nature biotechnology

Article

https://doi.org/10.1038/s41587-026-03238-6

A blinded, prospective benchmark of in silico antibody discovery anchored to experimental affinity and developability

Received: 28 February 2026

Accepted: 24 June 2026

Published online: 19 August 2026

Check for updates

A list of authors and their affiliations appears at the end of the paper

Experimentally validated prospective, blinded benchmarks are needed to separate durable advances from hype in computational antibody design. Here AIntibody, a challenge inspired by the Critical Assessment of Structure Prediction, tests 511 artificial intelligence (AI)-designed or predicted antibodies from 29 organizations on three tasks: in silico affinity maturation from phase 1 sequencing outputs, affinity ranking within heavy-chain complementarity-determining region 3 (HCDR3) clusters of a selection output and CDR design of proteins not included in a selection output. Validated with diverse experimental assays, several groups produced developable antibodies with affinities <100 pM. However, these successes were exceptions that did not transfer across tasks. Affinity-matured antibodies were modeled effectively. Except for one model, predicting high-affinity clones from clustered HCDR3 datasets was worse than random clone picking. Out-of-library design was highly variable for most method submissions, with many failing to outperform standard selections. The AIntibody challenge shows that AI can optimize antibodies in defined, biologically grounded regimes, in addition to highlighting critical gaps including affinity prediction and library-inspired antibody design and cross-task generalization.

Benchmarks are decisive in separating durable advances from claims lacking independent experimental validation in computational biology. When evaluations are prospective, blinded and tied to experimental readouts, they create shared standards, expose method limitations and accelerate iterative improvements. The Critical Assessment of Structure Prediction (CASP) provided that template for protein structure prediction[1], culminating in the Nobel Prize winning advances of AlphaFold (AF) in CASP XIII (ref. 2) and the subsequent demonstration that near-experimental accuracy generalized beyond the assessment set[3]. CASP evaluated hundreds of diverse protein targets across multiple rounds with independent governance. Other similar benchmarking challenges include D3R[4], SAMPL[5], Adaptyv[6] and CACHE[7].

Despite intense activity, most performance benchmarks for computational antibody design rely on retrospective analyses without additional experiments. We previously highlighted[8] how this limits external validation and obscures the true state of the art and we argue for blind, prospective evaluation as a prerequisite to meaningful assessment. This becomes essential as the number of reported computational antibody design methods grows, particularly in non-peer-reviewed preprints and technical reports.

We launched the AIntibody challenge to address this gap. While AIntibody was inspired by the prospective, blinded CASP philosophy, there are important differences. For example, AIntibody's first iteration evaluated a single antigen, was organized by the same consortium now reporting the results and relied on organizer integrity rather than informatic safeguards for blinding. Therefore, AIntibody represents a first step toward establishing CASP-caliber benchmarking in antibody discovery and future iterations will further address these structural limitations. By anchoring evaluation in experimentally measured affinity and developability, this study, as well as other recent efforts such as

e-mail: mfrank.erasmus@iqvia.com; andrew.bradbury@iqvia.com

the Adaptyv EGFR binder competition, which is not restricted to antibodies, offers a reality check for the field, in addition to providing the underlying sequencing datasets for future comparative benchmarking.

Whereas structural prediction seeks a single geometric solution, therapeutic antibody engineering requires multiparametric optimization, which is only assessable experimentally. Beyond affinity and kinetics, these include a set of properties known as 'developability'[9,10], such as expression, melting temperature (T_m), polyreactivity and aggregation. The challenge lies in improving one property without diminishing others. Although there are many solutions to improve a specific antibody, there are many more ways to destroy it.

Unlike internet text, amino acid sequence is not a functionally readable 'language'; humans cannot infer functionality and pharmacology of an antibody sequence without additional experimentation. Moreover, antibody sequences are inherently biased, as the most frequent residue at any position in large datasets is typically germline. The exceptions are the heavy and light third complementarity-determining regions (CDR3s), which are a product of recombination, random nucleotide addition or subtraction and somatic hypermutation. Unlike natural language, the statistically dominant token is often not optimal for affinity improvement[11]. Training data require costly, time-consuming assays rather than abundant human-annotated and decipherable text or images. Prediction of affinity changes will require orders of magnitude of highly diverse experimental data, beyond those presently available[12]. Unfortunately, such discovery data are frequently siloed, encumbered by intellectual property constraints and not machine readable. Notwithstanding heroic efforts[13,14], public corpora remain both relatively small and biased in their content.

AIntibody was conceptualized as a prospective, blinded evaluation in which designed sequences were synthesized as full IgGs and tested under uniform protocols by independent laboratories: high-throughput surface plasmon resonance (HT-SPR) and KinExA for affinity and a standardized five-assay developability panel[8]. In contrast to retrospective studies, this approach directly compares outputs of different modeling strategies under the same conditions. Two challenges (affinity maturation and hit identification) directly address contemporary antibody discovery and optimization pipelines, while the third represents limited de novo antibody design. Can better antibodies be designed by modifying CDRs or CDR combinations when experimental sequencing data are provided?

The SARS-CoV-2 receptor-binding domain (RBD) is the most extensively characterized protein ever, with abundant structural and sequence data; accordingly, it was used in the first AIntibody challenge. To further favor computational success, sequencing datasets were deeper and richer than typically available at equivalent discovery stages and the second and third challenges included experimentally measured affinities normally available only later in a pipeline. The results presented here should be interpreted as an upper bound on current computational capability for this target class, not as evidence of general-purpose performance across arbitrary antigens or data regimes.

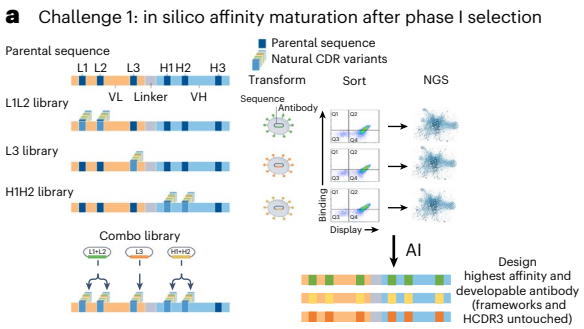
Results

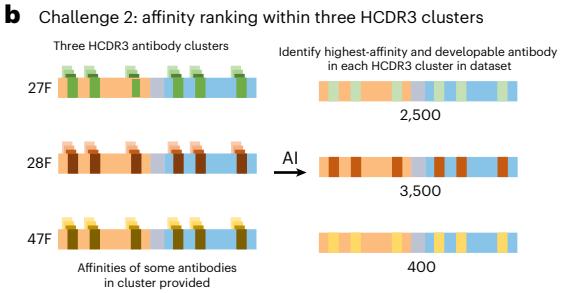
The AIntibody challenge comprised three tasks[8]. First, given the sequencing datasets of the final sorting outputs of an affinity maturation campaign with diversity in LCDR1 + LCDR2 (Fig. 1a), LCDR3 and HCDR1 + HCDR2 (Supplementary Data 1), generate high-affinity developable antibody sequences, modifying only HCDR1, HCDR2 and LCDR1–LCDR3, but not HCDR3 or the framework. Second, given the sequencing datasets of an HCDR3-clustered selection output (Supplementary Data 2 and Fig. 1b), identify the sequences of the highest-affinity developable antibodies in each of three different identified HCDR3 clusters. Third, given the full selection output of dataset 2 (Supplementary Data 2 and Fig. 1c), generate sequences of the highest-affinity developable antibodies that do not appear in the dataset by modifying CDRs but not frameworks (Supplementary Note 1). The results of these challenges are detailed below with all organizations assigned a challenge identifier (masked ID) except for challenge winners (notable exception is ProBioGen who agreed to be disclosed, for reasons described below), which are disclosed. Winning algorithms unsuccessfully used in other challenges are also disclosed to provide additional context for cross-challenge performance. A brief description of all computational methodologies across the different submissions and organizations is provided in the Methods (for challenge winners) and Supplementary Notes 2–6. Lastly, all challenge submissions distinguished by 'paper ID' across challenges, alongside controls, affinity measurements and developability, are provided in Supplementary Data 3.

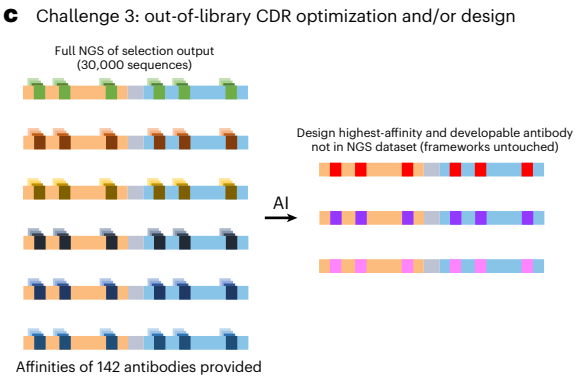
Figure 1d shows the number of organizations participating in each challenge. Figure 1e stratifies organizations by type, with the distribution of the 511 antibody sequence submissions by challenge and organization type in Fig. 1f. AI biotechs contributed the greatest number of sequences for challenge 3. In challenge 1 (affinity maturation), 2.5% of submitted variable light and heavy (V_L–V_H) amino acid sequences were independently submitted by more than one group, indicating minimal design convergence. The highest sequence submission redundancies (5–16%) were found in the prediction challenge, constrained by the 410–3,554 sequences in each cluster. No identical sequences were submitted for the out-of-library CDR design (challenge 3) (Fig. 1g), which, despite being consistent with the large designable sequence space, may also reflect algorithmic differences, the stochastic nature of generative sampling or a combination thereof. Figure 1h tabulates the total number of submissions, the number tested by SPR and a kinetic exclusion assay (KinExA) and the number of unique amino acid sequences by region. Some submissions were disqualified for not

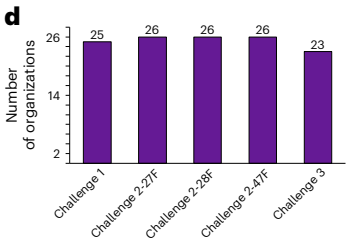
Fig. 1 | General submission details. a–**c**, Schematic depicting overview of challenge 1—in silico affinity maturation after phase I selection (design and optimization; **a**), challenge 2—affinity ranking with three HCDR3 clusters (prediction; **b**) and out-of-library CDR optimization and/or design (design and optimization; **c**). **d**, Distribution of the 29 participating organizations across the three distinct challenges: (1) in silico affinity maturation; (2) cluster-constrained affinity ranking; and (3) de novo out-of-library CDR design. **e**, Stratification of participants by organization type, showing representation from academic or nonprofit labs ($n = 10$), AI biotechs ($n = 8$), small or mid-size biotechs ($n = 5$), large pharma ($n = 5$) and big tech ($n = 1$). **f**, Total number of antibody sequences submitted ($n = 511$) broken down by organization for each challenge. **g**, Uniqueness of submitted V_L and V_H amino acid sequences, highlighting higher redundancy in challenges 1 and 2 compared to the unique, novel sequences in challenge 3. **h**, Summary table of total submissions versus unique sequences tested by SPR and KinExA. **i**, Snapshot of the winners for the five subcompetitions (challenges 1, 2-27F, 2-28F, 2-47F and 3) providing insights into their organization assignments, number of submissions, challenges that the group participated in and the challenges in which they were assigned a no. 1 ranking. **j**–**l**, Two-sided Spearman's rank correlation, with no adjustment for multiple comparisons; each AI-designed challenge submission was measured once on each platform. Pair-specific n, Spearman statistics and linear-fit slope on $\log_{10}(K_D)$: **j**, single-point KinExA versus SPR (x axis = K_D by SPR, y axis = K_D by single-point KinExA), $n = 95$, $\rho = 0.94$, $P = 2.0 \times 10^{-46}$, slope = 0.93; **k**, single-point KinExA versus standard KinExA (x axis = K_D by standard KinExA, y axis = K_D by single-point KinExA), $n = 52$, $\rho = 0.97$, $P = 6.5 \times 10^{-33}$, slope = 0.76; **l**, standard KinExA versus SPR (x axis = K_D by SPR, y axis = K_D by standard KinExA), $n = 52$, $\rho = 0.92$, $P = 6.8 \times 10^{-22}$, slope = 1.19. Pairs in **k**,**l** used the same $n = 52$ construct with both standard and single-point KinExA measurements available. **m**, Scoring criteria for developability across HIC, BVP, AC-SINS (dPW-adjusted), T_m and Tagg assays assigning bounds for passing (0), questionable (1) or failing (2) score according to direction of improved developability. **n**, HIC-HPLC data are well correlated between Mosaic and Jain et al.[9] for 15 IgG1 constructs (two-sided Pearson's correlation, $R^2 = 0.94$, $P = 2.5 \times 10^{-9}$). **o**, A composite score ≤ 3 indicates a developable candidate and a composite score > 3 indicates a fail or questionable status.

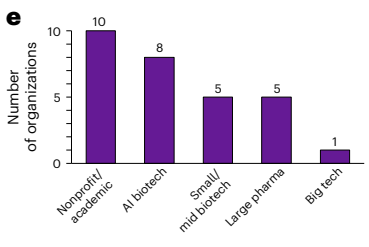
https://doi.org/10.1038/s41587-026-03238-6

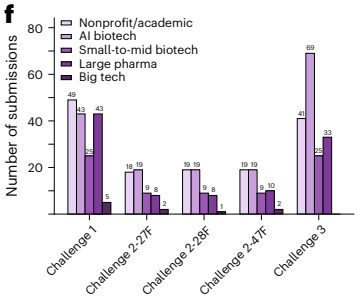


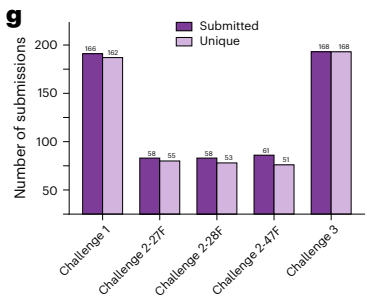












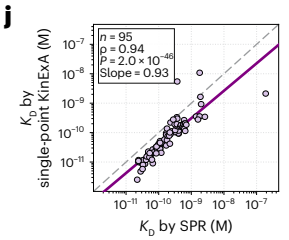
h

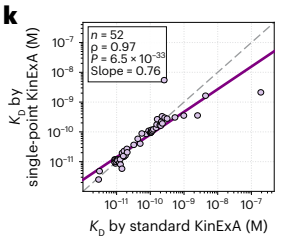
	Challenge	No. of submissions	No. of tested SPR	No. of tested KinExA	VL + VH	LCDR1	LCDR2	LCDR3	HCDR1	HCDR2	HCDR3
Source	1	–	31*	5*	185,097	5,184	7,337	6,055	2,528	1,023	1
Submitted		166	166	16	162	69	64	70	74	56	9
Source	2.27F	–	22	–	2,524	253	251	138	110	70	12
Submitted		58	58	20	55	26	20	10	11	10	3
Source	2.28F	–	14	–	3,554	311	301	180	142	88	5
Submitted		58	58	25	53	20	19	9	16	11	2
Source	2.47F	–	3	–	410	111	90	63	41	28	9
Submitted		61	61	12	51	23	15	11	9	7	5
Source	3	–	142	–	32,059	762	709	614	513	399	755
Submitted		168	168	18	168	85	89	70	76	65	89

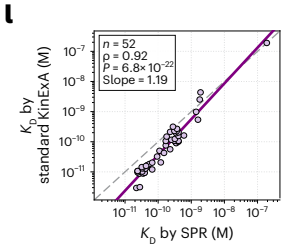
*Not provided to participants at time of challenge start

i

Company	Organization	No. of submissions	Challenges participated	challenge winner(s)
Aureka Biotechnologies	AI biotech	29	1, 2, 3	1
UCSD	Nonprofit /academic	18	1, 2, 3	2-27F
Scripps Research Institute	Nonprofit /academic	29	1, 2, 3	2-27F
Washington University in St. Louis	Nonprofit /academic	22	1, 2, 3	2-28F and 2-47F
Xencor	Small-to-mid biotech	29	1, 2, 3	2-28F and 3

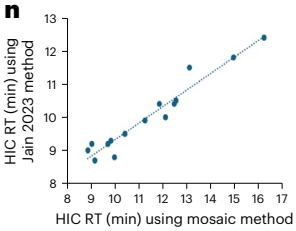






m

	Construct performance criteria		
	Passing (<)	Questionable	Failing (>)
HIC RT (min)	13.9	13.9–14.9	14.9
BVP score	5.3	5.3–10	10
AC-SINS	6.7	6.7–9.5	9.5
	Passing (>)	Questionable	Failing (<)
T_m	65	60–65	60
T_{agg}	64	60–64	60



o

Individual assay criteria	
Performance	Score
Passing	0
Questionable	1
Failing	2

Cumulative score criteria	
Performance	Score
Passing	≤3
Failing	≥4

following the rules in the announcement manuscript[8] (Supplementary Note 1). Figure 1i provides a snapshot (more details below) of challenge winners for the five subcompetitions, with the prediction challenges yielding multiple winners. While the datasets were generated in scFv format, antibodies were expressed and tested as full length IgGs.

Assay concordance and developability framework

Affinity measurements were first carried out on all submitted IgGs with SPR, followed by rank-order single-point KinExA and standard KinExA for the highest-affinity antibodies. Figure 1j shows strong rank-order correlation between SPR and single-point KinExA ($\rho = 0.94$, $P = 2.0 \times 10^{-46}$) and between single-point KinExA and standard KinExA ($\rho = 0.97, P = 6.5 \times 10^{-33}$) (Fig. 1k) and a slightly weaker correlation ($\rho = 0.92$, $P = 6.8 \times 10^{-22}$) between standard KinExA and SPR (Fig. 1l), confirming that definitive rank order could be established. Absolute affinities were less well correlated between either of the KinExA formats and SPR, with KinExA generally providing higher affinities.

While affinity was used to establish winning sequences, they also had to pass a developability threshold. Each antibody was assessed on five developability assays (affinity-capture self-interaction nanoparticle spectroscopy (AC-SINS), baculovirus particle (BVP) polyreactivity assay, hydrophobic interaction (HIC) retention, T_m and aggregation temperature (T_{agg})) and given a score of pass (0), questionable (1) or fail (2) for each assay (Fig. 1m,n) on the basis of clinical-stage reference antibody values[9]. Scores across the five assays were summed and constructs with a score ≤ 3 were considered developable, while those with a score > 3 were flagged as not developable and failing (Fig. 1o).

Challenge 1: in silico affinity maturation after phase I selection

The goal was to computationally affinity-mature a parental antibody recognizing SARS-CoV-2 RBD[15,16], without compromising developability, using sequencing datasets of improved antibody CDR sequences selected from HCDR1, HCDR2, LCDR1, LCDR2 and CDR3 libraries by yeast display (the first phase of a published affinity maturation protocol[17]) (Fig. 1a). Apart from the parental antibody, no affinities were provided. Antibody affinities were first established by SPR (Supplementary Fig. 1a,b), with corresponding isoaffinity plots in Fig. 2a. The affinities of the parental, experimentally affinity-matured[17] (not provided to participants) and best experimental antibodies are indicated, as well as submissions meeting (score ≤ 3) or failing (score > 3) developability criteria. Experimentally affinity-matured antibodies were selected from a combination library (generated by combining the selection outputs of the CDR sublibraries). Many of the highest-affinity designs were developable according to the challenge criteria but some of the strongest binders failed in at least one assay, predominantly BVP polyreactivity. Supplementary Fig. 2 shows the distribution of the developability properties of the designed affinity-matured antibodies, parental antibodies and top five experimentally derived control antibodies.

The SPR affinities of all AI antibodies are shown as box-and-jitter plots in Fig. 2b. Plots are ordered by the best affinity of each participating organization and compared to experimental antibodies. Although the overall distribution (for example, interquartile range (IQR)) of AI antibodies underperformed notably compared to experimental antibodies[17], the top computational design had an SPR affinity of 340 pM, compared to 517 pM for the best experimental antibody. The KinExA affinities (Fig. 2c) of the top five experimental antibodies ranged from 113 to 1,230 pM (95% confidence interval of the 113 pM antibody: 67.7–178 pM). The five best AI antibodies ranged from 95 to 984 pM (95% confidence interval of the 95 pM antibody: 69.1–127 pM). The overlapping 95% confidence intervals indicate that these values are statistically indistinguishable (Supplementary Fig. 3).

Figure 2d shows the cumulative affinity (SPR or KinExA) distributions of all and developable designed antibodies, with vertical lines representing the parental and best experimental affinities. The companion table (Fig. 2e) goes into greater detail, illustrating that most designed antibodies had higher affinity than the parental antibody and were developable.

The top five experimental and AI antibodies were all developable (Supplementary Fig. 2). Six CDR weblogos contrast the parental antibodies with strong binders, weak binders and nonbinders (Extended Data Fig. 1a), with the highest-affinity antibodies showing specific motifs (tyrosine at position 6 of LCDR1 and tryptophan at position 8 of LCDR3) suggesting that substitution of critical residues may greatly improve affinity; moreover, given the provided dataset, some AI algorithms were able to identify and include these substitutions in effective designs. The final list of the top ten designs, ranked first by KinExA and then by SPR, are shown in Extended Data Fig. 1b with complete ranking by masked ID, alongside submission profiles highlighted in Supplementary Table 1. Disqualified submissions (poor developability or not following rules; Supplementary Note 1) remain in the list, unranked. To assess the novelty of generated sequences, we calculated Levenshtein distances (LDs) of top submissions from the top 20 organizations compared to the parental antibody and the experimental dataset (Extended Data Fig. 1c). Notably, Aureka's winning submission was highly distinct, showing a concatenated CDR distance of 19 substitutions from the parental antibody and at least 12 from any sequence in the experimental dataset. Other top performers were more conservative, with concatenated CDR distances of 2–15 substitutions from the parental antibody and 1–9 substitutions from the experimental dataset. Similar comparisons to a consensus sequence comprising the most abundant residue at each CDR position from each CDR library's sort population (Extended Data Fig. 1d) showed that one of ProBioGen's submissions (ID 22), third overall by KinExA (540p M) with a perfect developability score of 0, was an exact match across all six CDRs (LD = 0). ProBioGen described their method not as based on AI but effectively a consensus sequence from the sequencing data, demonstrating that statistical analysis of the selection data, without machine learning (ML), was sufficient to produce one of the top-performing designs in challenge 1.

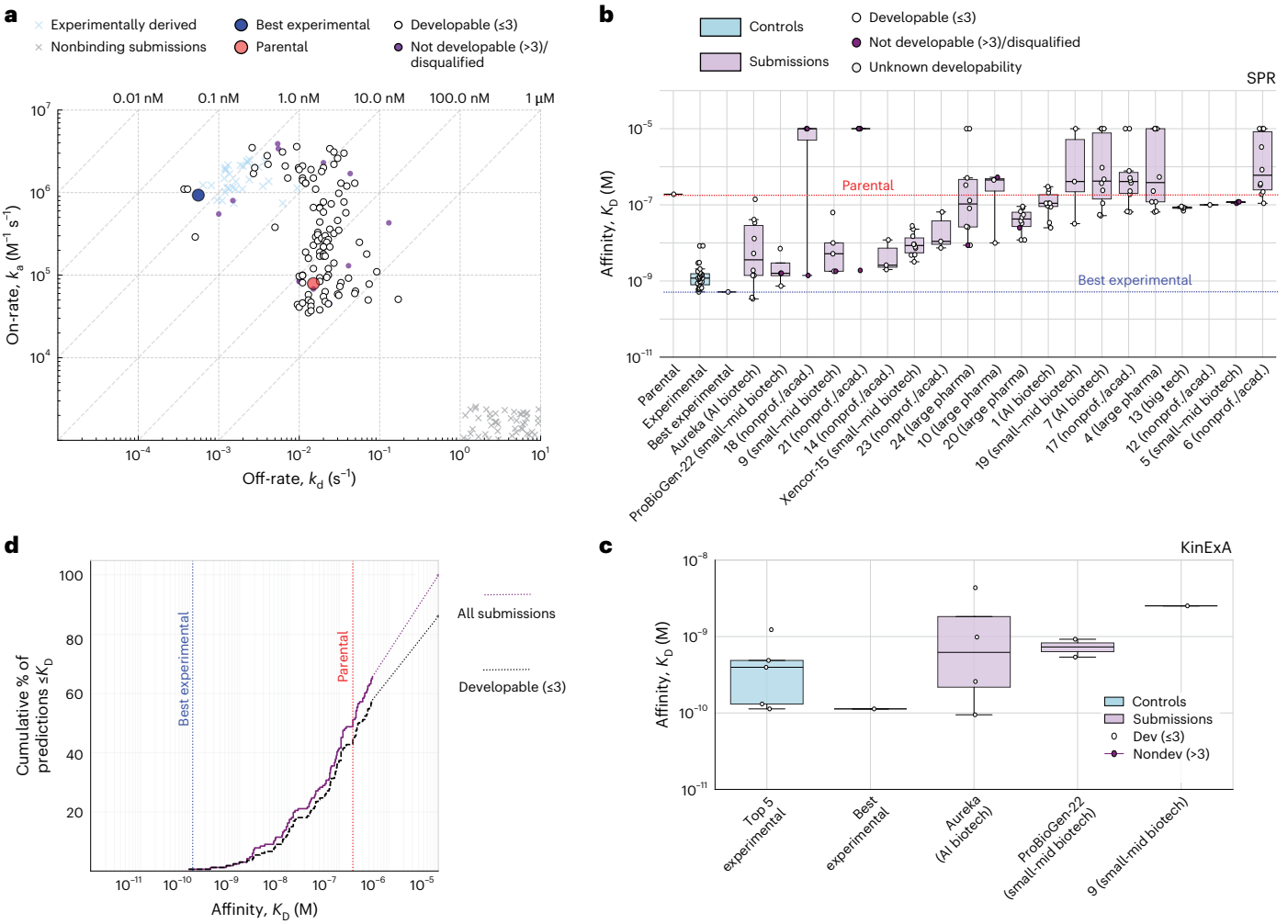
Across 25 organizations and 165 submissions, 118 (71.5%) bound the RBD and 47 (28.5%) were nonbinders, with 22 (13.3%) achieving single-digit nM affinity, five (3.0%) achieving subnanomolar affinity and one (0.6%) achieving sub-100 pM affinity; 86.1% passed developability (median = 1; 13.9% disqualified for dev ≥ 4) and 63.0% were developable binders. The winning entry from Aureka achieved a 95 pM affinity, a 2,000-fold improvement over the parental antibody. The methodological basis for the winning entry is detailed in Extended Data Fig. 1e and Supplementary Note 2.

Challenge 2: affinity ranking within three HCDR3 clusters

For the rank prediction challenge, the goal was to identify the highest-affinity developable antibody in each of three HCDR3 clusters (27F, 28F, 47F), defined by shared or similar HCDR3s, from an RBD binding output selected from a semisynthetic antibody library in which germline matched natural and mutant CDRs, purged of sequence liabilities, are embedded within developable therapeutic antibody scaffolds[18]. The three clusters comprised 2,524 (27F), 3,554 (28F) and 410 (47F) unique sequences, with diversity within CDRs and variable developability profiles. Affinities were provided for some antibodies in clusters 27F ($n = 22$), 28F ($n = 14$) and 47F ($n = 3$) (Fig. 1h and corresponding Supplementary Data 2), including the 'cluster control', the most abundant sequence, which, despite always being a binder, does not necessarily have the highest affinity[15,16].

Affinities were measured for predicted sequences using SPR (Supplementary Fig. 4a,b and Fig. 3a). Cluster 27F shows many submissions with slower dissociation rates and similar association rates relative to the cluster control (Fig. 3a, left), cluster 28F gains skew toward association rate improvement (Fig. 3a, middle) and cluster 47F shows minimal overall improvement except for a few submissions

https://doi.org/10.1038/s41587-026-03238-6



e

Best experimental: 113 pM (95% CI = 67.7–178 pM)
Best AI subm (Aureka): 95 pM (95% CI = 69.1–127 pM)

Affinity bin	*n* (all)	% all	*n* (developable)	% developable
<200 nM	84	50.90%	74	44.80%
<100 nM	68	41.20%	61	37.00%
<50 nM	48	29.10%	42	25.50%
<10 nM	27	16.40%	22	13.30%
<1 nM	5	3.00%	5	3.00%
<100 pM	1	0.60%	1	0.60%

Fig. 2 | Affinity maturation challenge and winning method overview. **a**, Isoaffinity plot (association rate versus dissociation rate) measured by SPR for all challenge 1 AI sequence submissions. Points are color-coded by developability as passing (≤3, white) or failing (>3, purple). Reference points (red) include the parental antibody, the best experimentally derived binder (blue) and the experimentally derived outputs from final combined library (light-blue 'x'). **b**, Box-and-jitter plots ranking organizations by their best achieved SPR K_D (light-purple fill), overlaid with developability status. The 'experimental baseline' (far left; light-blue fill) shows the distribution from clones derived by fluorescence-activated sorting. Each plotted point is one independent antibody construct (SPR K_D is the mean of three technical replicates per construct, range: 1–8). Boxes show the median (center line), 25th–75th percentiles (bounds) and 1.5× the IQR (whiskers); all individual points are overlaid as jitter. Controls (left to right): parental, $n = 1$; experimental, $n = 31$ (developability not measured, shown in gray); best experimental, $n = 1$. Here, 20 of the 25 participating organizations are ranked by best SPR K_D, with total AI submissions plotted ($n = 119$). **c**, Box-and-jitter plots ranking organizations (only top SPR candidates tested) by their best achieved KinExA K_D. The convention is identical to **b**, except that KinExA equilibrium fits replace the SPR K_D ($n = 8$ AI submissions tested by KinExA, $n = 5$ experimental). **d**, Cumulative distribution of best measured affinities (SPR or KinExA) for all designs (purple line) and the developable subset (black line). The vertical line reference marks the best affinity achieved by experimental in vitro display methods (blue) and the parental antibody (red). **e**, Tabular breakdown of submission performance by affinity bin (fold improvement over parental), showing the counts and percentages of developable candidates outlined in **b**.

(Fig. 3a, right). Although cluster controls had composite developability scores ≤ 3 (Supplementary Fig. 5), developability was strongly associated with the corresponding HCDR3 cluster, with clusters 27F and 28F in particular showing more poorly developable antibodies, while all antibodies in cluster 47F were developable (Fig. 3a and Supplementary Fig. 5). Developable SPR clones were then subjected to KinExA affinity measurement (Supplementary Fig. 6).

Rank-ordered box-and-jitter plots arrange the submissions of the top 20 organizations from strongest to weakest affinity (Fig. 3b) for SPR (top) and KinExA (bottom) for each cluster relative to cluster controls. These results highlight the difficulties of predicting strong developable binders by cluster group, with cluster 28F being particularly challenging (Fig. 3b, middle row), and confirm the importance of the HCDR3 cluster in developability (Fig. 3c).

We previously showed that the cluster control is always a binder, but not necessarily the best, with approximately 30% of randomly picked clones exhibiting improved affinities[15,16]. For clusters 27F, 28F and 47F, 10.3%, 13.8% and 9.8%, respectively, of all submissions were higher affinity relative to the cluster control (Fig. 3d). The best predicted antibody from cluster 27F was 9.2 pM, a 4.0-fold improvement over the cluster control (36.6 pM). Cluster 47F showed the greatest improvement at 7.6-fold, identifying a 50 pM antibody relative to 378 pM for the cluster control. Cluster 28F showed only minimal improvement with a 1.8-fold improvement (101 pM versus 177 pM). Per-cluster 'top ranking' tables summarize the strongest affinity predictions (Supplementary Fig. 6) while flagging those disqualified for developability or sequence violations (Extended Data Fig. 2a).

The LD of the top submissions from the cluster control using concatenated CDRs (Extended Data Fig. 2b) or HCDR3s (Extended Data Fig. 2c) shows that, for cluster 27F, the winning prediction (University of California, San Diego (UCSD) and Scripps) used the same HCDR3 and was only one substitution away for the concatenated CDRs. This HCDR3 cluster comprised two additional HCDR3s; however, all top submissions used the same cluster control HCDR3. This is in contrast to the cowinners of cluster 28F, for which Xencor predicted a distinct sequence (LD for concatenated CDRs = 12, LD for HCDR3 = 4), whereas WashU remained closer to the most abundant clone (LD for concatenated CDRs = 2, LD for HCDR3 = 0). Lastly, for cluster 47F, the sole winner (WashU) predicted a highly divergent sequence with a concatenated CDR LD = 26 and HCDR3 LD = 8 relative to the cluster control, exceeding the range of conservative predictions.

It is notable that all but one of the AI prediction strategies were worse than random picking at identifying high-affinity binders; only 9.8–13.8% of all AI submissions (and 11–50% of the winners) improved upon the cluster control, compared to 39% for randomly picked clones. For most participants, using their AI models to select antibodies was actively counterproductive relative to the baseline strategy of selecting the most abundant clone and picking additional random clones. Individually, WashU (D.H.F. Lab) was the only group to exceed the random baseline (50% versus 39%) (Extended Data Fig. 2d). Furthermore, AI prediction strategies did not extend across all clusters for any of the winning algorithms.

For cluster 27F, across 26 organizations and 58 submissions, 57 (98.3%) bound RBD and one (1.7%) was a nonbinder, with 12 (20.7%) at single-digit nM, 44 (75.9%) at sub-nM and 14 (24.1%) at sub-100 pM affinity; 79.3% passed developability (median = 2; 20.7% disqualified for dev ≥ 4) and 77.6% were both binders and developable. UCSD and Scripps tied for the top rank (Extended Data Fig. 3a,b), both predicting an identical 9.2 pM antibody, a 4.0-fold improvement over the cluster control, with 10.3% and 8.6% of submissions showing 1.5-fold and 2.0-fold improvement, respectively, with an acceptable developability score. The methodological basis for the cowinning predictions is detailed in Supplementary Note 3.

For cluster 28F, across 26 organizations and 58 submissions, 57 (98.3%) bound RBD and one (1.7%) was a nonbinder, with ten (17.2%) at single-digit nM, 44 (75.9%) at sub-nM and two (3.4%) at sub-100 pM affinity; 41.4% were disqualified for poor developability (dev ≥ 4, median = 3), including the lone binder, leaving 34 (58.6%) developable binders. Xencor (Extended Data Fig. 3c) and WashU (Extended Data Fig. 3d) tied for the top rank with indistinguishable 105–106 pM binders, a 1.7-fold improvement over the cluster control. The methodological basis for the cowinning predictions is detailed in Supplementary Note 4.

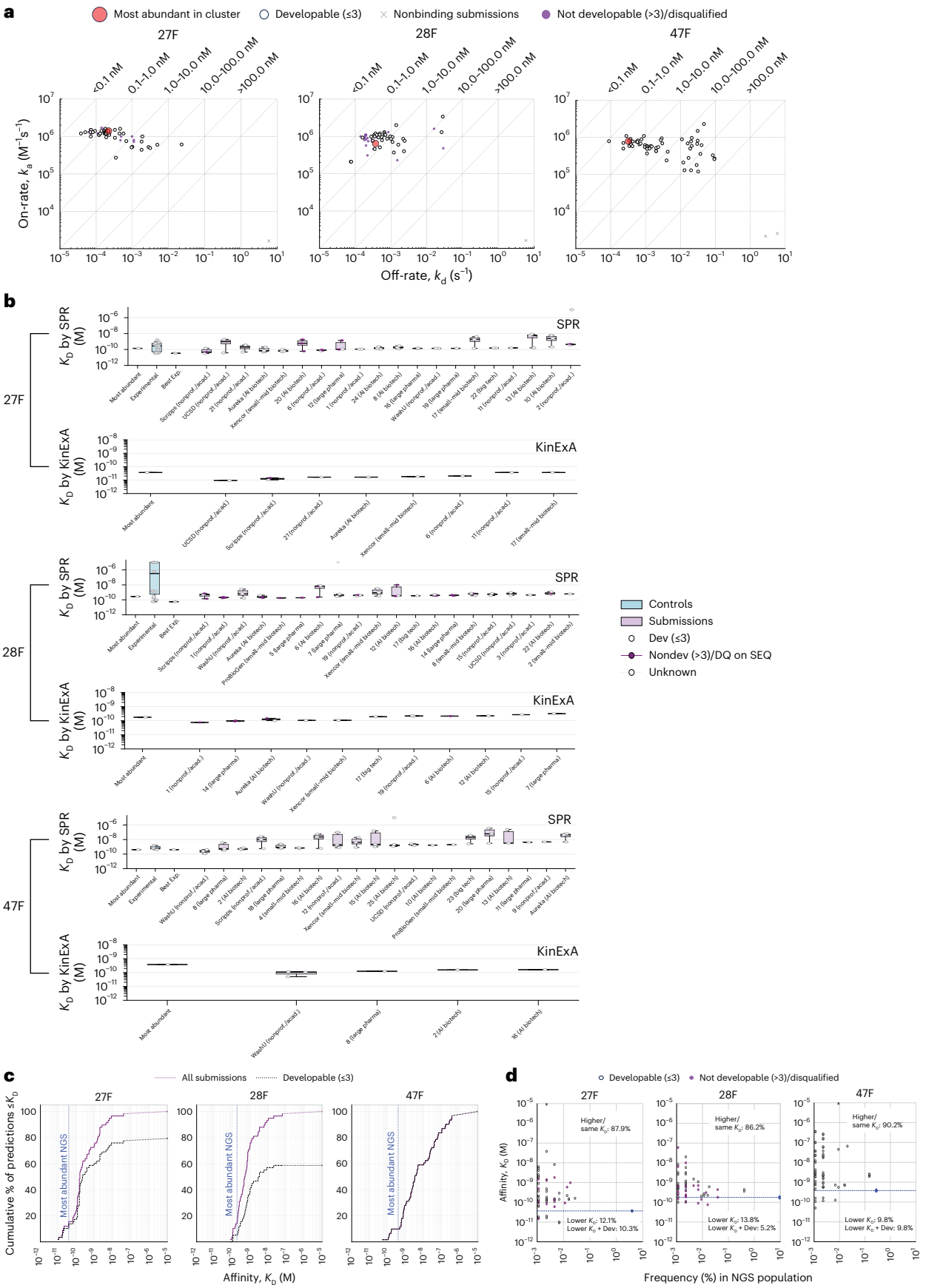
Lastly, for cluster 47F, across 26 organizations and 61 submissions, 59 (96.7%) bound RBD and two (3.3%) were nonbinders, with 19 (31.1%) at single-digit nM, 17 (27.9%) at sub-nM and one (1.6%) at sub-100 pM affinity; 100% passed developability (median = 1). The winning entry from WashU achieved a 50 pM affinity, which was a 7.6-fold improvement over the cluster control. The methodological basis for the winning entry is detailed in Extended Data Fig. 3d and additional details can be found in Supplementary Note 5.

Challenge 3: out-of-Library CDR optimization and/or design

In this task, teams used the full 32,059 unique sequences from datasets 2 and 3 (Fig. 1c), combined with the provided affinities of 142 antibodies (Fig. 1h and Supplementary Data 2), to generate high-affinity developable antibodies by designing novel CDRs or combinations thereof without modifying frameworks. Figure 4a shows the isoaffinity SPR plot of all submissions (sensorgrams in Supplementary Fig. 7a,b) compared to experimentally derived antibodies from prior studies[16,19] or predicted in challenge 2. Generally, there were many more nonbinders among AI-designed antibodies and, for binders, affinities tended to be worse than experimental antibodies. That said, there were 33.9% subnanomolar submissions, of which ~65% (22.0% of total) were developable (Supplementary Fig. 8). Across the entire dataset, 53.6% of all submissions were binders (≤1 μM) and developable, with the proportion of developable antibodies dropping notably within the highest-affinity pools. Figure 4b shows the mean SPR affinities for all submissions by organization, while Fig. 4c shows the KinExA affinities of a smaller subset of the strongest SPR affinity antibodies (2.2–9.0 times stronger than corresponding SPR affinities). Cumulative K_D distributions combining best affinities, whether KinExA or SPR (Fig. 4d), and the corresponding table (Fig. 4e) illustrate the challenge of simultaneously optimizing affinity and developability in unexplored sequence space.

Fig. 3 | Cluster prediction challenge overview. a, Isoaffinity plots for predicted sequences within three HCDR3 clusters: 27F (left), 28F (middle) and 47F (right). Each plot is centered on the cluster-specific experimental control (red color; most abundant sequencing clone). Points are colored by developability according to passing (≤3; white) or failing (>3; purple) composite developability scores. **b**, Rank-ordered box-and-jitter plots of top 20 organizations for each cluster (27F, 28F and 47F) by best SPR (top) and KinExA (bottom) affinity (light-purple fill) relative to cluster controls and experimentally derived antibodies corresponding to each cluster (light-blue box fill). Each plotted point is one independent antibody construct (SPR K_D is the mean of three technical replicates per construct, range: 1–8; KinExA K_D is the equilibrium fit). Boxes show the median (center), 25th–75th percentiles (bounds) and 1.5× the IQR (whiskers); all individual points are overlaid as jitter. Top, SPR panels include the most abundant clone (n = 1), experimental antibodies (n = 23, 14 and 3 for 27F, 28F and 47F, respectively, mostly without developability scoring, shown in gray), best experimental antibody (n = 1) and the top 20 of 26 participating organizations ranked by best SPR K_D (total AI submissions plotted, n = 58, 57 and 59). Bottom, KinExA panels show only the most abundant clone (n = 1) and the participating organizations with KinExA measurements (total AI submissions plotted, n = 9, 14 and 6). **c**, Cumulative affinity distributions for each cluster, comparing all submissions (purple) versus developable ones (black). **d**, Improvement rate analysis, showing the percentage of AI submissions that improved upon the simple heuristic of selecting the most abundant clone (blue dot representing abundance superimposed on blue dashed line representing affinity) in the sequencing cluster (10.3% for 27F, 5.2% for 28F and 9.8% for 47F).

https://doi.org/10.1038/s41587-026-03238-6



The KinExA affinity of the best experimental antibody was 9.2 pM; two submissions (2.4% of the total) had higher affinities (2.9 pM and 3.1 pM) and four others had comparable affinities (8.7–9.6 pM) but only one of the highest-affinity antibodies (1.2% of total) met the composite developability criteria (score ≤ 3) (Fig. 4f). Importantly, this antibody failed to elute from the HIC column—a failure that, under standard therapeutic development criteria, would probably be disqualifying. It also showed relatively low specific activity (Supplementary Fig. 9) with cooperative binding (positive Hill coefficient of 1.41). The composite scoring framework used in this challenge (≤3 denoting developable) allowed a single assay failure to be offset by favorable performance elsewhere, giving this antibody the top rank. This exposes a limitation in the present scoring design and the need for critical future go/no-go criteria ('Discussion').

The winning sequence exhibited a concatenated CDR distance of five substitution from the nearest sequence in the provided dataset (Fig. 4g), with the remaining top five winners 1–3 substitutions away. Those that were ≥6 substitutions away exhibited lower rankings. All top submissions used an HCDR3 sequence in the experimental data (Fig. 4h).

In summary, across 23 organizations and 168 submissions, 117 (69.6%) bound RBD and 51 (30.4%) were nonbinders, with 22 (13.1%) at single-digit nM, 57 (33.9%) at sub-nM and 24 (14.3%) at sub-100 pM affinities; 71.4% passed developability (median = 2; 28.6% disqualified for dev ≥ 4) and 53.6% were binders and developable. The methodological basis for the winning out-of-library design described above, which achieved the 2.9 pM affinity, is showcased in Extended Data Fig. 4a–d (additional details in Supplementary Note 6).

Computational method class analysis

We grouped each organization's approach per challenge into four method classes plus one undisclosed group (Fig. 5a): (1) statistical or heuristic baselines (ProBioGen consensus analysis); (2) classical ML, covering engineered feature pipelines and simple neural networks without attention or pretraining (UCSD Gaussian process regression in challenge 2; long short-term memory and tree-based pipelines); (3) methods that use attention or a pretrained sequence-based protein language model (PLM) such as ESM-2 or ft-ESM, AbLang, BALM, AntiBERTy, custom self-supervised antibody language models (WashU AlignNet) and proprietary antibody large language models (LLMs); and (4) structure-aware methods using explicit 3D structure (AF or AF3, ProteinMPNN, AntiFold, RFdiffusion and pairformer). Organizations that combined methods within a single challenge are counted in each applicable class. Most multiclass entries paired PLM or attention with structure (for example, Aureka and Scripps), with two organizations pairing classical ML with structure, neither of which placed among the challenge winners. The best organization rank reached across all challenges was ≥6th and 12th submissions overall in challenge 3. Organizations placed in class 3 but not class 4 used sequence-based PLMs without explicit 3D structure prediction at inference. The within-class variation in backbone, prediction head and selection logic is substantial; thus, Fig. 5 captures broad trends rather than controlled comparisons. By class (Fig. 5b), the winning class differed by challenge; Aureka (PLM and attention plus structure) won challenge 1, UCSD (classical ML) tied with Scripps in cluster 27F, WashU (custom transformer) tied with Xencor (PLM) in cluster 28F, WashU won cluster 47F and Xencor (PLM) won challenge 3. ProBioGen's consensus baseline matched the library-specific consensus across all six CDRs (Fig. 2i) and ranked third by KinExA (540 pM) in challenge 1 with no developability concerns. Cross-challenge ranks (Fig. 5c) show that each disclosed group had at least one challenge where its best submission fell well outside the leaders.

Discussion

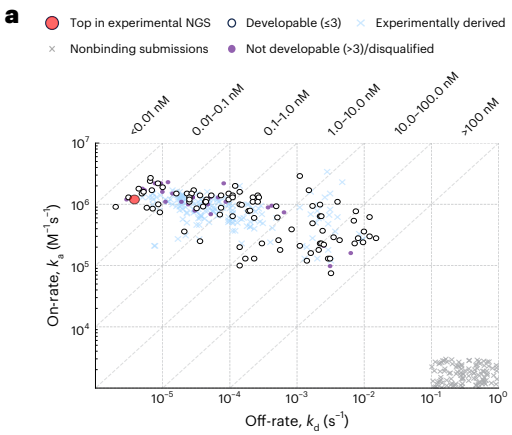
This prospective, blinded study provides a clear, experimentally anchored objective snapshot of computational antibody discovery as of Q1 of 2025. By judging every submission against a uniform suite of high-throughput SPR, orthogonal KinExA and a five-assay developability panel adapted from a previous study[9] (Fig. 1m–o), we removed retrospective variance to evaluate algorithmic performance directly, independently and without bias. Given the uncertainty of the value of AI in antibody discovery, this AIntibody challenge was deliberately primed for success by selecting a well-studied protein target (SARS-CoV-2 RBD) and providing rich deep sequencing datasets comprising extensive affinity information[8,15,16,19]. The three challenges, affinity maturation (Figs. 1a and 2), cluster-constrained hit expansion and ranking (Figs. 1b and 3) and library-inspired CDR design (Figs. 1c and 4), are based on steps in the antibody discovery pipeline that could be substantially improved by reliable success. However, of the three challenges, only the first used an experimental sequencing pipeline dataset like that generated under normal conditions, whereas the other two provided extensive affinity data not normally available until later in the discovery pipeline.

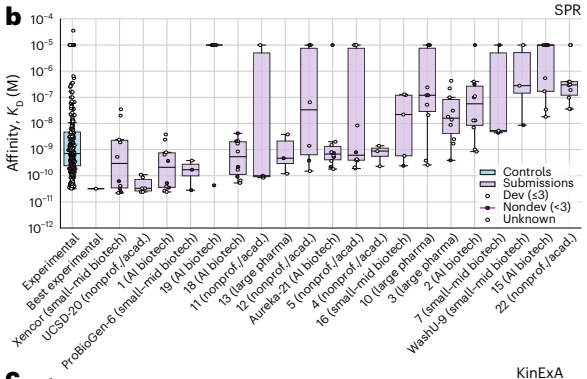
The clearest case for AI use lies in timeline reductions for lead optimization in challenge 1, where ~13% of submissions achieved a ≥20-fold affinity improvement (<10 nM) over the parental antibody with several subnanomolar antibodies (Fig. 2a–d). However, these were unevenly distributed with great variation between participants. Aureka designed six developable antibodies with affinities < 10 nM, including one (94.7 pM) statistically tied with the best experimental antibody (113 pM). Although six others designed at least one developable antibody with affinity < 10 nM, eight participants were unable to design a single developable binder. If Aureka's model was widely applicable (and available), it would reduce experimental time by 2–3 weeks, replacing the experimental combination library with a single in silico step, aligning with findings that deep learning can drive optimization given sufficient mutational diversity[12]. Almost as good (538 pM) was a consensus sequence approach that used statistics and no AI. Although hit rates of the top few models are generally lower than the near-100% success of fluorescence-activated sorting of an experimental combination library (Fig. 2a), synthesizing and testing 10–20 designs would be sufficient to identify useful high-affinity variants.

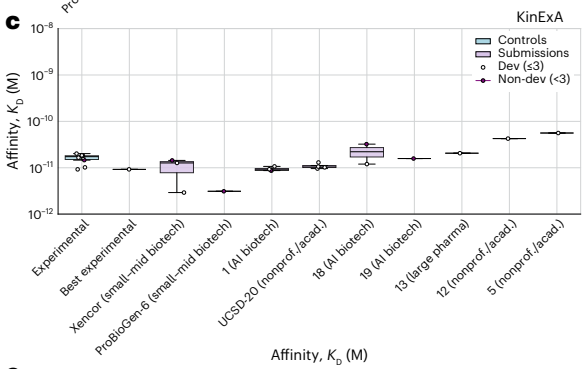
Distinct from affinity maturation, challenge 2 tested hit expansion within heterogeneous HCDR3-clustered pools. Unlike challenge 1,

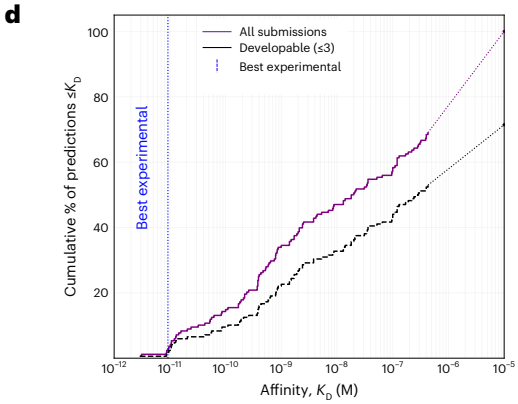
Fig. 4 | Out-of-library CDR design challenge overview. a, Isoaffinity plot for out-of-library designs relative to experimental benchmarks used within the provided dataset (light-blue 'x') including best experimentally derived affinity (red). Points of submissions are colored by developability score as passing (≤3; white) or failing (>3; purple). **b**,**c**, Rank-ordered box-and-jitter plots of the top 20 of 23 participating organizations by best SPR (**b**) and KinExA (**c**) affinities, with both panels showing the distribution of affinity per organization (light-purple fill) relative to the experimental baseline (light-blue fill). There were $n = 347$ independent antibody constructs combining 164 antibodies from prior in-house experimental work (developability not measured, shown in gray) with 183 challenge 2 submissions experimentally characterized in this study (149 developable (score ≤ 3) and 36 not developable (score > 3); one submission had no scoring). Each plotted point is one independent antibody construct (SPR K_D is the mean of three technical replicates per construct, range: 1–8; KinExA K_D is the equilibrium fit). Boxes show the median (center line), 25th–75th percentiles (bounds) and 1.5× the IQR (whiskers); all individual points are overlaid as jitter. AI submissions are plotted ($n = 117$ by SPR (**b**) and $n = 18$ by KinExA (**c**)). **d**, Cumulative frequency of submissions achieving corresponding affinity (SPR or KinExA) for all submissions (purple) versus only developable ones (black) relative to best experimentally derived affinities (blue vertical line). **e**, Tabulation of values from the plot in **d** showing the best affinities (SPR or KinExA) obtained, along with the number of submissions (and those developable) that fall into each affinity bin. **f**, Final ranking of the top ten designs by KinExA K_D and then SPR K_D. **g**,**h**, Minimum LD, using concatenated CDRs (LCDR1–LCDR3 and HCDR1–HCDR3; **g**) and HCDR3 (**h**) of the top-ranked antibody submission in challenge 3 to the closest experimental sequence provided in the dataset during release of the data.

https://doi.org/10.1038/s41587-026-03238-6









e

Affinity range	No. of submissions	No. of developable (≤3)	% Submissions	% Developable
<10 pM	6	4	3.60%	2.40%
<100 pM	24	16	14.30%	9.50%
<1 nM	57	37	33.90%	22.00%
<10 nM	79	55	47.00%	32.70%
<100 nM	96	72	57.10%	42.90%
<1 µM	117	90	69.60%	53.60%
≥1 µM/nonbinders	51	30	30.40%	17.90%

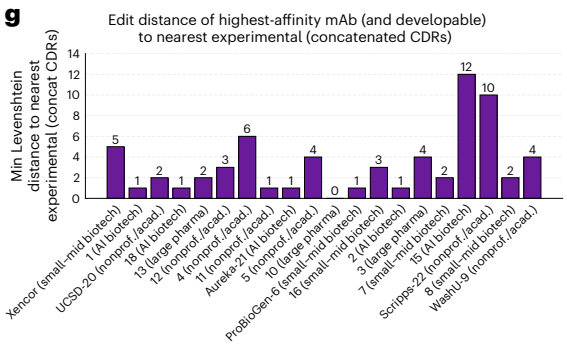
f

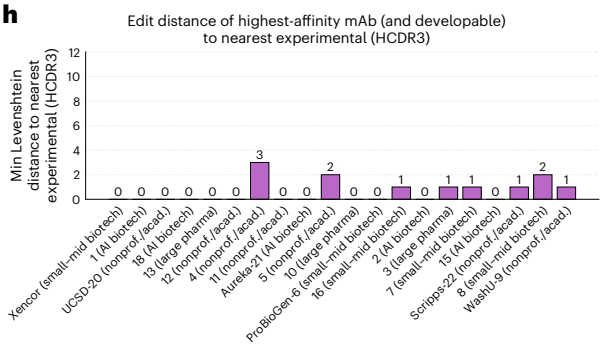
PLM = Protein language model/attention-based
SA = Structure-aware
CML = Classical ML
Stat = Statistical/heuristic

Bold/light purple = winner
DQ = disqualification reason
DEV = developability score > 3

Best experimental: 9.2 pM (95% c.i. = 7.5–11 pM)

Rank	Org.	Type	Method	DQ	LCDR1	LCDR2	LCDR3	HCDR1	HCDR2	HCDR3	K_D KinExA (M) (95% CI)	K_D SPR (M) (± s.d.)
1	Xencor	Small/mid bio	PLM		QDIQHY	ATISRPS	HQYATYPWT	GYSFTAHS	INPYIGDT	ARDGYGDYRGLYGMDV	2.92×10^{-12} (2.13×10^{-12}–3.87×10^{-12})	2.20×10^{-11} ($\pm 8.0 \times 10^{-12}$)
-	ProBioGen-6	Small/mid bio	Stat	DEV	QNIHSY	EASTLKP	HQYAAYPWT	GYTSPTT	ISPYRGDT	ARDGYSSGYYGMDV	3.11×10^{-12} (1.71×10^{-12}–5.03×10^{-12})	2.80×10^{-11} ($\pm 7.9 \times 10^{-12}$)
2	1	AI bio	PLM		QSIGSF	EASTLKP	HQYAAYPWT	GCTSPTT	ISPYRGDT	ARDGYSSGYYGMDV	8.69×10^{-12} (6.77×10^{-12}–1.1×10^{-11})	2.70×10^{-11} ($\pm 1.0 \times 10^{-11}$)
-	1	AI bio	PLM	DEV	QSIGSF	EASTLKP	HQYATYPWT	GYTSPTT	ISPYKGDT	ARDGYSSGYYGMDV	8.78×10^{-12} (5.76×10^{-12}–1.26×10^{-11})	3.50×10^{-11} ($\pm 3.2 \times 10^{-12}$)
3	1	AI bio	PLM		QDIGSF	EASTLKP	HQYAAYPWT	GYTSPTT	ISPYKGDT	ARDGYSSGYYGMDV	9.30×10^{-12} (5.9×10^{-12}–1.37×10^{-11})	3.90×10^{-11} ($\pm 3.0 \times 10^{-12}$)
4	UCSD-20	Academic	CML		QDIGNH	ASSNLAS	HQYATYPWT	GYTFTDHS	INPYVGDT	ARDGYGDYRGLYGMDV	9.60×10^{-12} (6.93×10^{-12}–1.29×10^{-11})	2.40×10^{-11} ($\pm 4.4 \times 10^{-12}$)
5	UCSD-20	Academic	CML		QDIGNH	EANKRPS	HQYATYPWT	GYSFTDHS	INPYVGDT	ARDGYGDYRGLYGMDV	1.02×10^{-11} (7.48×10^{-12}–1.34×10^{-11})	3.30×10^{-11} ($\pm 1.1 \times 10^{-11}$)
6	UCSD-20	Academic	CML		QDIGNH	ASSNLAS	HQYATYPWT	GYSFTDHS	INPYVGDT	ARDGYGDYRGLYGMDV	1.06×10^{-11} (8.35×10^{-12}–1.34×10^{-11})	2.40×10^{-11} ($\pm 4.1 \times 10^{-12}$)
7	1	AI bio	PLM		QAIGSF	EASTLKP	HQYATYPWT	GYTSPTT	ISPYRGDT	ARDGYSSGYYGMDV	1.07×10^{-11} (8.11×10^{-12}–1.38×10^{-11})	2.40×10^{-11} ($\pm 2.8 \times 10^{-12}$)
8	18	AI bio	PLM		QSIGTY	SGSTLQS	HQYATYPWT	GYSFTDHS	INPYIGDT	ARDGYSYGYLSGMDV	1.20×10^{-11} (8.82×10^{-12}–1.6×10^{-11})	6.60×10^{-11} ($\pm 2.1 \times 10^{-12}$)





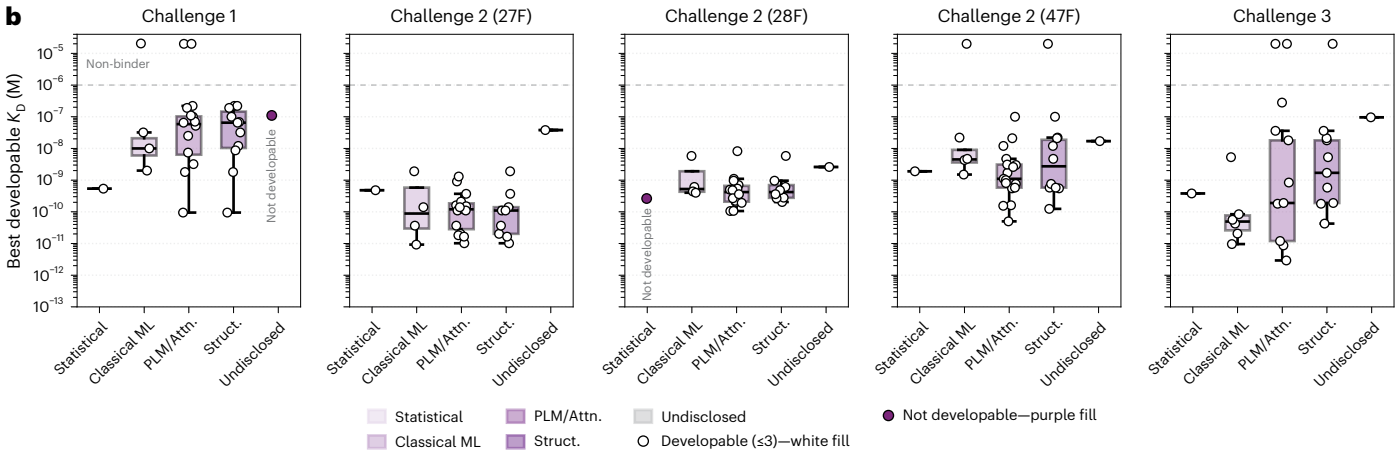
https://doi.org/10.1038/s41587-026-03238-6

a

Method Class	*n**	Classification basis	Example
Statistical/Heuristic	1	Consensus analysis, positional enrichment from NGS abundance; no learned models	ProBioGen (consensus)
Classical ML	6*	GP regression, AutoML with engineered features; LSTMs, CNNs, MLPs (no attention)	UCSD Ch2 (GP regression);
PLM-/attention-based	18*	Transformer/attention models—pretrained PLMs (ESM-2/ft-ESM, AbLang, BALM, AntiBERTy) or custom self-supervised antibody language models	Xencor (ft-ESM); WashU (AlignNet); UCSD Ch1/Ch3 (antibody LM)
Structure-aware	16*	Uses explicit 3D structure (AlphaFold/AF3, ProteinMPNN, AntiFold, RFdiffusion, pairformer); often combined with PLM-/attention-based or classical ML	Aureka (pairformer + DPO); Scripps (AF/AntiFold)
Undisclosed	1	Method not disclosed by submitter	—

n = number of organizations. *Submissions by organizations may be defined by multiple method classes

b



c

Rank (post-disqualification) / rank (pre-disqualification) submission rank

Organization	Method(s)	Ch1	Ch2-27F	Ch2-28F	Ch2-47F	Ch3
Xencor	PLM/Attn.	10/14	4/6	1/6	16/16	1/1
UCSD	PLM/Attn., classical ML	106/119 (nonbinder)	1/1	10/25	20/20	4/6
Washington U. St. Louis	PLM/Attn.	64/70	15/23	1/5	1/1	84/109
Scripps	PLM/Attn., structure-aware	47/53	1/1	12/29	8/8	67/91
Aureka	PLM/Attn., structure-aware	1/1	DQ/5	DQ/3	34/34	19/29
ProBioGen	Statistical	3/3	31/40	DQ/9	22/22	24/38

Rank shown is top submission rank (before DQ/after DQ)—that is, the rank of the org's best individual submission, not the organization rank. Purple = challenge winner or cowinner; red = disqualified.

Fig. 5 | Computational method classification and cross-challenge performance. a, Four method classes plus an undisclosed group, with defining characteristics and unique organization counts (29 organizations total; an organization that used different methods across challenges or combined methods within a single challenge was counted in each applicable class). **b**, Best developable affinity (K_D) per organization by method class for challenge 1, challenge 2 clusters 27F, 28F and 47F and challenge 3, shown as a composite of five subpanels with a shared *y* axis. Each plotted point is one organization's single best developable affinity (one biological replicate per organization × method class cell); organizations using multiclass methods (for example, PLM + structure-aware method) appear in every applicable single-class column. Boxes show the developable binder distribution, showing the center line (median), bounds (25th–75th percentile) and whiskers (1.5× the IQR). The numbers of points per box are as follows (challenge 1/27F/28F/47F/challenge 3): statistical, 1/1/1/1/1; classical ML, 4/4/4/5/6; PLM-based or attention-based, 14/16/15/17/11; structure-aware methods, 12/10/9/11/11; undisclosed methods, 1/1/1/1/1. White-filled black-outlined circles indicate developable submissions (score ≤ 3). Developable organizations whose best submission was a nonbinder are plotted at the assay ceiling above the dashed 10^{-6} M reference line. In two cells where every participating organization lacked any developable binder (ProBioGen in the statistical class of 28F and the undisclosed group in challenge 1), the best nondevelopable (score > 3) affinity is shown as a purple-filled circle labeled 'not developable'. Box statistics exclude these ceiling and purple points. **c**, Top submission rank (after disqualification/before disqualification) for each disclosed group across the five challenges. Purple cells indicate the winner or cowinner; red cells indicate disqualification. UCSD[†]: methods that used attention or a pretrained PLM in challenges 1 and 3 and classical ML (Gaussian process regression) in challenge 2 (the winning method).

a notable number of antibody affinities were provided. The highest-abundance clones are typically binders but rarely possess the highest affinities[16]. While AI models successfully identified binders (<1 µM) at high rates (>96%, reflecting the functionality of the library), improving upon our standard approach of selecting the most abundant clone proved challenging, with only 9.8–13.8% of all submissions showing improvements (Fig. 3h), worse than the 39% of randomly picked clones from these clusters with higher affinities than the cluster control (Extended Data Fig. 2d). With the exception of WashU (50% versus 39%), even the winning algorithms were worse than random picking (11–25% versus 39%) (Extended Data Fig. 2d) and did not transfer across all clusters, further limiting their utility. Performance was further complicated by HCDR3-driven biophysical liabilities. Despite using developability-optimized libraries[18], the failures in 28F versus the success in 47F confirm the critical role of HCDR3 in developability and the challenges of computational prediction (Fig. 3c). In summary, with the exception of one participant (WashU), AI algorithms were worse than just picking random clones when attempting to identify improved antibodies within a cluster and even that algorithm was ineffective for cluster 27F.

Challenge 3 results were highly diverse. Although Xencor's 2.9-pM antibody won the competition according to published rules[8], its fatal flaws (discussed above) would have probably doomed it as a clinical candidate. The next best developable antibody (8.69 pM) was statistically tied with the best experimental antibody (9.2 pM). A total of 16 (9.5%) antibodies had affinities < 100 pM and 37 (22%) antibodies had affinities < 1 nM (Fig. 4e). These results contrast with the large number of nonbinders (30.4%) and nondevelopable binders (16%), together comprising 46.4% of all submissions. For most organizations, the affinity spreads of their submissions were extremely broad, comprising nonbinders and binders of variable affinities (Fig. 4b). Six organizations bucked this trend with sequences within a 100-fold affinity range, three of which were within tenfold with at least one subnanomolar design.

This uneven performance across and within tasks (Extended Data Fig. 2d and Fig. 5c) and distinct submission distributions (Figs. 2b–4b) highlight both strategic tension (peak performance versus consistency) and the great diversity of algorithms used. Models were highly tuned to the specific data regimes provided (for example, local epistasis versus heterogeneous ranking), creating a trade-off in which difficult targets may justify the 'one-off' rare winner (even if representing stochastic success), whereas industrial pipelines require consistent manufacturable leads. These results argue against 'one model for everything' and point toward a portfolio approach where algorithms are applied opportunistically to specific tasks and available data. Some of the winning groups used the same algorithm for more than one challenge, with generally inconsistent results (Extended Data Fig. 2d and Fig. 5c).

When evaluated by method class (Fig. 5), the most effective approach varied with the task. Challenge 1 (in library affinity optimization) was won by Aureka's combined PLM and structure pipeline (pairformer), with ProBioGen's consensus baseline ranking third by KinExA, ahead of most AI-generated designs. Challenge 2 (cluster-level ranking) had no single-class winner across all three clusters. Winners drew from classical ML, custom transformer methods and combined PLM and structure-aware methods, suggesting that local epitope or cluster idiosyncrasies, rather than method class, determined which model best ranked candidate sequences. Challenge 3 (out-of-library design) was won by a PLM pipeline (Xencor), consistent with pretrained language models contributing most where the design space extends beyond the training library and broader sequence priors are useful. Thus, PLM and attention-based methods, including custom self-supervised antibody language models and proprietary antibody LLMs, were most effective in select contexts such as novel sequence generation and in combination with structure for in library design. These results support matching method choice to task: consensus or statistical methods for in library affinity optimization, PLM-based methods for novel sequence design and ensemble or task-tuned approaches for cluster-level ranking.

That a simple statistical summary of selection data, without ML, outperformed most AI-generated designs in challenge 1 establishes an important baseline for the field. While consensus engineering has been used to improve antibody[20–22] and protein[23,24] stability and create new functional fluorescent proteins from scratch[25], there is only one report describing its use in affinity maturation[26], likely reflecting the need for extensive datasets from which consensuses can be derived. As datasets increase in size to accommodate AI training, it will be interesting to compare the properties of consensus antibodies to those generated by AI.

Our edit distance analysis highlights a dichotomy in generative strategies dependent on design constraints. In challenge 1, where design was restricted to LCDR1–LCDR3, HCDR1 and HCDR2 (with HCDR3 fixed), winning models demonstrated aggressive exploration, generating sequences distant (>12 substitutions) from the experimental pool (Extended Data Fig. 1c). In sharp contrast, despite the absence of design constraints in challenge 3, winning solutions converged on HCDR3 sequences in the experimental dataset (LD = 0), limiting innovation to conservative refinements (1–5 substitutions) in surrounding CDRs (Fig. 4g,h). Distinct from these design tasks, challenge 2 functioned as a prediction benchmark. In cluster 27F, all top-affinity clones shared the dominant HCDR3; hence, identifying the high-affinity HCDR3 was trivially achievable. The winning UCSD and Scripps prediction differed from the most abundant clone by only a single residue in HCDR2. In cluster 28F, WashU and Xencor tied for first using contrasting strategies; WashU's prediction differed from the most abundant by only two residues (one CDR; same dominant HCDR3 motif), while Xencor's deviated by 12 residues across CDRs, picking an HCDR3 variant 10^5 rarer than the dominant motif. WashU's winning prediction for cluster 47F deviated by 26 residues across all six CDRs from the most abundant clone and used the less abundant of two HCDR3 motifs in the cluster. While these select cases of deviation from the most abundant validate the models' capacity to predict fitness peaks distinct from the population center, this reliance on existing motifs raises questions about transferability to data-sparse scenarios. In particular, the provision of extensive affinity data in challenges 2 and 3 is not representative of a typical early-stage discovery campaign where AI would be most useful, as affinities are usually assessed at the end of campaigns. It will be important to reassess the successful algorithms on sequencing data alone to see whether they can perform equally well without affinity data, as they did in challenge 1.

Distinct from this strategic consideration, the challenge 3 outcomes exposed a critical shortcoming in our benchmark design. The top-ranked design in the challenge failed to elute from the HIC retention column yet secured the win because the challenge's composite scoring system allowed for high affinity to offset a single biophysical failure. To align computational incentives with therapeutic reality, future benchmarks will enforce critical go/no-go criteria, where severe red flags would trigger disqualification.

The absence of several prominent groups that have publicly reported strong computational antibody design capabilities is noted. While there may be legitimate reasons for nonparticipation—including concerns about blinding procedures, method disclosure requirements or timeline constraints—broader participation would strengthen the benchmark's representativeness. Benchmarks of this kind provide the most direct way for organizations and their investors to understand the true performance of their algorithms relative to alternative approaches. A valid reason for not participating was the result blinding, which relied on the integrity of the organizers and was not secured informatically. To encourage as many participants as possible, this will be addressed in the next AIntibody challenge. The requirement to disclose some method details also likely drove out some potential competitors, as many organizations

have legitimate trade secrets and intellectual property constraints around full disclosure.

The data and results of this AIntibody challenge are available to the community to allow others to test their models and compare them to participant algorithms. It is striking that only one of the winners (Aureka) describes itself as an AI company. The others were academic groups (Scripps, UCSD and WashU) and a medium-sized biotech (Xencor), which, although it has extensive antibody knowledge, does not claim AI expertise and won one of the cluster challenges and the out-of-library CDR design challenge using two different algorithms. This is promising, suggesting that the field remains open and deep pockets are not required to excel.

The value of using AI to expand the range of antibodies against a target of interest depends upon the properties of those additional antibodies. If AI-designed antibodies can improve on the affinity of antibodies derived experimentally (while presumably binding to the same epitopes), as is the case for at least some of the antibodies described here, AI is already useful in the antibody discovery process. Equally, if AI can design developable versions of nondevelopable antibodies or design antibodies binding epitopes not represented within a panel of experimental antibodies, this will increase the use of AI in the antibody discovery pipeline. The results described here indicate that, when presented with deep, targeted, high-quality data, the best AI algorithms (but not the others) can already provide real value in affinity maturation and for some HCDR3s (WashU) in hit expansion, by identifying likely high-affinity binders within identified HCDR3 clusters and designing out-of-library antibodies directed to the target of the data. Whether algorithms will always need such deep datasets, including affinity, for each target of interest or will be able to use such datasets to train models transferable to data-sparse situations, thereby transitioning from optimization to true discovery, remains to be seen and will be the subject of the next AIntibody challenge for 2026. This will involve competitions to generate high-affinity antibodies against a target to be revealed using any discovery method, including in vivo, in vitro or in silico and a second competition similar to the one here, with sequencing datasets but without affinity data.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-026-03238-6.

References

1. Moult, J. Predicting protein three-dimensional structure. *Curr. Opin. Biotechnol.* **10**, 583–588 (1999).
2. Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K. & Moult, J. Critical assessment of methods of protein structure prediction (CASP)—round XIII. *Proteins* **87**, 1011–1020 (2019).
3. Senior, A. W. et al. Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706–710 (2020).
4. Gathiaka, S. et al. D3R grand challenge 2015: evaluation of protein–ligand pose and affinity predictions. *J. Comput. Aided Mol. Des.* **30**, 651–668 (2016).
5. Guthrie, J. P. A blind challenge for computational solvation free energies: introduction and overview. *J. Phys. Chem. B* **113**, 4501–4507 (2009).
6. Cotet, T.-S. et al. Crowdsourced protein design: lessons from the Adaptyv EGFR binder competition. Preprint at *bioRxiv* https://doi.org/10.1101/2025.04.17.648362 (2025).
7. Ackloo, S. et al. CACHE (critical assessment of computational hit-finding experiments): a public–private partnership benchmarking initiative to enable the development of computational methods for hit-finding. *Nat. Rev. Chem.* **6**, 287–295 (2022).
8. Erasmus, M. F. et al. AIntibody: an experimentally validated in silico antibody discovery design challenge. *Nat. Biotechnol.* **42**, 1637–1642 (2024).
9. Jain, T. et al. Biophysical properties of the clinical-stage antibody landscape. *Proc. Natl Acad. Sci. USA* **114**, 944–949 (2017).
10. Chennamsetty, N., Voynov, V., Kayser, V., Helk, B. & Trout, B. L. Design of therapeutic proteins with enhanced stability. *Proc. Natl Acad. Sci. USA* **106**, 11937–11942 (2009).
11. 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).
12. Hummer, A. M., Schneider, C., Chinery, L. & Deane, C. M. Investigating the volume and diversity of data needed for generalizable antibody–antigen $\Delta\Delta G$ prediction. *Nat Comput Sci* **5**, 635–647 (2025).
13. Kovaltsuk, A. et al. Observed antibody space: a resource for data mining next-generation sequencing of antibody repertoires. *J. Immunol.* **201**, 2502–2509 (2018).
14. 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).
15. Erasmus, M. F. et al. Determining the affinities of high-affinity antibodies using KinExA and surface plasmon resonance. *MAbs* **15**, 2291209 (2023).
16. Erasmus, M. F. et al. Insights into next generation sequencing guided antibody selection strategies. *Sci. Rep.* **13**, 18370 (2023).
17. Teixeira, A. A. R. et al. Simultaneous affinity maturation and developability enhancement using natural liability-free CDRs. *MAbs* **14**, 2115200 (2022).
18. Teixeira, A.A.R. et al. Drug-like antibodies with high affinity, diversity and developability directly from next-generation antibody libraries. *MAbs* **13**, 1980942 (2021).
19. Ferrara, F. et al. Author Correction: A pandemic-enabled comparison of discovery platforms demonstrates a naive antibody library can match the best immune-sourced antibodies. *Nat. Commun.* **13**, 2097 (2022).
20. Ohage, E. C., Wirtz, P., Barnikow, J. & Steipe, B. Intrabody construction and expression. II. A synthetic catalytic Fv fragment. *J. Mol. Biol.* **291**, 1129–1134 (1999).
21. Wirtz, P. & Steipe, B. Intrabody construction and expression III: engineering hyperstable V_H domains. *Protein Sci.* **8**, 2245–2250 (1999).
22. Ohage, E. & Steipe, B. Intrabody construction and expression. I. The critical role of V_L domain stability. *J. Mol. Biol.* **291**, 1119–1128 (1999).
23. Albalat, R., Atrian, S. & Gonzalez-Duarte, R. *Drosophila lebanonensis* ADH: analysis of recombinant wild-type enzyme and site-directed mutants. The effect of restoring the consensus sequence in two positions. *FEBS Lett.* **341**, 171–176 (1994).
24. Steipe, B. Consensus-based engineering of protein stability: from intrabodies to thermostable enzymes. *Methods Enzymol.* **388**, 176–186 (2004).
25. Dai, M. et al. The creation of a novel fluorescent protein by guided consensus engineering. *Protein Eng. Des. Sel.* **20**, 69–79 (2007).
26. Mankowska, S. A. et al. A shorter route to antibody binders via quantitative in vitro bead-display screening and consensus analysis. *Sci. Rep.* **6**, 36391 (2016).

M. Frank Erasmus[1] ✉, Daniel Bedinger[2], Elizabeth Hopkins[3], Ginger Ferguson[3], Justine Strickler[3], Christilyn P. Graff[4], Samantha R. Summers[4], Stacy L. Capehart[4], Joshua D. Slocum[4], Crystal Richardson[5], Sumit Kumar[5], Zhifei Sun[5], Yujie Shang[5], Jixian Zhang[6], Ming Gu[6], Lixia Yi[6], Alon Wellner[6], Shuangjia Zheng[6], Wei Lu[6], Pietro Sormanni[7], Matthew Greenig[7], Haiping Zhang[8], Brendan T. Mann[9], Mahdi Baghbanzadeh[9], Ali Rahnavard[9], Gregory L. Moore[10], Huaiyu Sun[10], Ying Ding[10], Alex Nisthal[10], Jitendra Kanodia[10], Matthew J. Bernett[10], Aurélien Pélissier[11,12], Yanjun Shao[11], Maria Rodriguez Martinez[11], Karthik Ramesh[13], Horacio Nastri[13], Andreas Evers[14], Anhar Abdelatif[15], Andrew J. Bordner[15], Mykola Bordyuh[15], Lim Heo[15], Brian A. Kidd[15], H. Serhat Tetikol[15], Shuai Wei[15], Jung-Eun Shin[16], Ryan Peckner[16], Leigh Manley[16], Ajitesh Lunge[17], Yashas Devasurmutt[17], Bora Guloglu[18], Liviu Copoiu[18], Miles McGibbon[18], Monica L. Fernandez-Quintero[19], Nitesh Mishra[19], Sean M. Callaghan[19], Olivia M. Swanson[19], Daniel L. V. Bader[19], James A. Ferguson[19], Sai S. R. Raghavan[19], Benjamin Nemoz[19], Colleen A. Maillie[19], Charles Bowman[19], Bryan Briney[19], Andrew B. Ward[19], Paolo Marcatili[20], Rahmad Akbar[20], Bing He[21], Fandi Wu[21], Jianhua Yao[21], Bin Hu[22], Michal Kucer[22], Kaetlyn Rose Gibson[22], Rahul Somasundaram[22], Li-Wei Hung[22], Tomasz Kaszuba[23,24,25], Daved H. Fremont[23,24,25,26,27], Hyeongsun Jeong[28], Vinodh Babu Kurella[29], Shipra Malhotra[29], Satyendra Kumar[29], Yanyun Liu[30], Lingling Xu[30], Joshua Misa[30], Alexander Nicholas St. John[31], Jeff Vogt[32], Fátima A. Dávila-Hernández[32], Da Xu[32], Michael Chungyoun[32], Zyaja D. Huggan[32], Jeffrey J. Gray[32], Jonathan Parkinson[33], Young Su Ko[33], Wei Wang[33], Franziska Geiger[34], Jonathon D. Ziegler[34], Nikhil Haas[35], Chance Challacombe[35], Ahmad Qamar[35], Akshita Singh[36], Yi-Ching Tang[36], Zhiqiang An[36], Xiaoqian Jiang[36], Yejin Kim[36], Xinyan Zhao[36], Erik Swanson[37], Jürgen Klattig[38], Karsten Winkler[38], Tschimegma Bataa[38], Volker Sandig[38], Lilian Denzler[12,39], Chunan Liu[39], Randall J. Brezski[40], Laura Spector[1], Katheryn Perea-Schmittle[1], Sara D'Angelo[1], Fortunato Ferrara[1] & Andrew R. M. Bradbury[1] ✉

[1]Specifica, an IQVIA Business, Santa Fe, NM, USA. [2]Carterra, Salt Lake City, UT, USA. [3]Sapidyne Instruments, Inc. GmbH, Boise, ID, USA. [4]Mosaic Biosciences, Boulder, CO, USA. [5]GENEWIZ, Azenta Life Sciences, South Plainfield, NJ, USA. [6]Aureka Biotechnologies, Laguna Hills, CA, USA. [7]University of Cambridge, Cambridge, UK. [8]Shenzhen University of Advanced Technology (SUAT), Shenzhen, China. [9]The George Washington University, Washington, D.C., USA. [10]Xencor, Inc., Pasadena, CA, USA. [11]Biomedical Informatics and Data Science, Yale School of Medicine, New Haven, CT, USA. [12]Institute of Computational Life Sciences, Zurich University of Applied Sciences (ZHAW), Zurich, Switzerland. [13]Incyte, Inc., Wilmington, DE, USA. [14]Merck KGaA, Darmstadt, Germany. [15]Bristol-Myers Squibb, Princeton, NJ, USA. [16]Seismic Therapeutic, Watertown, MA, USA. [17]Locksmith Bio, Dover, DE, USA. [18]InstaDeep, London, UK. [19]Scripps Research Institute, San Diego, CA, USA. [20]Novo Nordisk A/S, Bagsværd, Denmark. [21]Tencent AI for Life Sciences Lab, Shenzhen, China. [22]Los Alamos National Laboratories, Los Alamos, NM, USA. [23]Washington University in St. Louis, St. Louis, MO, USA. [24]Department of Pathology & Immunology, Washington University School of Medicine, St. Louis, MO, USA. [25]Department of Biomedical Engineering, Washington University McKelvey School of Engineering, St. Louis, MO, USA. [26]Department of Molecular Microbiology, Washington University School of Medicine, St. Louis, MO, USA. [27]Department of Biochemistry & Molecular Biophysics, Washington University School of Medicine, St. Louis, MO, USA. [28]BioNote, Hwaseong-si, South Korea. [29]Takeda, Inc., Cambridge, MA, USA. [30]Zymo Research, Irvine, CA, USA. [31]Apoha, London, UK. [32]Johns Hopkins University, Baltimore, MD, USA. [33]University of California, San Diego, CA, USA. [34]Cradle, Zürich, Switzerland. [35]BioLM, Oakland, CA, USA. [36]UTHealth, Houston, TX, USA. [37]Manifold Bio, Boston, MA, USA. [38]ProBioGen, Berlin, Germany. [39]Structural and Molecular Biology, Division of Biosciences, University College, London, UK. [40]The Antibody Society, Framingham, MA, USA. ✉e-mail: mfrank.erasmus@iqvia.com; andrew.bradbury@iqvia.com

https://doi.org/10.1038/s41587-026-03238-6

Methods

Experimental approaches

Antibody production and purification. V_H-encoding sequences were synthesized and subcloned into pS-GWZ293 human IgG1 expression vectors, while V_L-encoding sequences were cloned into pS-GWZ293 human IgG1 lambda or kappa light-chain expression vectors (GENEWIZ, Azenta Life Sciences). Heavy-chain and light-chain plasmids were cotransfected into HEK293F cells at a 1:1 ratio and cultured in a volume of 3 ml with a 5-day expression. Recombinant antibodies were purified from clarified culture supernatants using protein A affinity chromatography. Elution was performed with 0.1 M glycine (pH 3.0), followed by immediate neutralization with 1 M Tris-HCl (pH 8.0).

Developability assessment. To assess the developability profiles of the AIntibody challenge and parental antibodies, we assayed their thermal stability, hydrophobicity, self-interaction and binding to BVPs following protocols outlined by Jain et al.[9]. The performance of the antibodies in each assay was deemed passing, questionable or failing on the basis of an adaptation of Jain et al.'s developability assay thresholds to our assay protocols. To facilitate this, we expressed and purified 16 constructs from Jain et al., where the variable regions of clinical-stage antibodies were formatted onto human IgG1 isotype. The resulting thresholds for construct performance are shown below.

	Construct performance thresholds		
	Passing (<)	**Questionable**	**Failing (>)**
HIC-HPLC retention time (min)	13.9	13.9–14.9	14.9
BVP score	5.3	5.3–10	10
$\Delta\lambda_{max}$(nm)	6.7	6.7–9.5	9.5
	Passing (>)	**Questionable**	**Failing (<)**
T_m(°C)	65	60–65	60
T_{agg}(°C)	64	60–64	60

Each AIntibody and parental antibody received a performance score for each assay of 0 for passing, 1 for questionable and 2 for failing, which was used to define a total score as the sum of performance in each assay. Constructs that had a performance score ≥ 4 were not considered top candidates for each challenge.

For each assay, three constructs from Jain et al. that had passing, questionable and failing developability performance were included, which served as assay controls and as comparators for the AIntibody challenge antibodies.

T_m and T_{agg}. Thermal stability was determined with an Unchained Labs Uncle by using differential scanning fluorimetry to measure T_m and static light scattering to measure T_{agg}. Samples at 0.5 mg ml^{-1} were mixed with 5× SYPRO dye (Thermo, S6651) followed by exposure to a thermal ramp from 25 to 95 °C at a heating rate of 0.5 °C min^{-1} (continuous). Melting and aggregation curves were used to determine T_m and T_{agg} using the Uncle Analysis software. Where multiple transitions were observed, the first transition is reported.

Blosozumab, ponezumab, and bavituximab from Jain et al. were used as the passing, questionable and failing assay controls for T_m and T_{agg} measurements, respectively.

HIC high-performance liquid chromatography (HPLC). Relative hydrophobicity was measured by HIC-HPLC. Constructs (5 µg) were diluted 1:1 in buffer A (1.8 M ammonium sulfate and 0.1 M sodium phosphate, pH 6.5) to achieve a final ammonium sulfate concentration of 0.9 M. Samples were spin-filtered before injection onto a Sepax Proteomix HIC Butyl-NP5 5-µm nonporous column (4.6 × 100 mm; 431NP5-4610). A linear gradient from 100% buffer A to 100% buffer B (0.1 M sodium phosphate, pH 6.5) over 15 min at a flow rate of 0.6 ml min^{-1} was used to elute the constructs and absorbance was monitored at 280 nm. The presence or absence of an eluting peak and the retention time of the main eluting peak in the gradient for each construct was used to judge construct performance. The main peak area and number of peaks observed were also recorded for each construct.

Mepolizumab, onartuzumab and cixutumumab from Jain et al. were used as the passing, questionable and failing assay controls for HIC-HPLC, respectively.

BVP ELISA. Polyspecificity was assessed by BVP ELISA. High-binding plates were coated with BVPs (Medna, E3001) in 50 mM sodium carbonate pH 9.6 overnight at 4 °C while rocking. Plates were washed with 1× PBS followed by blocking for 1 h at room temperature with 2% BSA in 1× PBS. Samples at 1 µM were incubated with the BVP-coated plate for 1 h, followed by washing with 1× PBS to remove unbound sample. Horseradish-peroxidase-conjugated anti-human IgG secondary antibody (Jackson Immunoresearch, 109-035-008) was incubated in the plate for 1 h. The plate was washed with 1× PBS, followed by development with TMB Ultra (Thermo, 34028) for 8–10 min before adding stop solution. Absorbance was read at 450 nm on a Tecan plate reader and BVP score was calculated as follows:

$$\text{BVP Score} = \frac{\text{Sample Absorbance}}{\text{Blank Absorbance}}$$

Panitumumab, ponezumab and fulranumab from Jain et al. were used as the passing, questionable and failing assay controls for BVP ELISA, respectively.

AC-SINS. Self-interaction was assessed by AC-SINS. Gold nanoparticles (Ted Pella, 15705-20) were coated with anti-human goat IgG1 Fc (Jackson Immunoresearch, 109-005-098) overnight in 2 mM sodium acetate pH 4.3. Particles were further prepared by incubating with PEG-thiol (NOF America, SUNBRIGHT ME-020SH) for 1 h. The antibody samples (0.11 mg ml^{-1}) were added to goat nonspecific antibody (Jackson Immunoresearch, 005-000-003) and incubated with particles for 2 h in 10 mM sodium citrate and50 mM NaCl (pH 6.0). Absorbance spectra from 475–625 nm were measured on a Tecan plate reader and the plasmon wavelength λ_{max} was determined using the open-source AC-SINS Analysis R Shiny application provided by Phan et al. The shift in plasmon wavelength ($\Delta\lambda_{max}$) of sample from buffer was used to assess construct performance.

Eculizumab, blosozumab and bavituximab from Jain et al. were used as the passing, questionable and failing assay controls for the AC-SINS ELISA assay, respectively.

Binding kinetics and affinity by HT-SPR. Binding kinetics and affinity analysis were performed by HT-SPR using the Carterra LSA-XT instrument at 25 °C. An SAP sensor chip, which has a planar dextran surface derivatized with streptavidin, was used (Carterra, 4295). The instrument was primed with 1× HBSTE buffer (10 mM HEPES, 150 mM NaCl, 3 mM EDTA and 0.05% Tween-20, pH 7.4; Carterra, 3630) and the chip surface was conditioned with three 1-min injections of 20 mM NaOH and two 1-min injections of 10 mM sodium acetate pH 5.5 (Carterra, 3638 and 3622, respectively). Next a biotinylated anti-human IgG-Fc nanobody monoclonal antibody (mAb; CaptureSelect biotin anti-IgG-Fc (human) conjugate; Thermo Fisher, 71303262100) was injected at 20 µg ml^{-1} in HBSTE running buffer for 15 min to prepare a capture surface with 300–400 RU of the immobilized anti-Fc V_HH-Biotin. The surface was conditioned with two 30-s pulses of 0.85% H_3PO_4 (Carterra, 3637). The instrument was then primed into the assay running buffers used for the capture kinetics sets, which was 1× HBSTE buffer for the 96-channel side and 1× HBSTE + 0.5 mg ml^{-1} BSA (Sigma, A9085) for the single-channel side. Here, 1× HBSTE + 0.5 mg ml^{-1} BSA is

referred to as running buffer and was used as the diluent for both the mAb dilutions and the RBD analytes. mAbs were diluted in running buffer to three concentrations each 1, 0.2 and 0.04 μg ml^{-1} to create different densities of captured mAbs on the surface. A total of 96 unique mAbs were tested per capture kinetics set, with all three capture densities of 96 mAbs captured within a single array and analyzed in parallel. mAbs were captured with 10 min of cycling time using the 96-channel flowcell. Then, the single flowcell was docked for the analyte injections. SARS-Cov2 RBD was injected in a nonregenerative kinetic series at five increasing concentrations from 1.95 nM to 500 nM in a fourfold dilution series. Injections used a 1-min baseline, 5-min association and 15-min dissociation phase. In between capture sets, the surface was regenerated with two sets of three 40-s pulses (six pulses total) of the 0.85% H_3PO_4 and one set of three 15-s pulses of 20 mM NaOH. Ten buffer cycles were injected after array capture before the antigen titration injections. Data were double-referenced using interspot locations as reference surface, which had the streptavidin and the anti-human capture mAb; the leading buffer blank was used as the double reference cycle. Data were *y*-aligned in serial mode and fitted with a 1:1 Langmuir model using the float T_0 option to determine the association rate constant (k_a), the dissociation rate constant (k_d) and R_{max}. Where appropriate, a mass transport term (K_m) was added to the fit. k_a, k_d and R_{max} terms were globally fitted for each individual capture location. A k_d limit was applied to the fits at 2.0×10^{-5} s^{-1}, as this was near the limit of what could be measured in a 15-min dissociation allowing for a 1.8% dissociation of bound analyte. Interestingly, only one mAb, 1291, demonstrated a dissociation rate slower than this limit. The affinity or equilibrium dissociation constant is calculated as $K_D = (k_d/k_a)$. In most cases all three captured concentrations of the test mAbs resulted in RBD binding levels appropriate for analysis, spanning from 300 RU to 30 RU of binding. All usable replicates had the mean and s.d. calculated for the fit parameters and these are reported. In several instances where kinetics required clarification, additional tests were run with additional replicates; these values have $n > 3$ in the analysis table and the same assay conditions were used for the repeated analysis. The mean K_D value reported was calculated as mean k_d/mean k_a and the reported s.d. of K_D was calculated using the propagation of error approach where $K_D\text{ErrorValue} = K_D * \sqrt{\left(\frac{k_a\text{errorvalue}}{k_a}\right)^2 + \left(\frac{k_d\text{errorvalue}}{k_d}\right)^2}$. The SARS-Cov2 RBD construct was generously provided by the La Jolla Institute for Immunology and used the Genbank MN908947.3 sequence and comprised of amino acids 319–591, representing an RBD-XL or RBD + SD1 domain and a 2× Strep tag with a molecular weight of 36.5 kDa. This material was mammalian expressed, affinity purified using the Strep tag and further cleaned with size-exclusion chromatography to ensure a monomeric product.

KinExA measurements. All KinExA measurements were performed at 25 °C with a KinExA 4000 installed in a TC1000 temperature control chamber (both from Sapidyne Instruments). PBS (1×) with 0.02% sodium azide at pH 7.4 was used as the running buffer. The samples were prepared in a sample buffer consisting of running buffer supplemented with 1.0 mg ml^{-1} BSA.

To coat the solid phase, NHS-activated glass beads (Sapidyne Instruments, 440205) were coated with 10 μg of RBD (Bio-Techne) per vial, following Sapidyne Instrument's suggested protocol (Sapidyne Instruments, 'how to' guide HG209).

All antibodies were provided by Azenta Life Sciences and the SARS-CoV-2 RBD used in solution was from the La Jolla Institute for Immunology (LJI).

Equilibrium measurements

The traditional KinExA method was used to measure the K_d (ref. 15). Briefly, a constant concentration of antibody was mixed with a twofold titration series of RBD. The mixtures were incubated until equilibrium was reached then run on the KinExA 4000 to measure the free antibody concentration. For each antibody, two binding curves, at different antibody concentrations, were analyzed in Sapidyne Instruments KinExA Pro software (version 4.7.6) n-Curve module. Information regarding each n-Curve experiment can be found in the Supplementary Information.

KinExA screening

For screening measurements, a sample containing 20 pM antibody only (Sig100) and a sample with 20 pM antibody + 100 pM RBD (inhibited) were prepared for each antibody. A sample with no antibody and only sample buffer was run to generate the nonspecific binding signal (NSB). The concentration of RBD was chosen for the desired cutoff for the screening assay. The samples were incubated at 25 °C for approximately 24 h, until equilibrium was reached. The same capture and detection materials used for the equilibrium measurements were used for screening measurements on a KinExA 4000 instrument. Signals from the two samples and the NSB signal were used to calculate the screening ratio for each antibody: screening ratio = (inhibited signal − NSB)/(Sig100 − NSB). The antibodies were then placed in rank order on the basis of the screening ratio calculation, with a lower screening ratio considered a tighter-binding antibody.

Computational approaches

The challenge 1 winner (Aureka Biotechnology) methodology is outlined below.

Training data curation. The training data comprise both the datasets from the AIntibody competition rounds and our internal synthetic coevolution datasets, which are unrelated to the competition target and use different antibody frameworks. Raw reads were collapsed to unique sequences within each library and each sort population. We removed any sequence containing noncanonical residue symbols such as 'X' or the stop symbol '*'. Variable regions were aligned to their parental frameworks using the International Immunogenetics Information System (IMGT) numbering and records were excluded if framework residues deviated from the parental reference after quality control. For phase-constrained libraries, we enforced phase-specific constancy rules to prevent the inclusion of unintended diversity across CDRs. For example, in phase1_h1_h2_am, H3 and all light-chain CDRs (L1, L2 and L3) were required to match the wild-type sequence exactly. Analogous rules were applied for other phases that varied only a subset of CDRs.

After filtering, sequences were collapsed by exact amino acid identity within the relevant grouping key for the assay, typically by library and sort bin. For each unique sequence, we computed redundancy as the total count of identical reads. The supervised label for model training was defined as label = $\log_{10}$(redundancy). To reduce confounding by batch, we standardized labels within library and round. All curated sequences were stored together with metadata fields including library name, parental identity, phase, antigen, sort gate descriptors and experimental round. This schema enabled phase-aware masking during training and strict separation of training, validation and held-out test sets at the library level.

AuraBind architecture. AuraBind is a structure-informed antibody–antigen modeling framework derived from the Protenix architecture and extended for sequence design. It integrates a pairformer-based structural module with a fitness prediction adapter trained by direct preference optimization (DPO) on large-scale binding datasets. This design allows the model to predict both atomic-level complex structures and relative binding fitness within a unified framework[27–29].

Given antibody and antigen sequences, AuraBind encodes residues and interfaces into shared pairwise representations that capture interchain geometry and epitope–paratope relationships. The structure module refines these representations using a diffusion process to produce 3D coordinates and confidence metrics (predicted local distance difference test and predicted alignment error). The fitness

adapter learns to align model predictions with experimental enrichment data, enabling accurate, structure-aware affinity estimation.

Aurabind training. AuraBind is trained using a DPO objective to align its predictions with experimentally derived fitness rankings. Rather than regressing absolute affinity values, the model learns to reproduce the relative ordering of antibody variants observed in high-throughput enrichment assays. Each minibatch contains antibody–antigen complexes filtered by structural confidence and the model is optimized to assign higher fitness scores to sequences experimentally known to be more enriched. The DPO formulation remains fully differentiable and integrates seamlessly with the model's structure-aware encoder, allowing gradients to propagate through both the sequence and geometric representations.

Antibody generation and selection. For each design task, we sampled 10,000 antibody variants conditioned on the target antigen and the fixed framework and nonvariable regions of the parental antibody, with the designated CDRs masked during generation. The candidate pool was assembled by recombining frequently observed substitutions and mutational patterns derived from the curated training data, ensuring that generated sequences reflected realistic and experimentally supported diversity. Each candidate was then evaluated using the AuraBind fitness predictor, which estimates relative binding affinity from structure-informed representations. Variants were ranked according to their predicted fitness and the top-scoring antibodies were selected for testing. Additional methodological details can be found in Supplementary Note 2.

Code details

All data preprocessing, model architecture and training details necessary for reproducing our results are described in the Methods. Major components of the AuraBind framework, including the pairformer and diffusion modules, can be reproduced using AF3-like architectures such as Protenix (https://github.com/bytedance/Protenix). A comprehensive codebase implementing the methods presented in this work, together with example datasets and inference scripts, is publicly available from GitHub (https://github.com/AurekaBio/Aureka-AIntibody-Challenges).

The challenge 2 cluster 27F cowinner 1 (Scripps; B. Briney and A. Ward lab) methodology is outlined below.

To evaluate the structural convergence of antibody–antigen complex predictions, we used a quantitative ensemble analysis. The metric used, sum of squared root-mean-square deviations (ss-r.m.s.d.), captures both the accuracy and internal consistency of structural ensembles generated by AF3. A lower ss-r.m.s.d. denotes tighter convergence toward a consistent structural solution, while higher values indicate divergent or unstable predictions.

Predicted antibody–antigen complexes (CIF format) were parsed using the Bio.PDB module of Biopython (version 1.83). To minimize sensitivity to side-chain variability, only Cα atom coordinates were retained for analysis, focusing the comparison on global backbone conformations.

Each predicted structure was aligned to the best-ranked AF3 model by least-squares superposition using the Superimposer class in Bio.PDB. After optimal rotation and translation, the r.m.s.d. between corresponding Cα atoms was computed as the square root of the mean squared interatomic distances.

For a given ensemble, individual r.m.s.d. values were squared and summed to produce a single aggregate metric (ss-r.m.s.d.). This squared-sum approach amplifies the influence of structurally divergent models, thereby penalizing ensembles with poor convergence. Lower ss-r.m.s.d. values indicate robust and reproducible predictions, while higher values denote increased uncertainty.

All analyses were performed in Python (version 3.11), using Biopython (version 1.83), NumPy and Matplotlib for data processing and visualization. K_D values were obtained from competition organizers and integrated with model metrics for correlation analysis. Additional methodological details can be found in Supplementary Note 3.

Code details

The code for the ss-r.m.s.d. method is available from GitHub (https://github.com/brineylab/ssrmsd).

The challenge 2 cluster 27F cowinner 2 (USCD; W.W. lab) methodology is outlined below.

For competition 27F, we first filtered the raw data to remove any sequences with framework mutations (as these are not permitted in final competition submissions). We next numbered and aligned the sequences using AntPack version 0.3.7 with the IMGT numbering scheme, thereby converting all sequences to a fixed-length multiple-sequence alignment (MSA). The heavy and light chains were concatenated to form a single MSA.

To train the affinity prediction model, we encoded each amino acid in each sequence using the corresponding row of the PFASUM-62 matrix[30] and then fitted the encoded data to a Gaussian process regression model with an RBF kernel. Hyperparameters were tuned by maximizing the marginal likelihood of the training data using the limited-memory Broyden–Fletcher–Goldfarb–Shanno optimization algorithm as implemented in the scikit-learn library (version 1.6.1).

To train the enrichment prediction model, we first calculated the probability of each competition sequence in the two sorts (more and less stringent), that is, the posterior probability of obtaining that sequence if randomly sampling from that sort using a uniform prior. Next, we calculated the enrichment for each sequence as the log of the ratio of its probability in the more stringent sort to its probability in the less stringent sort, where these values are clipped at a lower bound of 1×10^{-3} to avoid numerical issues. We again encoded the sequence data using the PFASUM-62 matrix[30] and fitted the data to an approximate Gaussian process with a Conv1d-RBF kernel with convolution width = 9 and square-root averaging as implemented in the xGPR library (version 0.4.7)[31]. Hyperparameters were tuned by maximizing the marginal likelihood by following the procedure in the xGPR library documentation (https://xgpr.readthedocs.io/en/latest/) and the model was fitted using conjugate gradients with a randomized Nystrom preconditioner as implemented in xGPR.

To model the distribution of all training set sequences, we fit the one-hot encoded sequences to a mixture of categorical distributions. In this model, the probability of a new sequence x is given by the following formula:

$$p(x) = \sum_{k}^{K} \pi_k \prod_{j}^{L} \prod_{m}^{21} \theta_{kjm}^{\mathbb{I}[x_j = m]}$$

where K is the number of clusters, L is sequence length and 21 is the number of possible amino acids (including a gap token '-'). As we showed previously, it is straightforward and relatively fast to fit this model to 100 million sequences or more using the expectation–maximization (EM) algorithm; therefore, for simplicity, we used EM here. We chose the number of clusters by varying the number of clusters in increments of 10 across the 1–100 range, calculating the Bayes information criterion (BIC) for each value and minimizing the BIC. We ultimately used 80 clusters for this task.

To model the distribution of sequences similar to the high-affinity sequences in the training set, we first selected the subset of sequences with measured affinity that we considered to be high (1×10^{-10} or better). Next, we added to this set all of the sequences from the training data that did not have measured affinity but had a Hamming distance from a high-affinity sequence ≤ 2. We then fitted a mixture of categorical distributions to this combined dataset in the same way we did for the full training set. This 'affinity likelihood' model estimates the probability that a new sequence could have been generated by the observed distribution of tight-binding and closely related sequences in the training set, thereby enabling us to flag sequences very different from

the high-affinity sequences observed during training when generating candidates for experimental evaluation. Lastly, candidates were checked for liabilities and developability using AntPack version 0.3.7 before submission. Additional methodological details and refs. 30–37 can be found in Supplementary Note 3.

Code details

The methodology described here makes use of the AntPack library for antibody sequence data analysis (https://github.com/Wang-lab-UCSD/AntPack), the xGPR library for approximate Gaussian process regression (https://github.com/jlparkI/xGPR) and the resp_protein_toolkit library for iterative sequence design using ML models (https://github.com/jlparkI/resp_protein_toolkit). The code required to reproduce data processing and sequence selection for submission for competitions 27F, 28F and challenge 3 is available from GitHub (https://github.com/Wang-lab-UCSD/aintibody_competition_code), together with instructions for installation and usage.

The challenge 2 cluster 28F cowinner 1 (Xencor) methodology is outlined below.

Sequence + abundance model of affinity

For challenge 2, we constructed an affinity prediction model from the provided data using the subset of 142 antibody V_H–V_L sequences (Nseq) with paired experimental affinity measurements (Extended Data Fig. 3c). The model integrated sequence-derived features with relative abundances from the 1 nM and 10 nM sort rounds obtained by next-generation sequencing. Each antibody sequence was represented numerically using residue embeddings from the final transformer layer of ft-ESM[38], a 650-million-parameter ESM-2 model trained with a masked language modeling objective on 1,335,854 natively paired antibody sequences. Residue embeddings were mean-pooled across sequence length to generate a 1,280-dimensional vector (Hsize) per sequence. The $-\log_{10}$-transformed abundances from both sort rounds were appended to these vectors, producing a 1,282-dimensional descriptor for each sequence. The resulting feature matrix (Nseq × (Hsize + 2)) was used as input to an AutoML workflow (AutoGluon-Tabular[39]), which performed automated model selection, feature selection and stacked ensembling using an 80:20 train–test split, with 20% of the dataset held out to assess model performance. The top-performing estimator (a level-2 weighted stacked ensemble metamodel; WeightedEnsemble_L2) was subsequently applied to score uncharacterized sequences in clusters 27F, 28F and 47F.

Developability assessment

Top-ranked sequences were further evaluated for developability before shortlist selection. Homology models of the V_H and V_L regions were built in MOE (version 2024.0601) using default framework and CDR selection and refinement parameters. The C termini of both chains were capped (amidated and methylated) to prevent artificial negative charges because of truncation. Structural descriptors were computed using MOE's 'protein properties' module. Sequence-level conformity was assessed using pseudoperplexity scores from ft-ESM, with higher values indicating more anomalous sequences.

Shortlist selection

Key structural descriptors (patch_cdr_hyd, ens_charge and Fv_chml) were applied as described previously[40] to eliminate sequences predicted to have biophysical liabilities relative to clinical-stage therapeutic antibodies. Outliers in these descriptors, along with high-pseudoperplexity sequences and redundant designs, were removed to produce the final shortlist.

Code details

A codebase implementing the methods presented in this work is publicly available from GitHub (https://github.com/XenInc/Xencor-AIntibody-Challenges).

The challenge 2 cluster 28F cowinner 2 (WashU, D.H.F. lab) methodology is outlined below.

Training data curation. For challenges 1–3, we used the competition-provided datasets in addition to a noncompetition dataset curated from studies performed by the Xie lab[41,42]. This noncompetition dataset comprised 3,073 SARS-CoV-2 antibodies and their measured neutralization half-maximal inhibitory concentration (IC_{50}) against the Wuhan-Hu-1, BA.1, BA.2, BA.2.75, BA.5, BQ.11 and XBB backgrounds (neutralization assays performed using SARS-CoV-2 variant pseudotyped VSV) and their escape score measured against single-amino-acid variants of the Wuhan-Hu-1 RBD (MACS-based yeast display screen). All data were transformed into log space for computational tractability.

Architecture. We used two transformer-based architectures named AlignNet and) DMS^2. AlignNet provides feature-rich antibody embeddings and addresses the length variability and positional ambiguity introduced by antibody CDRs. This model was trained on the 3,073 SARS-CoV-2 antibody sequences from the noncompetition dataset to reconstruct masked residues while projecting variable-length antibody sequences into a fixed-length reference frame, encouraging consistent positional representations despite CDR insertions, deletions and length heterogeneity. DMS^2 then used the unsupervised antibody embedding from AlignNet along with a paired RBD to predict fitness metrics and was trained jointly on the competition and noncompetition datasets.

Within AlignNet, antibody sequences are embedded using a learnable token embedding layer and then processed in contiguous blocks of 10 aa. This chunking strategy reduces sequence length while preserving local residue context, effectively converting short-range biochemical patterns into higher-level tokens. The model projects the chunked sequence into a fixed-length (300 positions) reference frame using a learned spatial projection. This projection is refined first through dynamic linear transformations along both feature and sequence dimensions and further refined by cross-attention with the original latent antibody representation, enabling flexible, content-dependent reparameterization of the aligned antibody sequence. An alignment prediction layers converts this aligned antibody representation into tokens for a first optimization target. These operations are then repeated to return the antibody sequence to its original representation, which was fed through a masked token prediction layer as an additional optimization target.

Within DMS^2, RBD and antibody sequences are processed by separate encoders and integrated through cross-attention to predict escape, IC_{50} and affinity. Antibody sequences are first passed through the pretrained AlignNet model, with gradients disabled, to extract the learned prelogit embeddings. These embeddings are further processed by a shallow convolutional layer with a kernel size and stride of 10, mirroring the upstream chunking strategy while allowing local smoothing across adjacent blocks. The resulting features are then refined using multiheaded self-attention, producing an antibody representation suitable for interaction modeling. RBD sequences (fixed length of 200 aa) were processed by a separate encoder optimized for relatively low sequence variability. Amino acid tokens were converted to one-hot representations and segmented into contiguous blocks of 10 aa using the same strategy as the antibody encoder. These blocks were then similarly refined using multiheaded self-attention. This design enables efficient modeling of local mutational effects while retaining global context across the RBD. To explicitly model antibody–antigen interactions, a cross-attention module is applied with the RBD representation as the query and the antibody representation as keys and values. This allows each RBD position to attend selectively to antibody features, capturing spatially and chemically relevant interaction patterns that may underlie binding, escape or neutralization potency. The antibody features, RBD features and cross-attended interaction tensor were then flattened and concatenated into a single joint representation. This

representation was passed through a multilayer perceptron (MLP) that serves as a shared trunk for downstream prediction heads (escape, IC_{50} and affinity). This multitask formulation encourages the shared representation to capture generalizable biochemical interaction features while allowing task-specific specialization. Additional methodological details can be found in Supplementary Notes 4 and 5.

Challenge predictions. In predicting the highest-affinity antibodies for challenge 2, antibodies from the dataset were scored by the sum of their predicted escape, IC_{50} and affinity and the top three antibodies from each cluster were used for submission.

Code details

The method implemented by the WashU lab can be found on GitHub (https://github.com/kaszubat/AIntibody_Competition_WashU_Uploads).

The challenge 2 cluster 47F winner (WashU, D.H.F. lab) methodology was the same as that outlined above. The challenge 3 winner (Xencor) methodology is outlined below.

Sequence-alone model of affinity. For challenge 3, we developed a sequence-only affinity model based solely on the same subset of 142 characterized V_H–V_L sequences with paired experimental affinity measurements that were used for challenge 2 (Extended Data Fig. 3c). The model combined ft-ESM residue embeddings with a lightweight regression head incorporating attention-based pooling. Instead of uniformly averaging residue information, the attention mechanism learned position-specific importance across the full set of sequences. The residue-embedding matrices (Nseq × Lseq × Hsize) were first projected through a trainable linear layer (Hsize → 1) and normalized using a softmax function to generate attention weights (Nseq × Lseq) that captured the relative relevance of positions across sequences. These weights were then used to produce attention-pooled sequence embeddings (Nseq × Hsize), which were subsequently processed by a two-layer MLP (architecture Hsize → Hsize → 1) to generate affinity predictions for the entire sequence set.

The complete architecture contained approximately 1.64 million trainable parameters (1,281 from the attention-pooling module and $1{,}280^2$ from the two-layer MLP). Training was performed using the mean squared error loss within the HuggingFace Trainer framework, applying an 80:20 train–test split stratified across five affinity quantiles. Model selection was guided by R^2, Pearson's r and Spearman's ρ on the test set, while overfitting was monitored by comparing training and test loss trajectories. After identifying optimal hyperparameters, the model was retrained on the full dataset and the resulting final model served as the scoring function for evaluating newly designed antibody sequences.

Out-of-library design. We applied both deterministic and stochastic strategies to generate new antibody designs in challenge 3, as discussed below.

Deterministic sequence generation

Candidate CDRs were extracted from the top 100 sequences of clusters 27F and 28F (as ranked by the sequence + abundance model). For 27F, two HCDR1, seven HCDR2, one HCDR3, 40 LCDR1, 33 LCDR2 and seven LCDR3 sequences were collected; for 28F, four HCDR1, 11 HCDR2, two HCDR3, 20 LCDR1, 41 LCDR2 and seven LCDR3 sequences were collected. All unique six-CDR combinations were enumerated, yielding 129,298 new V_H–V_L pairs for 27F and 505,049 for 28F. All novel sequences (without enrichment data) were scored using the sequence-alone model.

Stochastic sequence generation

To explore a multimodal mutation landscape, we used both Metropolis–Hastings Markov chain Monte Carlo (MCMC) and a modified Gibbs sampling scheme, with the sequence-alone model serving as the objective function. Both methods were initialized using the top-ranked challenge 2 antibody (from the sequence + abundance model). Two MCMC samplers were run: (1) unrestricted, with random position → random amino acid substitution (excluding cysteine), and (2) data-restricted, with random position → substitution restricted to amino acids observed at that position in clusters 27F and 28F. Each sampler was run for 100,000 Metropolis steps. For Gibbs sampling, at each iteration, we allowed up to three substitutions, applying top-k truncation ($k = 8$) to retain the highest-scoring substitutions plus the original amino acid. A total of 20,000 iterations were run.

Developability assessment. The same methodology used by Xencor for challenge 2 was applied.

Shortlist selection. The same methodology used by Xencor for challenge 2 was applied. Additional methodological details can be found in Supplementary Note 6.

Code details

A codebase implementing the methods presented in this work is publicly available from GitHub (https://github.com/XenInc/Xencor-AIntibody-Challenges).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All data generated in this study are provided as Supplementary Information: (1) all SPR kinetic parameters (k_a, k_d, K_D, R_{max}) for every submission; (2) KinExA equilibrium and screening data for all tested antibodies; (3) individual developability assay scores (T_m, T_{agg}, HIC retention time, BVP score and AC-SINS Δλ) for all submissions; (4) processed and annotated sequencing datasets used as input for all three challenges; (5) the complete sequence-to-result mapping table. All data are provided in standard, machine-readable formats (for example, Excel and CSV). There are no restrictions on data availability. No human-subject data, clinical data or other access-restricted datasets were used. Participant code repositories are listed below and in the Methods.

Code availability

The codebases implementing the methods for each challenge presented in this work are publicly available from GitHub: Challenge 1 (Aureka), https://github.com/AurekaBio/Aureka-AIntibody-Challenges; challenge 2-27F, cowinner 1 (Scripps, B. Briney and A. Ward lab), https://github.com/brineylab/ssrmsd; challenge 2-27F, cowinner 2 (UCSD, W.W. lab), https://github.com/Wang-lab-UCSD/AntPack, https://github.com/jlparkI/xGPR, https://github.com/jlparkI/resp_protein_toolkit and https://github.com/Wang-lab-UCSD/aintibody_competition_code; challenge 2-28F, cowinner 1 (Xencor), https://github.com/XenInc/Xencor-AIntibody-Challenges; challenge 2-28F and 2-47F, cowinner 2 (WashU, D.H.F. lab), https://github.com/kaszubat/AIntibody_Competition_WashU_Uploads; challenge 3 (Xencor), https://github.com/XenInc/Xencor-AIntibody-Challenges.

References

27. Abramson, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
28. Team, B. A. A. S. et al. Protenix—advancing structure prediction through a comprehensive AlphaFold3 reproduction. Preprint at *bioRxiv* https://doi.org/10.1101/2025.01.08.631967 (2025).
29. Rafailov, R. et al. Direct preference optimization: your language model is secretly a reward model. In *Proc. 37th International Conference on Neural Information Processing Systems* https://proceedings.neurips.cc/paper_files/paper/2023/file/a85b405ed65c6477a4fe8302b5e06ce7-Paper-Conference.pdf (2023).

30. Keul, F., Hess, M., Goesele, M. & Hamacher, K. PFASUM: a substitution matrix from Pfam structural alignments. *BMC Bioinformatics* **18**, 293 (2017).
31. Parkinson, J. & Wang, W. Linear-scaling kernels for protein sequences and small molecules outperform deep learning while providing uncertainty quantitation and improved interpretability. *J. Chem. Inf. Model.* **63**, 4589–4601 (2023).
32. Hastie, T., Tibshirani, R., Friedman, J. & Franklin, J. *The Elements of Statistical Learning: Data Mining, Inference and Prediction* 2nd edn (Springer, 2005).
33. Ko, Y. S., Parkinson, J., Liu, C. & Wang, W. TUnA: an uncertainty-aware transformer model for sequence-based protein–protein interaction prediction. *Brief. Bioinform.* **25**, bbae359 (2024).
34. Parkinson, J., Hard, R. & Wang, W. The RESP AI model accelerates the identification of tight-binding antibodies. *Nat Commun* **14**, 454 (2023).
35. Parkinson, J., Hard, R., Ko, Y. S. & Wang, W. RESP2: an uncertainty aware multi-target multi-property optimization AI pipeline for antibody discovery. *Adv. Sci. (Weinh.)* **12**, e04350 (2025).
36. Parkinson, J. & Wang, W. For antibody sequence generative modeling, mixture models may be all you need. *Bioinformatics* **40**, btae278 (2024).
37. Murphy, K. P. *Probabilistic Machine Learning: Advanced Topics* (MIT Press, 2023).
38. Burbach, S. M. & Briney, B. Improving antibody language models with native pairing. *Patterns (N. Y.)* **5**, 100967 (2024).
39. Erickson, N. et al. Autogluon-tabular: robust and accurate automl for structured data. Preprint at https://arxiv.org/abs/2003.06505 (2020).
40. Thorsteinson, N., Gunn, J. R., Kelly, K., Long, W. & Labute, P. Structure-based charge calculations for predicting isoelectric point, viscosity, clearance and profiling antibody therapeutics. *MAbs* **13**, 1981805 (2021).
41. Cao, Y. et al. BA.2.12.1, BA.4 and BA.5 escape antibodies elicited by Omicron infection. *Nature* **608**, 593–602 (2022).
42. Cao, Y. et al. Imprinted SARS-CoV-2 humoral immunity induces convergent Omicron RBD evolution. *Nature* **614**, 521–529 (2023).

Acknowledgements

We would like to send a notable acknowledgement to Johannes R. Loeffler from the Scripps Research Institute for his work contributing to the design and prediction challenges presented throughout this manuscript. We thank members of the Mosaic Protein Sciences Team (A. Barbour, M. Braun, P. Ceres, M. Weir, H. Barrett, N. Collins and A. Jackson) for their contributions to the developability analysis of the challenge antibodies. We also thank R. Chan, A. Person, N. Nath and H. Beernink at Bio-Techne for supply of the RBD antigen. We thank K. Hastie and LJI for supplying antigens.

Author contributions

A.R.M.B. and M.F.E. conceptualized and coordinated the AIntibody challenge, performed data analysis and interpretation and wrote the manuscript. M.F.E., L.S., K.P.-S., S.D., F.F. and A.R.M.B. (Specifica) administered the challenge submissions and provided the challenge data. Experimental methodology and execution, D.B. (Carterra, HT-SPR); C.P.G., S.R.S., S.L.C., and J.D.S. (Mosaic Biosciences, developability panel); E.H., G.F. and J.S. (Sapidyne Instruments, KinExA); C.R., Sumit K., Z.S. and Y. Shang (GENEWIZ from Azenta Life Sciences, antibody expression and supply). Winning method development, submission and writeup, J.Z., M. Gu, L.Y., A. Wellner, S.Z. and W.L. (Aureka, challenge 1—AuraBind); G.L.M., H.S., Y. Ding, A.N., J.K. and M.J.B. (Xencor, challenge 2-28F and challenge 3); M.L.F.-Q., N.M., S.M.C., O.M.S., D.L.V.B., J.A.F., S.S.R.R., B.N., C.A.M., C.B., B. Briney, A. Ward (Scripps, challenge 2-27F); J.P., Y.S.K. and W.W. (UCSD, challenge 2-27F); T.K. and D.H.F. (WashU, challenges 2-28F and 2-47F); J. Klattig, K.W., T.B. and V.S. (ProBioGen). Additional algorithm submission and methods writeup: P.S. and M. Greenig (Cambridge); H.Z. (Shenzhen University of Advanced Technology); B.T.M., M. Baghbanzadeh and A.R. (George Washington University); A.P., Y. Shao, C.L., L.D. and M.R.M. (Yale, Zurich University of Applied Sciences and University College London); K.R. and H.N. (Incyte); A.E. (Merck); A.A., A.J.B., M. Bordyuh, L.H., B.A.K., H.S.T. and S.W. (BMS); J.E.S., R.P. and L.M. (Seismic); A.L. and Y. Devasurmutt (Locksmith); B.G., L.C. and M.M. (InstaDeep); P.M. and R.A. (Novo Nordisk); B. He, F.W. and J.Y. (Tencent); B. Hu, M.K., K.R.G., R.S. and L.-W.H. (Los Alamos National Laboratories); H.J. (BioNote); V.B.K., S. Malhotra and S.K. (Takeda); Y.L., L.X. and J.M. (Zymo); A.N.S.J. (Apoha); J.V., F.A.D.-H., D.X., M.C., Z.D.H. and J.J.G. (Johns Hopkins); F.G. and J.D.Z. (Cradle); N.H., Y.K., C.C. and A.Q. (BioLM); A.S., Y.-C.T., Z.A., X.J. and X.Z. (UTHealth Houston); E.S. (Manifold Bio); R.J.B. (The Antibody Society). All authors reviewed and approved the final manuscript.

Funding

The Los Alamos National Laboratories authors gratefully acknowledge funding support from Los Alamos National Laboratory (Laboratory Directed Research and Development Director-Initiated Research, grant nos. 20250637DI, 20250638DI, 2020250639DI and 20240734DI). The George Washington University team's work was supported by the US National Science Foundation (grant no. 2109688 to A.R.). The Shenzhen University of Advanced Technology author's work was supported by the Shenzhen Medical Research Fund (grant no. B2404003). The Yale School of Medicine authors gratefully acknowledge the Swiss National Science Foundation (SNSF), through Sinergia grant CRSII5_193832 and project funding grant 200021_192128. The UTHealth team's work was supported by NIH (R01AG082721, R01AG084637) and the Welch Foundation (AU- 0042- 20030616). The Washington University in St. Louis team's work was supported in part by NIAID contract 75N93022C00035 (CSBID) and by NIH T32 HL134635.

Competing interests

The following authors are or were employees of the indicated companies during the conduct of this work: M.F.E., L.S., K.P.-S., S.D., F.F. and A.R.M.B. (Specifica, an IQVIA business); D.B. (Carterra); E.H., G.F. and J.S. (Sapidyne Instruments); C.P.G., S.R.S., S.L.C. and J.D.S. (Mosaic Biosciences); C.R., Sumit K., Z.S. and Y. Shang (GENEWIZ from Azenta Life Sciences); J.Z., M. Gu, L.Y., A. Wellner, S.Z. and W.L. (Aureka Biotechnologies); G.L.M., H.S., Y. Ding, A.N., J.K. and M.J.B. (Xencor); K.R. and H.N. (Incyte); A.E. (Merck); A.A., A.J.B., M. Bordyuh, L.H., B.A.K., H.S.T. and S.W. (Bristol-Myers Squibb); J.E.S., R.P. and L.M. (Seismic Therapeutic); A.L. and Y. Devasurmutt (Locksmith Bio); B.G., L.C. and M.M. (InstaDeep); P.M. and R.A. (Novo Nordisk); B. He, F.W. and J.Y. (Tencent); H.J. (BioNote); V.B.K., S. Malhotra and S.K. (Takeda); Y.L., L.X. and J.M. (Zymo Research); A.N.S.J. (Apoha); F.G. and J.D.Z. (Cradle); N.H., Y.K., C.C. and A.Q. (BioLM); E.S. (Manifold Bio); J. Klattig, K.W., T.B. and V.S. (ProBioGen); R.J.B. (The Antibody Society); A.R. (seqSight). The other authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41587-026-03238-6.

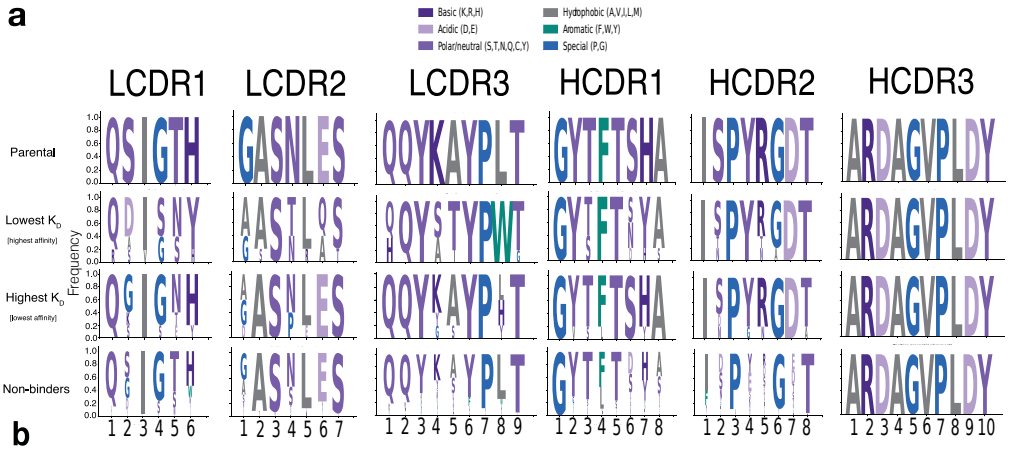
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-026-03238-6.

Correspondence and requests for materials should be addressed to M. Frank Erasmus or Andrew R. M. Bradbury.

Peer review information *Nature Biotechnology* thanks the anonymous reviewers for their contribution to the peer review of this work.

https://doi.org/10.1038/s41587-026-03238-6

a



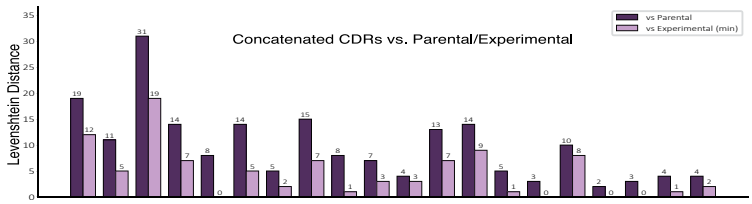
b

Rank	Org.	Type	Method	DQ	LCDR1	LCDR2	LCDR3	HCDR1	HCDR2	HCDR3	KD KinExA (M) (95% c.i.)	KD SPR (M) ± StDev
1	**Aureka**	**AI bio**	**PLM, SA**		**RDISSY**	**GASTRAS**	**HQYATYPWG**	**GYSFTNYA**	**ITPYNADT**	**ARDAGVPLDY**	**9.47e-11**	**3.40e-10**
											(6.9e-11–1.3e-10)	**(±1.9e-11)**
2	Aureka	AI bio	PLM, SA		QAIGNH	ASSNLAS	HQYATYPWT	GYSFTDHS	INPYTGDT	ARDAGVPLDY	2.58e-10	3.70e-10
											(1.7e-10–4.2e-10)	(±7.2e-11)
3	ProBioGen-22	Sm/Mid bio	Stat		QDISNY	AASTLQS	QQYSTYPWT	GYTFTSYA	ISPYRGDT	ARDAGVPLDY	5.38e-10	1.59e-09
											(3.0e-10–9.1e-10)	(±3.6e-10)
4	ProBioGen-22	Sm/Mid bio	Stat		QSVSSY	GASNLQT	QQYSIYPWT	GYTFTNYA	ISPYRGDT	ARDAGVPLDY	9.22e-10	7.40e-10
												(±5.6e-11)
5	Aureka	AI bio	PLM, SA		QDIGNY	AASTLQS	QQYSTYPWT	GYTFTSYA	ISPYRGDT	ARDAGVPLDY	9.84e-10	1.40e-09
											(5.8e-10–1.4e-09)	(±1.7e-10)
-	18	Academic	CML	DEV	GYIFST	SAHSIGT	HISPYRGDT	GSSILESQ	QYKAYPLT	ARDAGVPLDY	N/A	1.40e-09
												(±0.0e+00)
6	Aureka	AI bio	PLM, SA		QDISNY	AASTLQS	QQYSTYPWT	GYTFTSYA	ISPYRGDT	ARDAGVPLDY	N/A	1.40e-09
												(±3.6e-10)
-	ProBioGen-22	Sm/Mid bio	Stat	DEV	HNIGTY	AASTLHT	QQYSNYGWT	GYAFTDYA	ISPYRGDT	ARDAGVPLDY	N/A	1.60e-09
												(±6.9e-11)
-	9	Sm/Mid bio	PLM, SA	DEV	RDIRNY	AVSALQT	QHYSTFPWT	GGIFSSSA	ISPYRGDT	ARDAGVPLDY	N/A	1.80e-09
												(±1.9e-10)
-	21	Academic	CML, SA	DEV	QSIGTH	GASNLES	QQYKAYPLT	GYTFSTFY	MSPKSGYT	ARDAGVPLDY	N/A	1.90e-09
												(±0.0e+00)

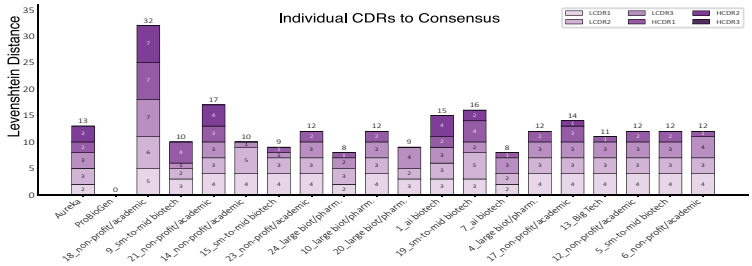
PLM = Protein Language Model/Attention-based
SA = Structure-aware
CML = Classical ML
Stat = Statistical / Heuristic

bold / light purple = winner
DQ = disqualification reason
DEV = developability score >3

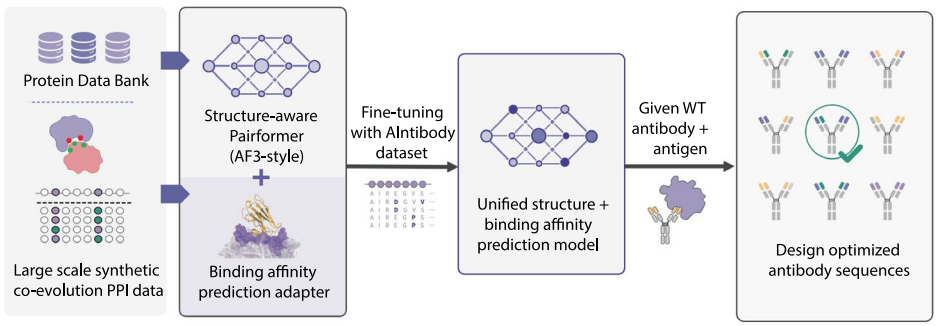
c



d



e



Extended Data Fig. 1 | See next page for caption.

https://doi.org/10.1038/s41587-026-03238-6

Extended Data Fig. 1 | Affinity Maturation Challenge & Winning Method Overview. (a) Weblogos comparing CDR sequence diversity (in order) for parental, lowest KD (highest affinity), highest KD (lowest affinity), and non-binders. **(b)** Final ranking of the top 10 designs by KinExA K_D, then SPR K_D. **(c)** Levenshtein distance using the concatenated CDRs (LCDR1-3 & HCDR1-3) of the top ranked antibody submission in challenge 1 to parental antibody (dark purple) and the minimum Levenshtein distance of top ranked antibody submission to closest sequence within the provided dataset. **(d)** Per-CDR Levenshtein distance to the composite consensus sequence for the top 20 Challenge 1 submissions. Edit distances were computed independently for each of the six CDRs (LCDR1-3, HCDR1-3) between each submission and a library-specific composite consensus derived from provided dataset. The composite consensus was derived by computing the positional majority-vote residue at each position within each CDR's diversified sort populations and ranked left to right by best measured affinity. **(e)** Schematic of the winning methodology (AuraBind by Aureka), utilizing a structure-aware pairformer and Direct Preference Optimization (DPO) to navigate the fitness landscape. Protein structure database and large-scale synthetic antibody–antigen co-evolution data are used to train a structure-aware pairformer architecture equipped with a binding affinity prediction adapter. After fine-tuning on the Antibody dataset, the unified model jointly predicts complex structures and binding fitness, enabling sequence optimization to propose high-affinity antibody variants given a WT antibody–antigen pair.

https://doi.org/10.1038/s41587-026-03238-6

a

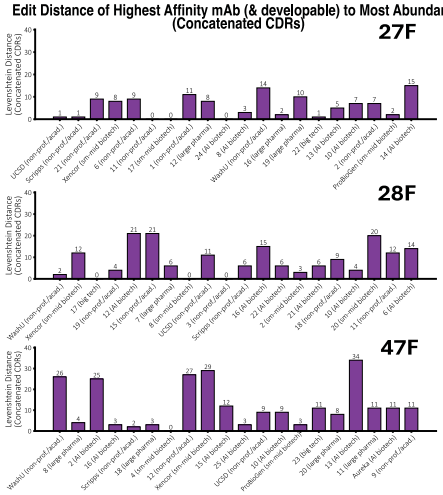
Most abundant, 27F: 37pM (95% c.i. = 18.8 - 60.1pM)
Most abundant, 28F: 177pM (95% c.i. = 131 - 237pM)
Most abundant, 47F: 378pM (95% c.i. = n.d.)

PLM = Protein Language Model/Attention-based
SA = Structure-aware
CML = Classical ML
Stat = Statistical / Heuristic

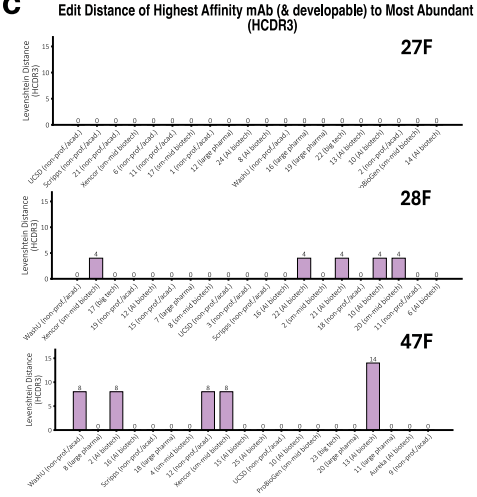
bold/light purple = winner
DQ = disqualification reason
DEV = developability score >3
SEQ = did not meet challenge rules

	Cluster	Rank	Org.	Type	Method	DQ	LCDR1	LCDR2	LCDR3	HCDR1	HCDR2	HCDR3	KD KinExA (M) (95% c.i.)	KD SPR (M) (± StDev)	
27F	**2-27F**	**1**	**UCSD**	**Academic**	**CML**		**QDIGNH**	**ASSNLAS**	**HQYATYPWT**	**GYSFTDHS**	**INPYIGDT**	**ARDGYGDYRGLYGMDV**	**9.24e-12 (7.5e-12–1.1e-11)**	**3.80e-11 (±1.0e-11)**	**1st**
	2-27F	**1**	**Scripps**	**Academic**	**PLM, SA**		**QDIGNH**	**ASSNLAS**	**HQYATYPWT**	**GYSFTDHS**	**INPYIGDT**	**ARDGYGDYRGLYGMDV**	**1.02e-11 (8.1e-12–1.3e-11)**	**3.20e-11 (±8.1e-12)**	**1st**
	2-27F	-	Scripps	Academic	PLM, SA	DEV	QDIFNY	DVNKRPS	HQYATYPWT	GYSFTDHS	INPYTGDT	ARDGYGDYRGLYGMDV	1.46e-11 (1.2e-11–1.8e-11)	4.80e-11 (±1.5e-11)	
	2-27F	3	21	Academic	PLM, SA		QDIDTY	DATIRAT	HQYATYPWT	GYSFTDHS	INPYTGDT	ARDGYGDYRGLYGMDV	1.65e-11 (1.4e-11–1.9e-11)	4.70e-11 (±8.6e-12)	
	2-27F	-	5-Aureka	AI bio	PLM, SA	SEQ	QDIDTY	ASSNLAS	HQYATYPWT	GYSFTDHS	INPYTGDT	ARDGYGDYRGLYGMDV	1.66e-11 (1.2e-11–2.2e-11)	4.80e-11 (±4.3e-12)	
	2-27F	4	4-Xencor	Sm/Mid bio	PLM		QDIDTY	MGSGRVS	HQYATYPWT	GYSFTDHS	INPYTGDT	ARDGYGDYRGLYGMDV	1.82e-11 (1.3e-11–2.5e-11)	5.40e-11 (±4.1e-12)	
	2-27F	5	6	Academic	PLM, SA		QDIDTY	EANKRPS	HQYATYPWT	GYSFTDHS	INPYTGDT	ARDGYGDYRGLYGMDV	2.02e-11 (1.6e-11–2.5e-11)	6.60e-11 (±3.2e-12)	
	2-27F	6	17	Sm/Mid bio	PLM, SA		QDIGNH	ASSNLAS	HQYATYPWT	GYSFTDHS	INPYTGDT	ARDGYGDYRGLYGMDV	3.66e-11 (1.9e-11–6.0e-11)	1.40e-10 (±1.6e-11)	
	2-27F	6	11	Academic	CML		QDIGNH	ASSNLAS	HQYATYPWT	GYSFTDHS	INPYTGDT	ARDGYGDYRGLYGMDV	3.66e-11 (1.9e-11–6.0e-11)	1.60e-10 (±1.7e-11)	
	2-27F	-	20	AI bio	PLM	DEV	RSIGTY	NVSNRFS	HQYATYPWT	GYSFTDHS	INPYTGDT	ARDGYGDYRGLYGMDV	N/A	6.00e-11 (±1.1e-11)	
28F	2-28F	-	1	Academic	PLM, SA	DEV	RGISTY	AASNLQT	HQYATYPWT	GYSFTDHS	INPYTGDT	ARDGYDFWSGSYGMDV	7.55e-11 (6.2e-11–9.1e-11)	1.70e-10 (±3.9e-11)	
	2-28F	-	14	Large pharma	SA	DEV	QTIGTY	GASTRAS	HQYATYPWT	GYSFTDHS	INPYTGDT	ARDGYDFWSGSYGMDV	8.97e-11 (6.7e-11–1.2e-10)	3.50e-10 (±6.2e-11)	
	2-28F	-	4-Aureka	AI bio	PLM, SA	SEQ	QDIGNH	QDNKRPL	HQYATYPWT	GYSFTDHS	INPYTGDT	ARDGYDFWSGSYGMDV	1.01e-10 (8.1e-11–1.3e-10)	4.10e-10 (±5.9e-11)	
	2-28F	-	14	Large pharma	SA	DEV	QTIGTY	GASTRAS	HQYATYPWT	GYSFTDHS	INPYTGDT	ARDGYDFWSGSYGMDV	1.03e-10 (8.5e-11–1.2e-10)	4.00e-10 (±6.6e-11)	
	2-28F	**1**	**WashU**	**Academic**	**PLM**		**QDIKNY**	**ASSNLAS**	**HQYATYPWT**	**GYSFTDHS**	**INPYTGDT**	**ARDGYDFWSGSYGMDV**	**1.06e-10 (8.3e-11–1.3e-10)**	**1.70e-10 (±2.6e-11)**	**1st**
	2-28F	**1**	**Xencor**	**Sm/Mid bio**	**PLM**		**QDIRHY**	**EASTLKP**	**HQYATYPWT**	**GYSFTDHS**	**INPYTGDT**	**ARDGYGSGSRSYGMDV**	**1.07e-10 (8.9e-11–1.3e-10)**	**3.10e-10 (±2.6e-11)**	**1st**
	2-28F	-	9-Scripps	Academic	PLM, SA	DEV	QSIGTY	SGSTLQS	HQYATYPWT	GYSFTDHS	INPYTGDT	ARDGYDFWSGSYGMDV	N/A	1.40e-10 (±4.7e-11)	
	2-28F	-	4	AI bio	PLM, SA	DEV	QNIDTF	ASSNLAS	HQYATYPWT	GYSFTDHS	INPYTGDT	ARDGYDFWSGSYGMDV	1.48e-10 (1.1e-10–2.0e-10)	1.80e-10 (±3.2e-11)	
	2-28F	-	ProBioGen-24	Sm/Mid bio	Stat	DEV	QDIDTY	ASSNLAS	HQYATYPWT	GYSFTDHS	INPYTGDT	ARDGYDFWSGSYGMDV	N/A	1.80e-10 (±4.2e-11)	
	2-28F	-	5	Large pharma	CML	DEV	QDIDTY	ASSNLAS	HQYATYPWT	GYSFTDHS	INPYTGDT	ARDGYDFWSGSYGMDV	N/A	1.90e-10 (±1.1e-11)	
47F	**2-47F**	**1**	**WashU**	**Academic**	**PLM**		**QDIRHY**	**ASSNLAS**	**HQYATYPWT**	**GYSFTDHS**	**INPYTGDT**	**ARDGYYYDSSGLPTGDAFDI**	**5.00e-11 (3.6e-11–6.7e-11)**	**1.10e-10 (±6.6e-12)**	**1st**
	2-47F	2	WashU	Academic	PLM		RDISHH	ASSNLAS	HQYATYPWT	GYSFTDHS	INPYTGDT	ARDGYYYDSSGLPTGDAFDI	1.07e-10 (7.6e-11–1.4e-10)	4.10e-10 (±2.4e-11)	
	2-47F	3	WashU	Academic	PLM		QNIDNY	ASSNLAS	HQYATYPWT	GYSFTDHS	INPYTGDT	ARDGYYYDSSGLPTGDAFDI	1.16e-10 (9.6e-11–1.4e-10)	2.50e-10 (±3.5e-11)	
	2-47F	4	8	Large pharma	SA		RDISSY	SGSTLQS	LQYSIYPWT	GYTSPAT	ISPYKGDT	AKDSYYYDSSGSGLGGGFDY	1.24e-10 (9.7e-11–1.6e-10)	3.20e-10 (±4.7e-11)	
	2-47F	5	2	AI bio	PLM		RDISHH	ASSNLAS	HQYATYPWT	GYSFTDHS	INPYTGDT	ARDGYYYDSSGLPTGDAFDI	1.55e-10 (1.1e-10–2.2e-10)	4.10e-10 (±3.8e-11)	
	2-47F	6	16	AI bio	PLM		RTIGTY	GASTRAS	LQYSIYPWT	GYTSPAT	ISPYKGDT	AKDSYYYDSSGSGLGGGFDY	1.61e-10 (1.1e-10–2.2e-10)	6.00e-10 (±6.3e-11)	
	2-47F	7	2	AI bio	PLM		QDIGNH	ASSNLAS	HQYATYPWT	GYSFTDHS	IGPGQGDT	ARDGYYYDSSGLPTGDAFDI	N/A	4.60e-10 (±5.6e-11)	
	2-47F	8	8	Large pharma	SA		QDISHY	GASTRAS	LQYSIYPWT	GYTSPAT	ISPYKGDT	AKDSYYYDSSGSGLGGGFDY	N/A	5.50e-10 (±9.0e-11)	
	2-47F	8	1-Scripps	Academic	PLM, SA		RDINKY	GASTRAS	LQYSIYPWT	GYTSPAT	ISPYKGDT	AKDSYYYDSSGSGLGGGFDY	N/A	5.50e-10 (±6.3e-11)	
	2-47F	10	4	Sm/Mid bio	PLM, SA		RDISSY	GASTRAS	LQYSIYPWT	GYTSPAT	ISPYKGDT	AKDSYYYDSSGSGLGGGFDY	N/A	5.80e-10 (±6.1e-11)	

b



c



d

	27F			28F			47F			% Abs better than control
	pM affinity	KinExA/SPR	Rank	pM affinity	KinExA/SPR	Rank	pM affinity	KinExA/SPR	Rank	
Cluster control	37	K	-	177	K	-	378	K	-	-
UCSD	9.2	K	1	400	S	10	-	-	-	25%
Scripps	10.2	K	1	480	S	12	550	S	8	22%
Xencor	18.2	K	4	107	K	1	810	S	16	22%
WashU	140	S	15	106	K	1	50	K	1	50%
Aureka	17*	K	*	101*	K	*	4,700	S	34	11%
Random	11/21 (52%) better than control			3/13 (23%) better than control			0/2 (0%) better than controls			39%

*disqualified by developability or sequence

light purple = winner in at least one challenge

Extended Data Fig. 2 | See next page for caption.

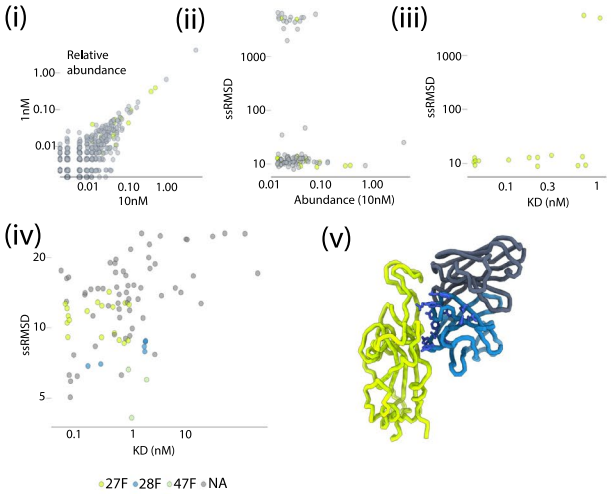
https://doi.org/10.1038/s41587-026-03238-6

Extended Data Fig. 2 | Cluster Prediction Challenge Overview. (a) Top-10 ranking tables for each cluster, flagging disqualified entries. Minimum Levenshtein distance of the top ranked antibody submissions in challenge 2 to **(b)** the concatenated CDRs (LCDR1-3 & HCDR1-3) and **(c)** the HCDR3 of the most abundant sequencing control in each corresponding cluster. **(d)** Performance of different winning algorithms across different clusters, with the winner in each cluster highlighted. For each algorithm the percentage of total antibodies predicted with affinities better than the cluster control is indicated. Aureka was a winner in Challenge 1, and used the same algorithm for Challenges 2 and 3. The two disqualifications were for the introduction of framework mutations.

https://doi.org/10.1038/s41587-026-03238-6

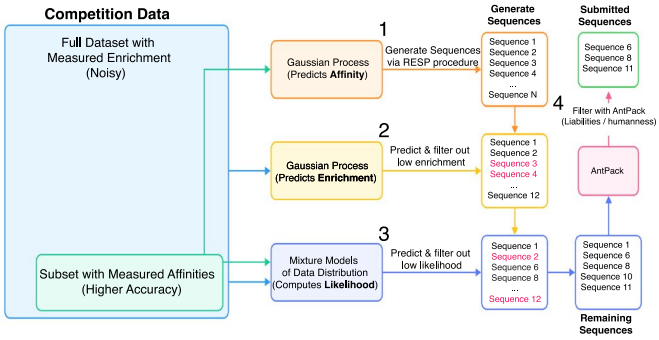
a - Challenge 2 (27F), co-winner #1

Scripps, Briney/Ward Lab



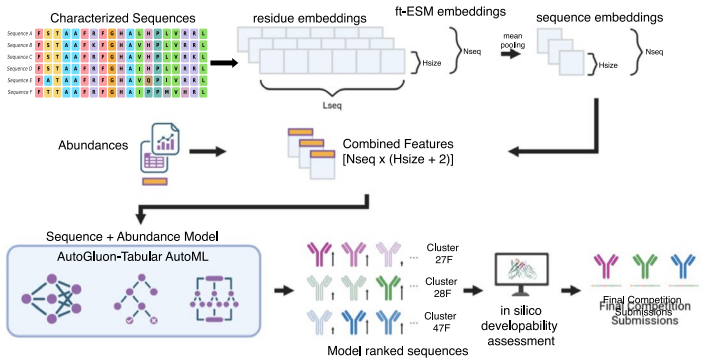
b - Challenge 2 (27F), co-winner #2

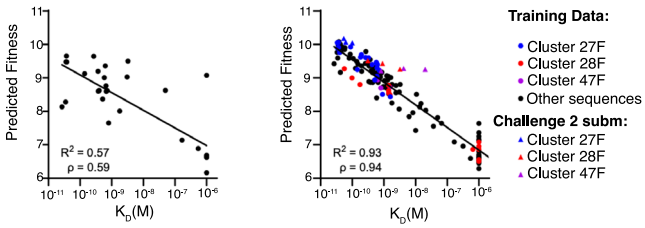
UCSD, Wang Lab



c - Challenge 2 (28F), co-winner #1

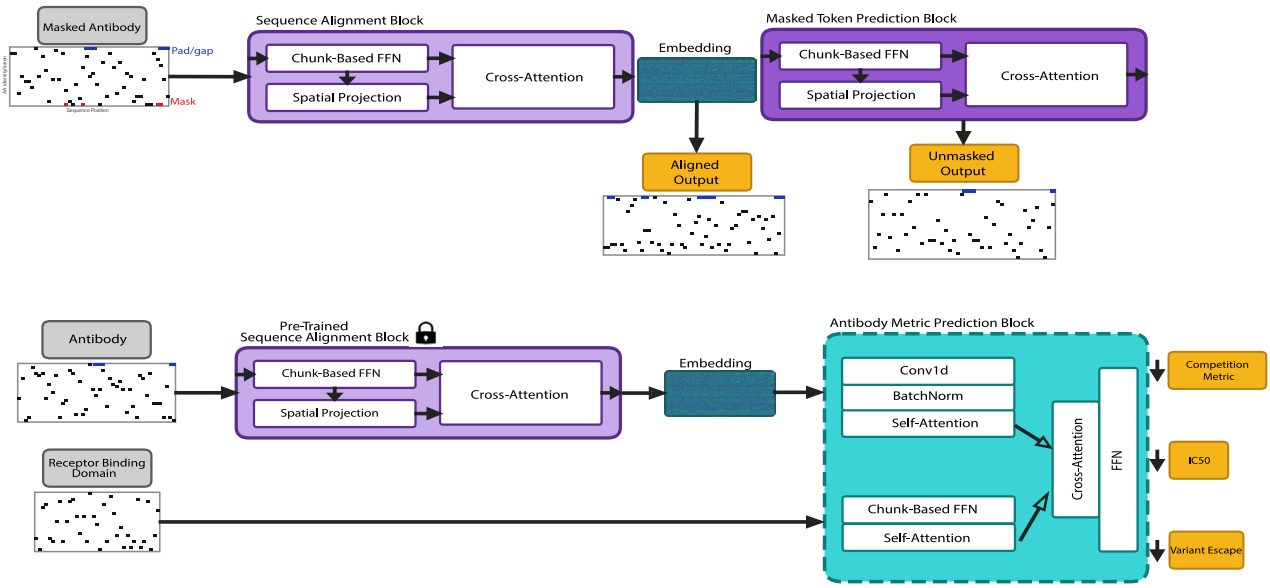
Xencor





d - Challenge 2 (28F), co-winner #2 & Challenge 2 (47F) winner

WashU, Fremont Lab

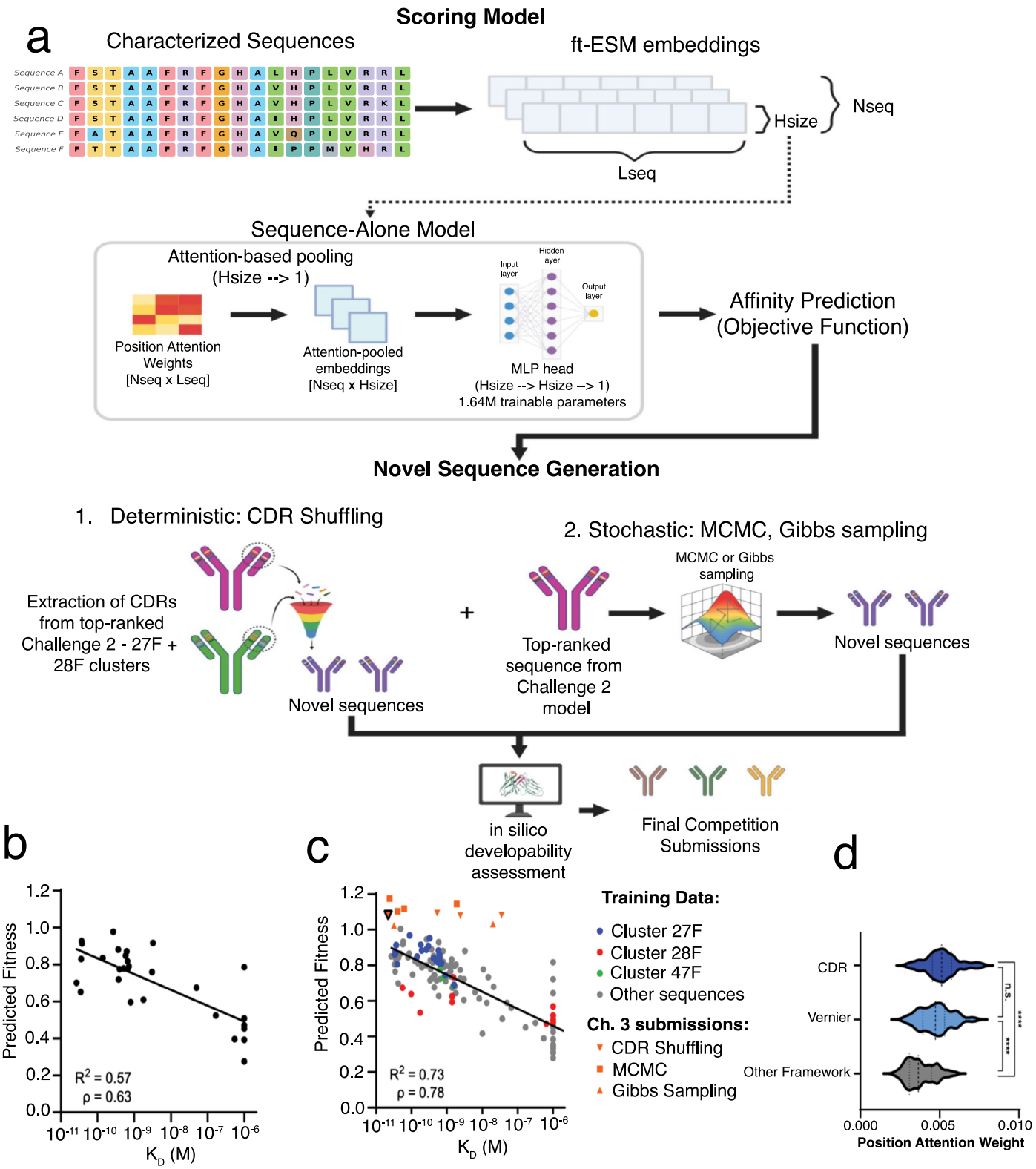


Extended Data Fig. 3 | See next page for caption.

https://doi.org/10.1038/s41587-026-03238-6

Extended Data Fig. 3 | Cluster Prediction Challenge Winning Schematics. **(a)** Scripps, Ward Lab (winner 27F) utilized structural convergence, using the "sum of squared RMSD" (ssRMSD) of the predicted ensemble as a proxy for physical realizability. **(a, (i))** Relative abundance of RBD-binding antibodies following sorting at 10 nM and 1 nM antigen concentrations. Antibodies displaying consistent enrichment (relative abundance > 0.01 at both concentrations) were selected for structural modeling using AlphaFold3 (AF3). **(a,(ii))** Relationship between the sum of squared root-mean-square deviations (ssRMSD) and relative antibody abundance at 10 nM antigen. Antibodies with experimentally determined binding affinities (K_D) are indicated by colored markers. **(a,(iii))** ssRMSD versus experimentally measured K_D for antibodies from cluster 27. Although most antibodies with nanomolar affinity exhibited low ssRMSD values (< 10), two outliers displayed high ssRMSD (> 1000), indicating structural divergence despite strong binding. **(a,(iv))** ssRMSD versus K_D across all antibodies from challenge 2, encompassing clusters 27, 28, and 47, as well as unclassified antibodies. The broader dataset reinforces the observed inverse relationship between ssRMSD and binding affinity. **(a,(v))** AF3-predicted structure of the antibody selected for experimental validation, chosen based on low ssRMSD and high relative abundance. The antigen (neon yellow) and antibody (heavy chain in blue, light chain in dark gray) are depicted as ribbon representations. CDRH2 and CDRH3 residues are shown in dark blue. **(b)** UCSD, Wang Lab (winner 27F) employed a Gaussian Process regression model coupled with "AntPack" to filter for developability liabilities. **(c)** Xencor (co-winner 28F) VH–VL sequences (Nseq = 142) are transformed into residue-level embeddings via a fine-tuned antibody language model (ft-ESM, 650 M parameters) and mean-pooled across sequence length (Lseq) into 1,280-dimensional vectors (Hsize). These vectors are fused with sequencing-derived relative abundances (1 nM and 10 nM) to form a [Nseq x (Hsize + 2)] feature matrix. An AutoML ensemble (AutoGluon-Tabular) identifies the optimal affinity predictor to score and rank uncharacterized sequence clusters (Sequence+Abundance Model). Top-ranked candidates undergo *in silico* developability filtering—comprising homology modeling (MOE) and ft-ESM pseudo-perplexity scoring—to produce the final competition submissions. Predicted fitness and provided experimental K_D values for the test (left regression plot) and full dataset (right regression plot), with sequences submitted for evaluation indicated. R^2 and Spearman's ρ are shown. **(d)** WashU (co-winner 28F & winner 47F) used a masked language model with a specialized prediction head, relying on the "biological plausibility" of the generative model to implicitly constrain developability without external filters.

https://doi.org/10.1038/s41587-026-03238-6



Extended Data Fig. 4 | Out-of-library CDR Design Challenge Overview. **(a)** ft-ESM residue embeddings are condensed via position-specific attention weights (Nseq x Lseq) to produce pooled vectors (Nseq x Hsize). These serve as input for a two-layer MLP (one-hidden layer; 1.64 M parameters) used to establish the Sequence-Alone Model objective function. This Sequence-Alone model evaluates novel candidates generated through two parallel approaches: deterministic CDR shuffling of top-ranked parents from clusters 27F and 28F, and stochastic MCMC/Gibbs sampling initialized from the single top-ranked sequence from the Challenge 2 model. High-scoring designs were subjected to *in silico* developability assessment, including MOE homology modeling and ft-ESM pseudo-perplexity, to identify the final competition shortlist. **(b – d)** Predicted fitness and provided experimental KD values for the test (**b**) and full dataset (**c**). R^2 and Spearman's ρ are shown. The winning Xencor submission, a 2.9 pM binder, is highlighted as an inverted orange triangle with a black border. **(d)** Violin plots showing the distribution of learned position-specific attention weights for CDR, Vernier, and other framework residues; horizontal lines within distributions indicate the median and interquartile range. These results illustrate the model's ability to prioritize regions known to influence paratope affinity, with significant enrichment observed in the CDR and Vernier regions relative to other framework residues with corresponding p-values as follows: framework vs. CDR (p = 1.59E-17), other framework vs. Vernier (p = 9.54E-07) and CDR vs. Vernier (p = 0.0949) (****p < 0.0001, n.s. = not significant).

nature portfolio

Corresponding author(s): M. Frank Erasmus; Andrew R.M. Bradbury

Last updated by author(s): May 12, 2026

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

HT-SPR kinetic data were acquired on a Carterra LSA-XT instrument (Carterra, Salt Lake City, USA) using the manufacturer's instrument-control software. KinExA equilibrium and screening measurements were acquired on a KinExA 4000 (Sapidyne Instruments, Boise, USA) housed in a TC1000 temperature-control chamber, using the manufacturer's acquisition software. Thermal stability (Tm) and aggregation (Tagg) data were acquired on an Unchained Labs Uncle by DSF (Tm) and static light scattering (Tagg). HIC-HPLC chromatograms were acquired on an HPLC system fitted with a Sepax Proteomix HIC Butyl-NP5 column (4.6 × 100 mm, 5 µm), with absorbance monitored at 280 nm. BVP ELISA and AC-SINS absorbance spectra (475–625 nm) were acquired on a Tecan microplate reader. NGS data were generated by Azenta Life Sciences (GENEWIZ). No custom software was used for data collection.

Data analysis

SPR sensorgrams were double-referenced (inter-spot reference; buffer-blank double reference), y-aligned in serial mode, and fitted to a 1:1 Langmuir model (with a mass-transport term Km added where appropriate) using the float T0 option on the Carterra platform; ka, kd and Rmax were globally fitted per capture location and means/SDs propagated for KD. KinExA equilibrium binding curves were globally fitted using Sapidyne KinExA Pro v4.7.6 (n-Curve module). Tm/Tagg curves were fitted with the Uncle Analysis software. AC-SINS λmax shifts were determined with the open-source AC-SINS Analysis R Shiny Application (Phan, S. et al.).

Participant computational methods used: Python 3.11, Biopython 1.83, NumPy, Matplotlib, AlphaFold3, Protenix, AntPack v0.3.7, scikit-learn v1.6.1, xGPR v0.4.7, resp_protein_toolkit, ft-ESM (ESM-2, 650M params), AutoGluon-Tabular, MOE 2024.0601, HuggingFace Trainer, and custom WashU architectures (AlignNet, DMS2). Per-team code repositories are listed in the "Code availability" subsections of the Methods. Figures were generated in Python with Matplotlib.

Code Availability: The codebases implementing the methods for each challenge presented in this work are publicly available at:
1. Challenge 1 (Aureka): https://github.com/AurekaBio/Aureka-AIntibody-Challenges

2. Challenge 2 – 27F, co-winner 1 (Scripps – Briney/Ward Lab): https://github.com/brineylab/ssrmsd
3. Challenge 2 – 27F, co-winner 2 (UCSD – Wang Lab): https://github.com/Wang-lab-UCSD/AntPack; https://github.com/jlparkI/xGPR); https://github.com/jlparkI/resp_protein_toolkit; https://github.com/Wang-lab-UCSD/aintibody_competition_code
4. Challenge 2 – 28F, co-winner 1 (Xencor): https://github.com/XenInc/Xencor-AIntibody-Challenges
5. Challenge 2 – 28F & Challenge 47F, co-winner 2 (WashU – Fremont Lab): https://github.com/kaszubat/AIntibody_Competition_WashU_Uploads
6. Challenge 3 (Xencor): https://github.com/XenInc/Xencor-AIntibody-Challenges

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

DATA AVAILABILITY: All data generated in this study are provided as supplementary data uploaded and include: (1) all SPR kinetic parameters (ka, kd, KD, Rmax) for every submission; (2) KinExA equilibrium and screening data for all tested antibodies; (3) individual developability assay scores (Tm, Tagg, HIC retention time, BVP score, AC-SINS delta-lambda) for all submissions; (4) processed and annotated NGS datasets used as input for all three challenges; (5) the complete sequence-to-result mapping table. All data in standard, machine-readable formats (e.g., Excel & CSV).

There are no restrictions on data availability. No human-subjects data, clinical data, or other access-restricted datasets were used. Participant code repositories are listed in the Software section and within each per-team "Code availability" subsection of the Methods.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	*Use the terms sex (biological attribute) and gender (shaped by social and cultural circumstances) carefully in order to avoid confusing both terms. Indicate if findings apply to only one sex or gender; describe whether sex and gender were considered in study design; whether sex and/or gender was determined based on self-reporting or assigned and methods used. Provide in the source data disaggregated sex and gender data, where this information has been collected, and if consent has been obtained for sharing of individual-level data; provide overall numbers in this Reporting Summary. Please state if this information has not been collected. Report sex- and gender-based analyses where performed, justify reasons for lack of sex- and gender-based analysis.*
Reporting on race, ethnicity, or other socially relevant groupings	*Please specify the socially constructed or socially relevant categorization variable(s) used in your manuscript and explain why they were used. Please note that such variables should not be used as proxies for other socially constructed/relevant variables (for example, race or ethnicity should not be used as a proxy for socioeconomic status). Provide clear definitions of the relevant terms used, how they were provided (by the participants/respondents, the researchers, or third parties), and the method(s) used to classify people into the different categories (e.g. self-report, census or administrative data, social media data, etc.) Please provide details about how you controlled for confounding variables in your analyses.*
Population characteristics	*Describe the covariate-relevant population characteristics of the human research participants (e.g. age, genotypic information, past and current diagnosis and treatment categories). If you filled out the behavioural & social sciences study design questions and have nothing to add here, write "See above."*
Recruitment	*Describe how participants were recruited. Outline any potential self-selection bias or other biases that may be present and how these are likely to impact results.*
Ethics oversight	*Identify the organization(s) that approved the study protocol.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	511 designed or predicted antibody variants were submitted across the three challenges by 29 participating organizations. Sample size reflects open, prospective submission volume and was not set by a power calculation. All 511 submissions were tested under uniform protocols by independent assay laboratories.
Data exclusions	No data were excluded from the analyses. All 511 submissions appear in every figure, table or supplementary data. Pre-established competition-rule violations (e.g. framework mutations in Challenge 2, where only CDR-resident variation was permitted) were flagged as "disqualified" (DQ) for winner determination only — they remain visible in all figures, tables or supplementary data (DQ marked in red in Fig. 3e and Fig. 5c). No data points were excluded post hoc.
Replication	Affinity was measured for every submission by two orthogonal methods in two independent labs: HT-SPR at Carterra (each mAb captured at three densities) and KinExA at Sapidyne (two-curve n-Curve global fit). Developability was measured at a third independent lab, Mosaic Biosciences, using a five-assay panel (Tm, Tagg, HIC-HPLC, BVP ELISA, AC-SINS) with Jain et al. clinical-stage IgG1 passing/questionable/failing controls in every batch. Cross-assay concordance was independently verified (Fig. 1i: ρ = 0.94 KinExA single-point vs HT-SPR; ρ = 0.88 standard KinExA vs HT-SPR). All replication attempts produced consistent results within the propagated uncertainty.
Randomization	Randomization of group allocation does not apply: this is a prospective benchmark where each organization self-selected which challenge(s) to enter and which sequences to submit. Within-assay sample order was set by the independent assay laboratories without reference to participant identity.
Blinding	The challenge was prospective and blinded: participants submitted sequences without prior knowledge of downstream affinity or developability outcomes, and the assay laboratories measured all submissions under uniform protocols, blinded to participant identity beyond a per-submission identifier. As stated in the Discussion, this first iteration relied on organizer integrity rather than informatic safeguards for blinding — a structural limitation being addressed in the next iteration.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	Capture / detection reagents: ThermoFisher Cat# 71303262100 (anti-human IgG-Fc nanobody, SPR); Jackson ImmunoResearch Cat# 109-005-098 (anti-human goat IgG1 Fc, AC-SINS); Jackson ImmunoResearch Cat# 005-000-003 (goat non-specific IgG, AC-SINS); Jackson ImmunoResearch Cat# 109-035-008 (HRP anti-human IgG, BVP ELISA). Test antibodies (n = 511): all designed or predicted antibodies submitted by the 29 participating organizations across Challenges 1, 2 and 3, plus the parental anti–SARS-CoV-2 RBD antibody, all produced as human IgG1 by Azenta Life Sciences (GENEWIZ). Developability controls: clinical-stage IgG1 antibodies from Jain et al. (PNAS 2017, ref. 9) — Blosozumab, Ponezumab, Bavituximab, Mepolizumab, Onartuzumab, Cixutumumab, Panitumumab, Fulranumab, Eculizumab — included as per-assay passing/questionable/failing controls (assay-specific assignments in Methods).
Validation	All 511 test antibodies were produced as human IgG1 by Azenta Life Sciences (GENEWIZ) under a standardized expression pipeline and assayed under uniform conditions by independent laboratories; affinity was confirmed by two orthogonal methods (HT-SPR and KinExA). Jain et al. clinical-stage IgG1 reference antibodies were used as developability controls in every batch, and the commercial capture/detection reagents (ThermoFisher 71303262100; Jackson ImmunoResearch 109-005-098, 005-000-003, 109-035-008) were used as supplied with manufacturer-validated specificity for human IgG/Fc.

Plants

Seed stocks	n/a
Novel plant genotypes	n/a
Authentication	n/a