequences were used, and for each starting sequence, 128 optimized variants were generated. The top 1 sequence was selected as the optimized variant for comparison. Each test was run with three different random seeds. For the optimized variants, we then checked downstream performance using an affinity predictor trained on the emibetuzumab DMS dataset, which achieved a Spearman correlation of 0.87 on held-out data. The resulting variants exhibited highly similar affinity distributions (Supplementary Fig. 6). **The mean affinity was 0.52 for perplexity-guided optimization and 0.54 for ipTM-guided optimization, with no statistically significant difference ($p = 0.67$), indicating comparable optimization performance under both guidance signals.** This observation suggests that, although these guidance signals originate from fundamentally different modeling paradigms, they provide similarly informative optimization signals for sequence search. Consequently, **AbOptimizer appears robust to the choice of generic surrogate, and its performance is not strongly dependent on a specific predictor when high-fidelity property models are unavailable.** Notably, ipTM-guided optimization was substantially more computationally expensive, with an inference speed approximately 15-fold slower than that of perplexity-guided optimization. Therefore, these results indicate that, in the context of fitness optimization, AbOptimizer using the ESM-IF1 GVP-Transformer achieves optimization performance comparable to that obtained with state-of-the-art antibody–antigen complex structure prediction models, but with substantially higher computational efficiency. Thus, it provides a more practical choice for efficient antibody optimization.



Supplementary Fig. 6. Comparison of the affinity of variants optimized using structure-based (ipTM) and evolution-based (perplexity) guidance. Structure-based guidance was derived from the ipTM score predicted by IntelliFold-2, whereas evolution-based guidance was derived from the perplexity score predicted by the ESM-IF1 GVP-Transformer. Results are shown for 90 affinity-optimized sequences. The embedded boxplots indicate the median and interquartile range. Statistical significance was calculated using a two-sided paired Student's t-test.

2.6 Multi-site mutagenesis achieved by AbOptimizer

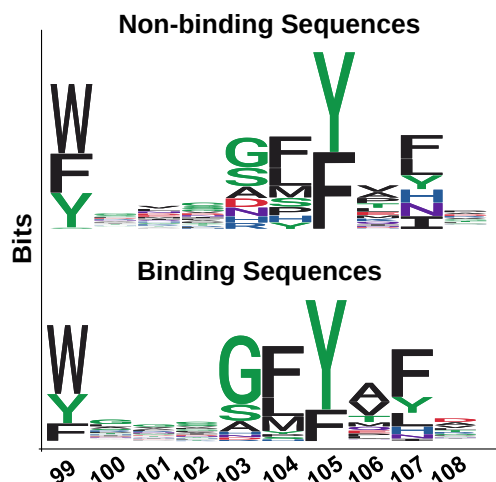
The sequence editing method employed in our AbOptimizer is inherently highly controllable, which facilitates an in-depth analysis of how the model performs parallel mutagenesis across multiple sites. We evaluated this property using both the emibetuzumab and trastuzumab datasets. For emibetuzumab, we selected a high-affinity antibody variant (“VRTEDAWA”)



Supplementary Fig. 7. Influence of the number of mutation sites on optimization performance for emibetuzumab and trastuzumab. **a**, Optimization performance and mutation patterns for emibetuzumab. **b**, Optimization performance and mutation patterns for trastuzumab. Red squares indicate substitution actions; green squares indicate retention actions; blue letters indicate beneficial point mutations; red letters indicate deleterious point mutations.

obtained from the affinity maturation task and systematically varied the number of mutated sites from 1 to 7 to assess improvements in affinity, while keeping all other experimental settings consistent with the original affinity maturation protocol. For trastuzumab, we employed the variant “WGGDGFYAMD” as the experimental subject and controlled the number of mutated sites from 1 to 9 to investigate the model’s capability in optimizing specificity.

Supplementary Fig. 7 shows the editing actions assigned by AbOptimizer at each position. Red and blue squares denote ‘substitution’ and ‘reservation’ actions, respectively. When the amino acid under a red square is identical to the original residue, the position was edited back to the wild-type residue, although the action was still classified as ‘substitution’. As the mutation number increased, AbOptimizer largely retained previously identified mutation sites, while varying the residue types at these sites. We further assessed the effect of each individual point mutation to determine whether AbOptimizer preferentially combined beneficial mutations; blue letters indicate beneficial mutations, whereas red letters indicate deleterious mutations. In emibetuzumab, when the number of mutation sites is set to 4, the optimized antibody attains the highest affinity (Supplementary Fig. 7a left). For mutation site counts exceeding 1, the relative affinity of the optimized antibodies consistently exceeds 0.99, indicating that the AbOptimizer algorithm effectively manages the number of mutation sites while maximizing the optimization of antibody properties. Supplementary Fig. 7a right depicts the specific changes in mutation sites across varying mutation site counts in emibetuzumab. In trastuzumab, a similar phenomenon was observed: with each increment in the number of mutation sites, previous mutations were generally retained. In Supplementary Fig. 7b (left), the strongest binding affinity was achieved at 3 mutation sites. Likewise, in Supplementary Fig. 7b (right), deleterious single-point mutations began to emerge when the number of mutation sites exceeded 4. These findings suggest that AbOptimizer’s parallel multi-site mutagenesis tends to combine beneficial single-point mutations. However, as the externally imposed mutation number increases, AbOptimizer also explores deleterious single-point mutations. Notably, when combined with other beneficial mutations, these deleterious mutations neither diminish nor enhance the binding affinity of the variants.



Supplementary Fig. 8. The sequence logo of existing wet-lab affinity dataset of trastuzumab.

3 Optimized candidates for bioassay experiments

3.1 Details of selection rules for experiment validation

Apart from the wild type emibetuzumab (Ab1), another lead sequence (Ab2) exhibiting high affinity and significant non-specific binding was randomly selected. For both lead sequences, 5 optimized variants were chosen; however, the selection criteria for Ab1 and Ab2 differ slightly.

For Ab1, the emibetuzumab wild type demonstrated an exceptionally high predicted affinity of 0.99866 (as indicated by the surrogate model) and a low predicted specificity of 0.00008. To address the optimization challenge, 2 iterations were conducted for Ab1. The top 10 variants of the first iteration, based on the cumulative affinity and specificity scores, was selected for the second optimization iteration. Subsequently, 5 variants were chosen from the results for wet experimental validation. Among these 5 variants, 2 were the best performers based on their ranking in the combined affinity and specificity scores. To ensure high affinity, 3 additional variants were selected that exhibited top 3 affinity scores while maintaining specificity scores higher than the starting sequence.

For Ab2, only one iteration was performed, and the 5 variants were selected according to the aforementioned criteria.

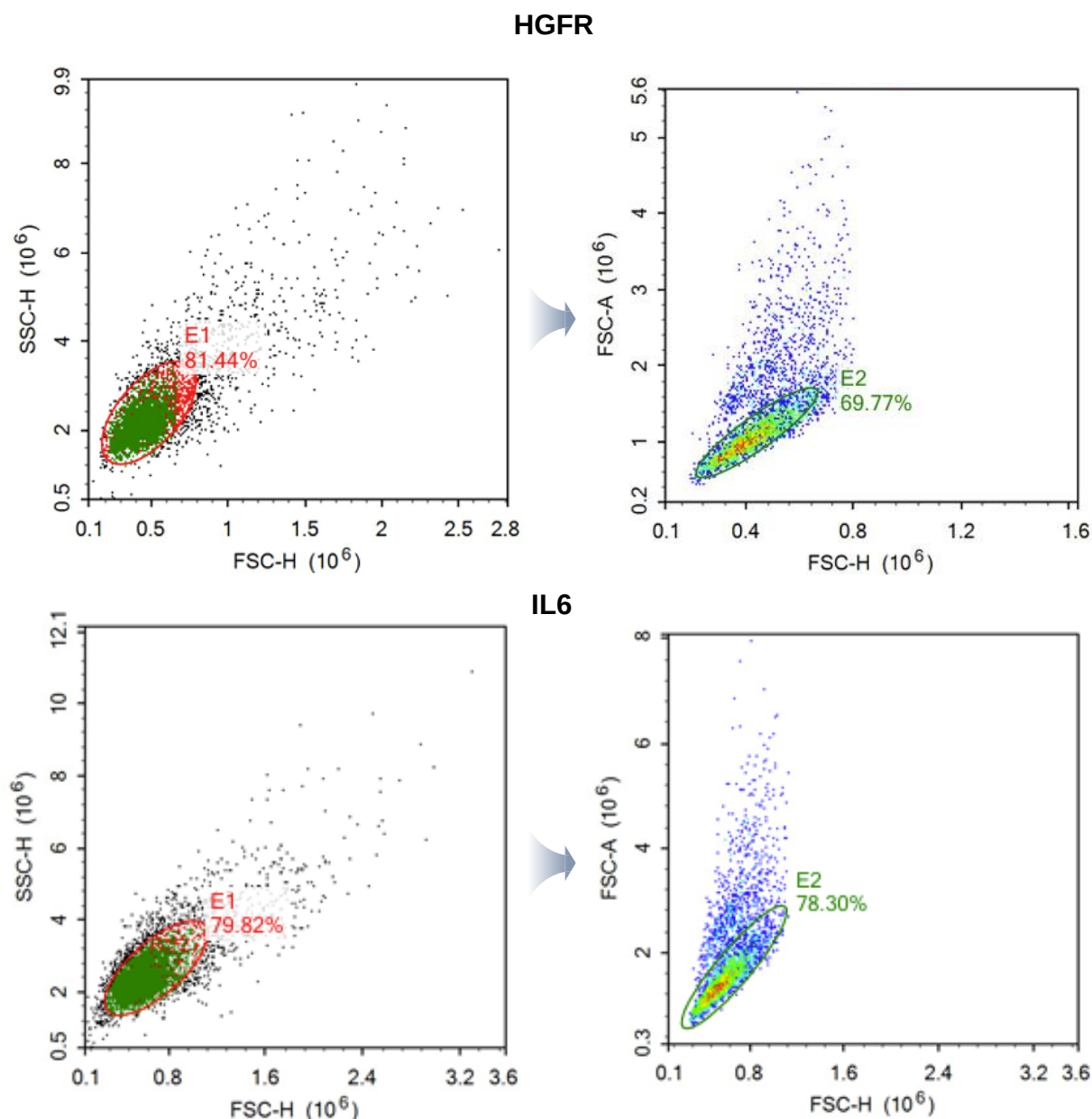
For Ab3, the therapeutic antibody siltuximab exhibits strong binding affinity. We guided its optimization using a general fitness-predictor model (see Supplementary section 1.3), which primarily indicates whether a mutation increases the probability of the mutated sequence folding into a stable backbone structure. Building on the already high affinity and leveraging this fitness predictor, we did not restrict mutations to the CDRH3 region, which largely contributes to affinity; instead, we conservatively introduced 1–3 mutations across CDRH1, CDRH2, and CDRH3. Furthermore, we applied 1–3 mutations spanning the entire CDRH region to explore multi-point mutations across CDRHs. In total, we generated four groups of mutations, selecting the Top-2 variants from each group for experimental validation, yielding eight sequences in total.

For Ab4, AbOptimizer was permitted to introduce four editing actions into trastuzumab. This substantially expanded the search space, thereby increasing both the opportunity to identify variants with superior properties and the risk of generating deleterious mutations. To improve candidate prioritization, we further re-ranked the surrogate-selected variants using the complex *ipTM* predicted by IntelliFold-v2³² flash and the binding $\Delta\Delta G$ calculated by Rosetta. By jointly incorporating a structural confidence metric and a physicochemical energy metric, we prioritized the variants identified by AbOptimizer and selected the top five candidates for experimental validation. The final composite score was defined as:

$$\text{Final score} = \text{Surrogate score} + ipTM - 0.01 \times \Delta\Delta G$$

3.2 Detailed results of molecular docking

In this study, we first performed docking of the antibody with HGFR using ZDOCK 3.0.2. Subsequently, the docking conformations generated by ZDOCK were rescored using ZRANK to identify a more stable binding complex. As shown in



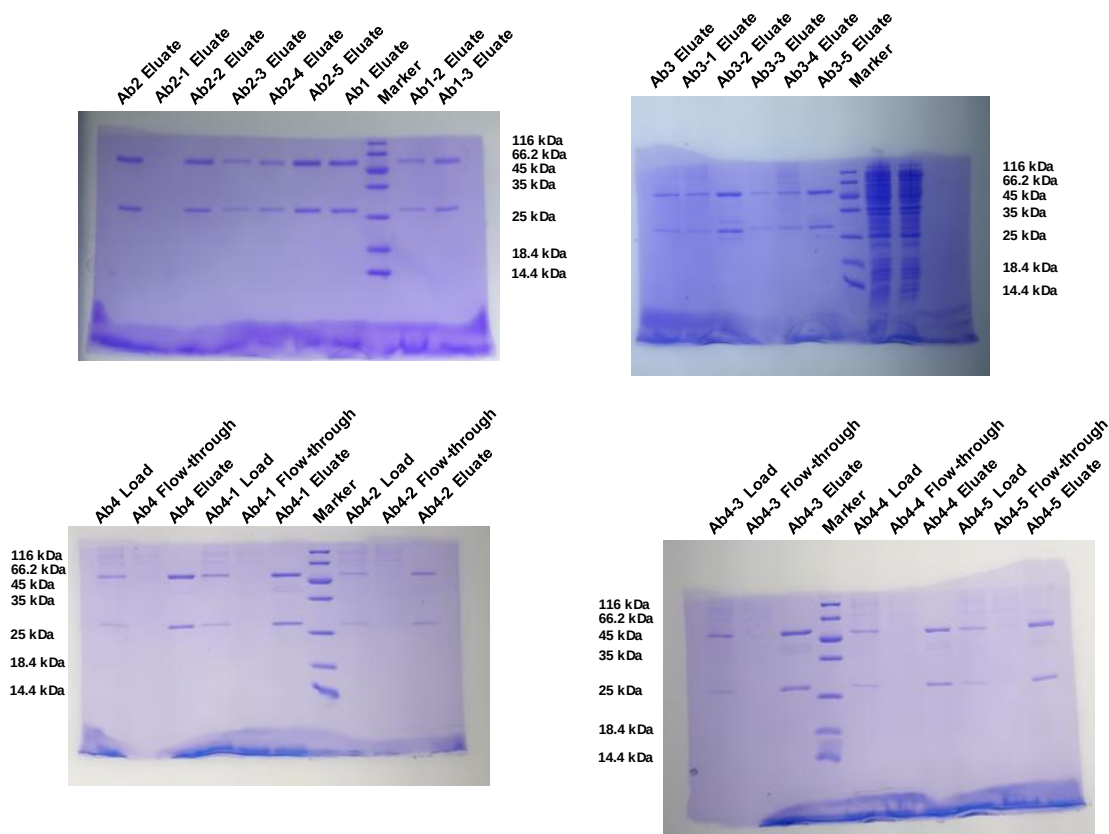
Supplementary Fig. 9. Representative sequential gating strategies for CHO cell-based flow cytometry. Representative gating strategies are shown for CHO cells used in the HGFR and IL6 antibody specificity assays. Cells were first gated on FSC-H versus SSC-H to identify the main cell population (E1), followed by singlet gating on FSC-H versus FSC-A to exclude cell aggregates and doublets (E2). The frequency of events within each gated population is indicated in the corresponding gate. The same gating strategy was applied to all samples within each experiment. The HGFR gating panels correspond to the CHO-cell-based flow cytometry MFI data shown in Fig. 5d and Fig. 5g in main text, and the IL6 gating panels correspond to the CHO-cell-based flow cytometry MFI data shown in Fig. 5j in main text. The experiment was repeated independently three times with similar results.

Supplementary Table 8, the ZDOCK scores for Ab1 and its variants are approximately 10, while their ZRANK scores are around -40. Given the potential deviations associated with evaluating molecular interactions through docking software, we conclude that Ab1 and its variants yield nearly identical docking results and will thus be retained for further validation. However, as illustrated in Supplementary Table 8, a significantly lower ZRANK score was observed for Ab2-5 (-3.95) compared to Ab2

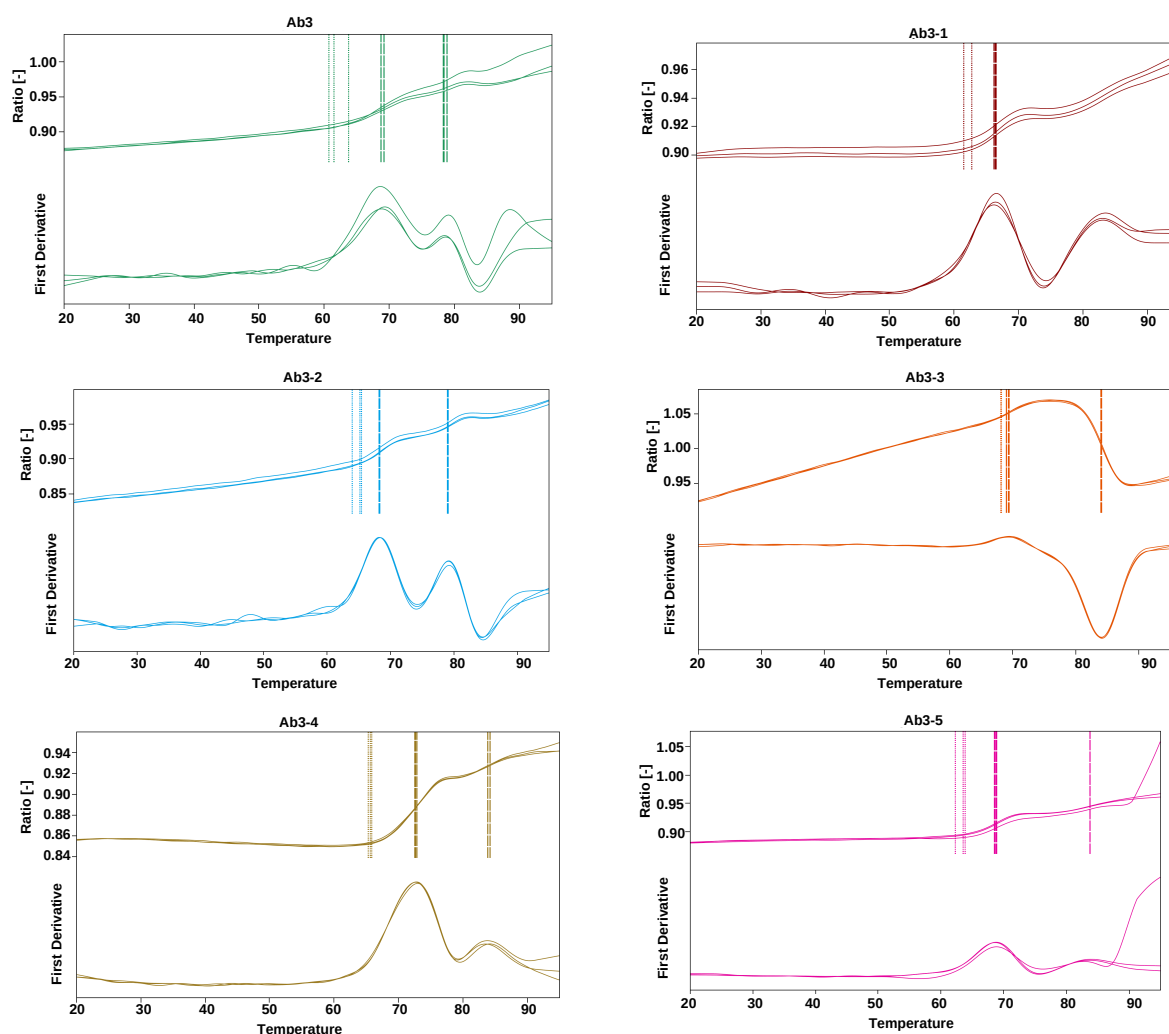
(-54.745). Consequently, Ab2-5 was excluded from further experimental validation due to its markedly inferior docking score.

Supplementary Table 8. Results of molecular docking score

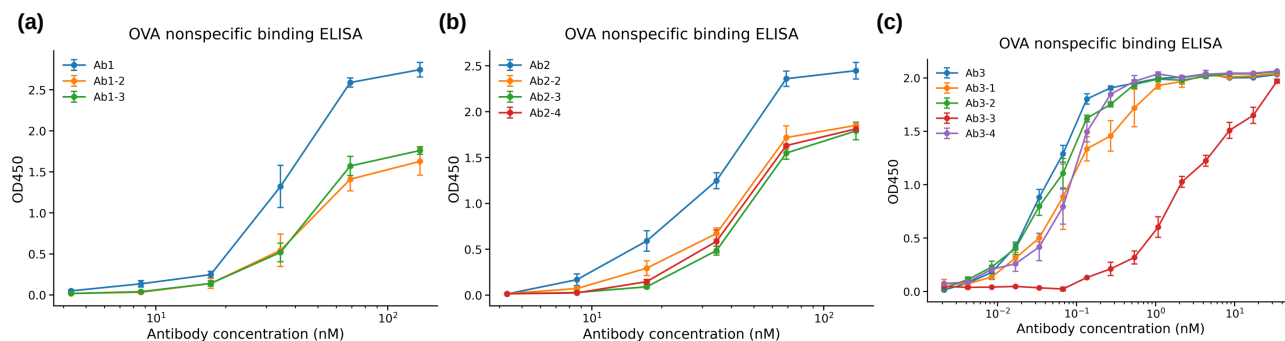
	ZDock score	ZRank Score
Ab1	10.12	-47.5
Ab1-1	11.2	-35.596
Ab1-2	10.98	-39.581
Ab1-3	10.74	-42.373
Ab1-4	9.64	-50.994
Ab1-5	10.14	-33.397
Ab2	10.4	-54.745
Ab2-1	9.94	-53.718
Ab2-2	9.22	-53.785
Ab2-3	10	-78.549
Ab2-4	10.62	-40.492
Ab2-5	9.62	-3.95



Supplementary Fig. 10. Representative reducing SDS-PAGE analysis of selected antibody preparations after Protein A purification. Load, flow-through and eluate fractions are shown where available. The purified eluate fractions exhibited the expected IgG heavy-chain and light-chain bands at approximately 50 kDa and 25 kDa, respectively, supporting successful antibody expression and purification. Molecular weight markers are indicated. The experiment was repeated independently three times with similar results.



Supplementary Fig. 11. Raw thermal unfolding profiles used for T_m determination of Ab3-derived variants.



Supplementary Fig. 12. OVA non-specific binding concentration-response profiles. OVA-based ELISA was used to visualize the concentration-dependent non-specific binding signals of the Ab1, Ab2, and Ab3 antibody systems against a non-cognate protein reagent. **a**, Ab1 and variants. **b**, Ab2 and variants. **c**, Ab3 and variants.

3.3 Details of SDS-PAGE and SEC characterization of purified antibody variants

Representative SDS-PAGE analysis was performed as part of the quality-control workflow for expressed antibody variants (Supplementary Fig. 10). Compared with the load and flow-through fractions, the eluate fractions showed clear enrichment of

bands corresponding to the expected IgG heavy and light chains, indicating successful capture and purification by Protein A chromatography. Although minor weak bands were observed in some lanes, the dominant band pattern was consistent with purified IgG preparations, supporting their suitability for subsequent SEC, ELISA and flow cytometry analyses.

SEC analysis was performed to further assess the monomeric purity of purified antibody variants following SDS-PAGE quality control (Supplementary Table 9). All antibody preparations showed monomeric species exceeding 90% of the total integrated SEC peak area, with monomer contents ranging from 90.07% to 98.85%. Most variants exhibited high monomeric purity, and several preparations exceeded 98%, indicating limited aggregation or degradation after purification. Together with the expected SDS-PAGE band patterns, these SEC results support the overall quality of the purified antibodies and their suitability for downstream ELISA, flow cytometry and thermal stability analyses.

Supplementary Table 9. SEC analysis of monomeric purity for purified antibody variants

Variants	Monomer (%)
Ab1	90.07%
Ab1-2	93.16%
Ab1-3	93.92%
Ab2	95.99%
Ab2-2	90.49%
Ab2-3	94.35%
Ab2-4	97.03%
Ab3	93.48%
Ab3-1	98.85%
Ab3-2	97.60%
Ab3-3	94.35%
Ab3-4	98.75%
Ab3-5	98.59%
Ab4	95.76%
Ab4-1	90.08%
Ab4-2	96.97%
Ab4-3	92.54%
Ab4-4	95.02%
Ab4-5	93.12%

3.4 Details of EC50 fitting parameters for antigen-binding ELISA curves

Antigen-binding ELISA curves were fitted using a four-parameter logistic model, and the corresponding EC50 values and fitting parameters are summarized in Supplementary Table 10. The optimized variants generally showed lower EC50 values than their corresponding parental antibodies, supporting improved apparent antigen-binding potency. In the Ab1- and Ab2-derived groups, the optimized variants showed reduced EC50 values relative to their parental antibodies, consistent with improved HGFR-binding activity. Among the Ab3-derived variants, Ab3-4 exhibited the lowest EC50 value, indicating the strongest apparent IL6-binding improvement. For the Ab4-derived variants, Ab4-2, Ab4-3 and Ab4-4 all showed lower EC50 values than parental Ab4, supporting enhanced HER2-binding activity. All fitted curves showed high goodness of fit, with R^2 values above 0.993, indicating reliable EC50 estimation and supporting the antigen-binding curves presented in the main figures.

3.5 Details of thermal stability analysis

Representative raw nanoDSF curves (Supplementary Fig. 11) for Ab3 and Ab3-derived variants are presented to show the original data used for melting temperature determination. T_m values were calculated from the first-derivative peak of the unfolding transition, and these values were summarized in the corresponding thermal stability analysis.

References

1. Liu, L. *et al.* LY2875358, a Neutralizing and Internalizing Anti-MET Bivalent Antibody, Inhibits HGF-Dependent and HGF-Independent MET Activation and Tumor Growth. *Clin. Cancer Res.* **20**, 6059–6070, DOI: [10.1158/1078-0432.CCR-14-0543](https://doi.org/10.1158/1078-0432.CCR-14-0543) (2014).

Supplementary Table 10. EC₅₀ fitting parameters of antigen-binding ELISA curves for optimized antibody variants, reported as mean values from three biologically independent replicates.

Sample	EC ₅₀ ± SD (nM)	95% CI (nM)	Hill slope	Bottom	Top	R ²
Ab1	1.327 ± 0.073	0.957–1.872	1.121	0.027	2.283	0.996
Ab1-2	0.920 ± 0.051	0.606–1.418	1.141	0.000	2.266	0.992
Ab1-3	0.777 ± 0.024	0.432–1.708	0.916	0.140	2.411	0.991
Ab2	1.517 ± 0.130	0.932–3.000	1.237	0.019	2.401	0.990
Ab2-2	0.960 ± 0.040	0.675–1.407	1.313	0.030	2.403	0.993
Ab2-3	0.851 ± 0.026	0.719–1.008	1.728	0.090	2.450	0.998
Ab2-4	1.095 ± 0.088	0.759–1.609	1.378	0.047	2.459	0.992
Ab3	0.215 ± 0.035	0.154–0.305	1.053	0.170	2.250	0.997
Ab3-1	0.095 ± 0.012	0.055–0.208	0.780	0.080	2.289	0.994
Ab3-2	0.051 ± 0.010	0.040–0.064	1.937	0.235	2.170	0.994
Ab3-3	0.056 ± 0.016	0.019–0.181	0.671	0.073	2.443	0.985
Ab3-4	0.016 ± 0.001	0.008–0.038	0.881	0.136	2.248	0.989
Ab4	2.607 ± 0.276	1.744–3.943	1.529	0.006	2.197	0.991
Ab4-2	1.316 ± 0.237	1.088–1.595	1.949	0.084	2.252	0.998
Ab4-3	1.363 ± 0.138	1.052–1.773	1.626	0.073	2.257	0.996
Ab4-4	0.824 ± 0.103	0.628–1.101	1.173	0.021	2.331	0.998

- Nahta, R. e. & Esteva, F. Trastuzumab: triumphs and tribulations. *Oncogene* **26**, 3637–3643 (2007).
- van Rhee, F. *et al.* Siltuximab, a Novel Anti–Interleukin-6 Monoclonal Antibody, for Castleman’s Disease. *J. Clin. Oncol.* DOI: [10.1200/JCO.2009.27.2377](https://doi.org/10.1200/JCO.2009.27.2377) (2010).
- Bajwa, A. *et al.* Siltuximab for chimeric antigen receptor T-cell therapy–related CRS and ICANS: A multicenter retrospective analysis. *Blood Adv.* **9**, 170–175, DOI: [10.1182/bloodadvances.2024013688](https://doi.org/10.1182/bloodadvances.2024013688) (2025).
- Makowski, E. K. *et al.* Co-optimization of therapeutic antibody affinity and specificity using machine learning models that generalize to novel mutational space. *Nat. communications* **13**, 3788 (2022).
- Mason, D. M. *et al.* Optimization of therapeutic antibodies by predicting antigen specificity from antibody sequence via deep learning. *Nat. Biomed. Eng.* **5**, 600–612 (2021).
- Shanehsazzadeh, A. *et al.* Unlocking de novo antibody design with generative artificial intelligence. *BioRxiv* 2023–01 (2023).
- Shanker, V. R., Bruun, T. U. J., Hie, B. L. & Kim, P. S. Unsupervised evolution of protein and antibody complexes with a structure-informed language model. *Science* **385**, 46–53, DOI: [10.1126/science.adk8946](https://doi.org/10.1126/science.adk8946) (2024).
- Chungyoun, M. & Gray, J. Fitness landscape for antibodies 2: Benchmarking reveals that protein ai models cannot yet consistently predict developability properties. *bioRxiv* 2025–12 (2025).
- Amini, S. *et al.* Application of protein language models for antibody developability prediction. In *Mabs*, vol. 18, 2647489 (Taylor & Francis, 2026).
- Mirdita, M. *et al.* Colabfold: making protein folding accessible to all. *Nat. methods* **19**, 679–682 (2022).
- U.S. Food and Drug Administration, Center for Drug Evaluation and Research. Pharmacology/toxicology bla review and evaluation: Sylvant (siltuximab); bla 125496 (2014).
- Reynisson, B., Alvarez, B., Paul, S., Peters, B. & Nielsen, M. Netmhcp4n-4.1 and netmhci4n-4.0: improved predictions of mhc antigen presentation by concurrent motif deconvolution and integration of ms mhc eluted ligand data. *Nucleic acids research* **48**, W449–W454 (2020).
- Xu, X. *et al.* Ab-gen: Antibody library design with generative pre-trained transformer and deep reinforcement learning. *Genomics, Proteomics & Bioinforma.* (2023).
- Sharma, V. K. *et al.* In silico selection of therapeutic antibodies for development: viscosity, clearance, and chemical stability. *Proc. Natl. Acad. Sci.* **111**, 18601–18606 (2014).
- Bachas, S. *et al.* Antibody optimization enabled by artificial intelligence predictions of binding affinity and naturalness. *bioRxiv* 2022–08 (2022).
- Khan, A. *et al.* Toward real-world automated antibody design with combinatorial bayesian optimization. *Cell Reports Methods* 100374 (2023).
- Liu, G. *et al.* Antibody complementarity determining region design using high-capacity machine learning. *Bioinformatics* **36**, 2126–2133 (2020).

19. Hie, B. L. *et al.* Efficient evolution of human antibodies from general protein language models. *Nat. Biotechnol.* (2023).
20. Olsen, T. H., Moal, I. H. & Deane, C. M. Ablang: an antibody language model for completing antibody sequences. *Bioinforma. Adv.* **2**, vbac046 (2022).
21. Yang, J. *et al.* Improved protein structure prediction using predicted interresidue orientations. *Proc. Natl. Acad. Sci.* **117**, 1496–1503, DOI: [10.1073/pnas.1914677117](https://doi.org/10.1073/pnas.1914677117) (2020).
22. Pierce, B. G., Hourai, Y. & Weng, Z. Accelerating protein docking in zdock using an advanced 3d convolution library. *PLoS one* **6**, e24657 (2011).
23. Pierce, B. & Weng, Z. Zrank: reranking protein docking predictions with an optimized energy function. *Proteins: Struct. Funct. Bioinforma.* **67**, 1078–1086 (2007).
24. Ippolito, J. A., Alexander, R. S. & Christianson, D. W. Hydrogen bond stereochemistry in protein structure and function. *J. molecular biology* **215**, 457–471 (1990).
25. Arunan, E. *et al.* Defining the hydrogen bond: An account (iupac technical report). *Pure Appl. Chem.* **83**, 1619–1636 (2011).
26. Dwork, C., Kumar, R., Naor, M. & Sivakumar, D. Rank aggregation methods for the web. In *Proceedings of the 10th international conference on World Wide Web*, 613–622 (2001).
27. Olsen, T. H., Boyles, F. & Deane, C. M. Observed antibody space: A diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. *Protein Sci.* **31**, 141–146 (2022).
28. Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R. & Wu, C. H. Uniref: comprehensive and non-redundant uniprot reference clusters. *Bioinformatics* **23**, 1282–1288 (2007).
29. Olsen, T. H., Moal, I. H. & Deane, C. M. Addressing the antibody germline bias and its effect on language models for improved antibody design. *Bioinformatics* **40**, btae618 (2024).
30. Dunbar, J. & Deane, C. M. Anarci: antigen receptor numbering and receptor classification. *Bioinformatics* **32**, 298–300 (2016).
31. Bennett, N. R. *et al.* Atomically accurate de novo design of antibodies with rfdiffusion. *Nature* **649**, 183–193 (2026).
32. Qiao, L., Yan, H., Liu, G., Guo, G. & Sun, S. Intellifold-2: Surpassing alphafold 3 via architectural refinement and structural consistency. *bioRxiv* 2026–02 (2026).