Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Library size and structure, not synthetic labels, explain apparent benchmark--experiment discrepancies in antibody machine learning

Dohun Kim

dohun@etri.re.kr

Electronics and Telecommunications Research Institute

Minho Bae

Electronics and Telecommunications Research Institute

Article

Keywords: antibody engineering, machine learning benchmarks, protein language models, active learning, phage display, confounding

Posted Date: November 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10857079/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Library size and structure, not synthetic labels, explain apparent benchmark–experiment discrepancies in antibody machine learning

Dohun Kim[1*] and Minho Bae[1]

[1*]Autonomous Intelligence DX Research Section, Electronics and Telecommunications Research Institute, Seongnam, Gyeonggi-do, Republic of Korea.

*Corresponding author(s). E-mail(s): dohun@etri.re.kr;

Abstract

Synthetic benchmarks are widely used to develop machine-learning methods for antibody–antigen binding, but the conditions under which their conclusions transfer to measured data have been examined narrowly. Typical synthetic and measured datasets differ not only in whether labels are simulated, but also in library size, library structure, alignability and measurement noise. We show that these covarying factors, not the labels, account for the apparent differences. Holding a lattice oracle fixed and rebuilding its library as 6,000 mutational variants of one parent, we compared it against measured affinities for four therapeutic targets and a third-party campaign of 119,828 measurements. Before matching, the regimes disagreed about the search advantage, the preferred representation and architecture, and learnability decomposition. After matching, three of the four disagreements disappeared and the fourth proved unmeasurable: the search advantage fell from 1,620-fold to single digits, 3.3-fold against 3.0-fold measured, and a protein language model overtook one-hot encoding on the matched lattice. Surviving effects include a threefold search advantage, a Spearman cost of 0.105 from measurement noise, and an 8.1-fold overstatement by random splits. Synthetic benchmarks are faithful proxies when conditions are matched; the problem lies in uncontrolled comparison.

Keywords: antibody engineering, machine learning benchmarks, protein language models, active learning, phage display, confounding

Introduction

Machine learning is now routine in antibody discovery. The task is usually to rank candidate sequences for binding to a given antigen, so that only a small fraction of a library need be measured, and campaigns organised this way have delivered antibodies with improved affinity and specificity [2, 9]. Related advances in structure prediction and in protein language modelling have made sequence-only prediction credible at scales that were out of reach a few years ago [10–12]. Progress of this kind depends on being able to tell which method actually works, and that depends in turn on the benchmarks used to compare methods.

Experimental affinity data are expensive, so the field has come to rely on synthetic benchmarks. Absolut! [1] is the most widely used of these in the antibody setting: it computes lattice binding energies for arbitrary CDRH3 sequences against 159 PDB-derived antigens, providing ground truth at a scale no experiment can match, and it has been used to develop design and optimisation methods that are intended for real campaigns [7, 13]. Benchmarks for the wider protein-fitness problem take the opposite approach and assemble measured assays [14–16], which avoids the question of label realism but leaves the available data few, heterogeneous and mostly retrospective.

The value of a synthetic benchmark depends on whether methodological conclusions drawn from it hold for real measurements. The original Absolut! study addressed this directly [1] and reported that the relative accuracy ranking of machine learning methods transfers to experimental data, finding a Spearman correlation of 0.9 between accuracies on Absolut! and on the HER2 dataset of Mason et al. [2]. That evidence, however, was obtained for one binary classification task on a single experimental dataset using classical classifiers, and it is narrower than the uses the benchmark has since been put to, which include the choice of sequence representation, the choice of model architecture, and the design of iterative search strategies.

Comparisons between a synthetic benchmark and a measured dataset carry a structural risk that has received little attention in this field, although it is familiar elsewhere in machine learning for biology and chemistry. Ligand-based virtual screening benchmarks were shown to reward memorisation of dataset composition rather than generalisation [17], and to need explicit bias control before they can be interpreted [18]. Genomic applications are vulnerable to subtle leakage between training and evaluation data [19], and a survey across seventeen fields found reported performance inflated by leakage of exactly this kind [20]. Models readily exploit whatever incidental regularity separates the classes rather than the intended signal [21], which is why adversarial and trivial controls are recommended when such a model is evaluated [22], and why reporting standards for machine learning in biology ask that they be stated [23]. The common thread is that a benchmark comparison is interpretable only when the compared arms differ in the intended way and in no other.

The two arms usually compared in this field differ in several ways at once. A typical synthetic dataset is a large collection of unrelated CDRH3 sequences of varying length. A typical measured dataset is a mutational neighbourhood around a single parent antibody, produced by designing all single mutants and a sample of double and triple mutants [3, 4]. These two collections differ in at least five respects simultaneously: the origin of the labels, the size of the library, its structure, whether positions can

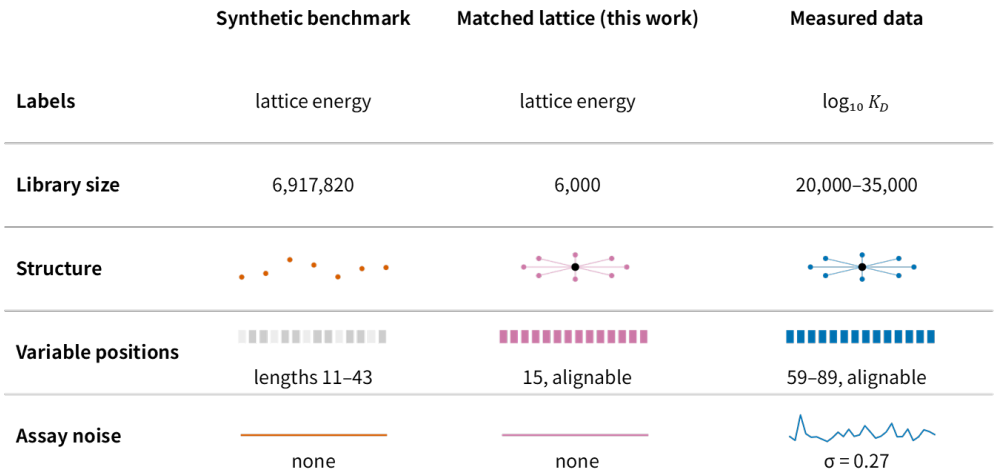


Fig. 1 The design of the comparison. A typical synthetic benchmark and a typical measured dataset differ on every row at once, so no difference between them can be attributed to the labels. The matched lattice built here keeps the synthetic label source and takes size, structure and alignability from the measured side, which is what makes the attribution testable. In the structure row each point is a library member and the black point is the parent; in the positions row a filled block is a position that can be aligned across members.

be aligned across members, and the presence of measurement noise. Any difference observed between them is therefore not attributable to the synthetic labels alone, and to our knowledge that attribution has not been tested.

Here we test it. We hold the oracle fixed, using the Absolut! binary to score sequences it has never seen, and rebuild the library so that it has the mutational structure and a comparable size to a measured dataset. This third arm shares its label source with the synthetic side and everything else with the measured side, so comparing it against both neighbours isolates what the synthetic labels are responsible for. We find that of four differences which appear when the two regimes are compared as they stand, three do not survive matching and the fourth cannot be measured once the arms are comparable. We then report the effects that do survive, and propose a short protocol that would make future comparisons interpretable.

Results

What differs between typical synthetic and measured datasets

Figure 1 summarises the comparison as it is normally made. On the synthetic side we used the Absolut! library of 6,917,820 murine CDRH3 sequences [1], which are drawn from observed repertoires [8], scored against six infectious-disease antigens (SARS-CoV spike RBD, influenza N9 neuraminidase, West Nile virus envelope, *Borrelia* OspA, *Plasmodium* AMA1 and HIV-1 p24). On the measured side we used AlphaSeq affinities for four targets (anti-SARS-CoV-2 VHH72, anti-TIGIT, anti-PD-1 and anti-HER2), comprising 707,713 measurements [3].

Table 1 The four differences that appear when the synthetic and measured regimes are compared as they stand, and what remains of each once library size and structure are matched. The lattice oracle is unchanged throughout; only the library differs. Win counts and intervals are given in Supplementary Table S2.

Apparent effect	Unmatched comparison	After matching size and structure	Verdict
Search advantage, active over random	1,620-fold	at a budget of 1,000, 3.3-fold on the matched lattice against 3.0-fold measured; single digits in both	Agrees
Representation, one-hot against ESM-2	one-hot ahead at every training size	the language model ahead at 3 of 5 training sizes, one-hot ahead at none	Reverses
Architecture, ridge against CNN	crossover seen only on the lattice	crossover at 2,000 labels in both regimes; ridge ahead below, the network ahead above	Agrees
Learnability split, noise against smoothness	54% / 46%	38% / 62%, with the two arms varying at different numbers of positions	Not measurable

The two collections differ in five ways at once. Labels are lattice energies versus $\log_{10} K_D$; library sizes are 6.9×10^6 versus 2–3.5×10^4; the synthetic library is a set of unrelated sequences whereas the measured libraries are mutational neighbourhoods of one parent; CDRH3 lengths vary from 11 to 43 residues on the synthetic side but positions are alignable on the measured side; and only the measured data carry assay noise, which we quantified as $\sigma = 0.27$ in $\log_{10} K_D$ from 100 replicate measurements of the wild-type parent.

Four differences appear when conditions are not matched

Comparing the two collections as they stand reproduces four differences (Table 1).

First, guided search appeared far more valuable on the synthetic side. With a budget of 2,000 measurements out of 6.9×10^6, active learning recovered 378 of the 691 sequences in the top 0.01%, against 0.23 for random sampling, a 1,620-fold ratio. On measured data the same procedure gave a threefold advantage.

Second, the preferred sequence representation differed. On the synthetic library a one-hot encoding of the sequence outperformed per-position ESM-2 embeddings [5] at every training-set size tested (5 of 5 intervals, mean difference in Spearman +0.075), whereas on measured targets the language model won 16 of 19 intervals.

Third, the preferred architecture appeared to differ. Ridge regression on one-hot features beat a max-pooling convolutional network across the whole range we had examined on measured data, which stopped at 5,000 training labels, whereas on the synthetic library the convolutional network overtook ridge within that range.

Fourth, a decomposition of the learnability gap attributed 54% to measurement noise and 46% to a residual we initially interpreted as intrinsic smoothness of the lattice energy function.

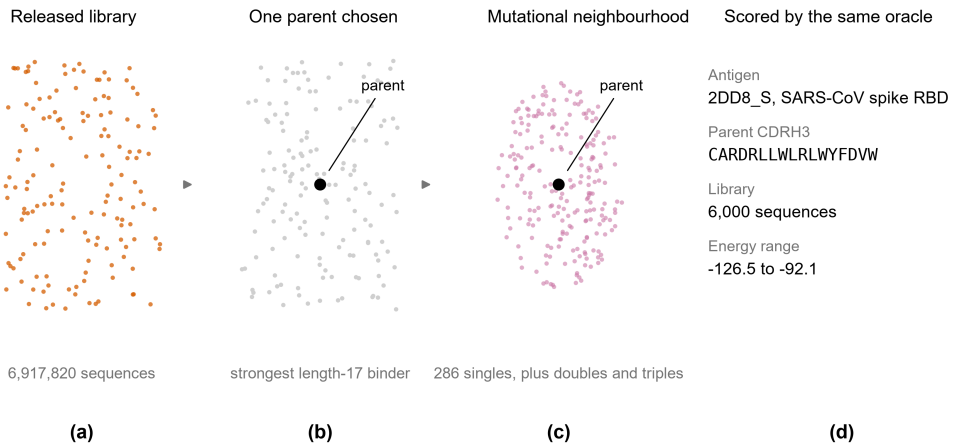


Fig. 2 Construction of the matched library. **(a)**, the released library of unrelated CDRH3 sequences. **(b)**, one parent is chosen, the strongest binder of length 17. **(c)**, the parent is perturbed the way a wet-lab library is perturbed, giving a mutational neighbourhood rather than a sample. **(d)**, the two libraries built this way. The resulting collection has the size and the mutational structure of a measured dataset while its labels still come from the lattice oracle. The Absolut! binary scores every sequence, so the oracle, the antigen and the scoring function are identical to those used for the released library. The matched and released libraries dock into the same pocket of the antigen, so they differ in sequence space rather than in binding site. A second library was built for influenza N9 neuraminidase but is withheld: its parent disagreed with the oracle and it could not be re-scored.

Three of the four disappear, and the fourth proves unmeasurable

We then rebuilt the synthetic library to match the structure of the measured ones (Fig. 2). We selected a strong binder of length 17 as the parent (`CARDRLLWLRLWYFDVW` for SARS-CoV spike RBD, energy −121.84), enumerated all 286 single mutants, and added random double and triple mutants to reach 6,000 sequences, the design recipe used to build the AlphaSeq libraries. A second library, for influenza N9 neuraminidase, did not pass the check described in Methods and is not used. Every sequence was scored with the Absolut! binary, so the labels remain lattice energies from the same oracle. The only variables changed are library size and structure.

Three of the four differences resolved and the fourth turned out not to be measurable (Table 1).

The search advantage collapsed by two orders of magnitude. On the matched lattice at a budget of 1,000 it was 3.3-fold, against 3.0-fold on measured libraries subsampled to the same size (Fig. 3); absolute recall levels still differ (0.66 against 0.43 at that budget), but the ratio over random sampling, which is the quantity method papers report, agrees. The two regimes are not identical at every budget: at 500 labels the matched lattice retains a larger advantage (6.3-fold against 3.0-fold), and at 200 neither strategy beat random sampling in either regime. What survives matching is therefore the order of magnitude, single digits rather than three, and not an exact correspondence.

The representation preference reversed on the lattice itself. Over twenty independent splits, one-hot held no significant advantage at any training-set size on the

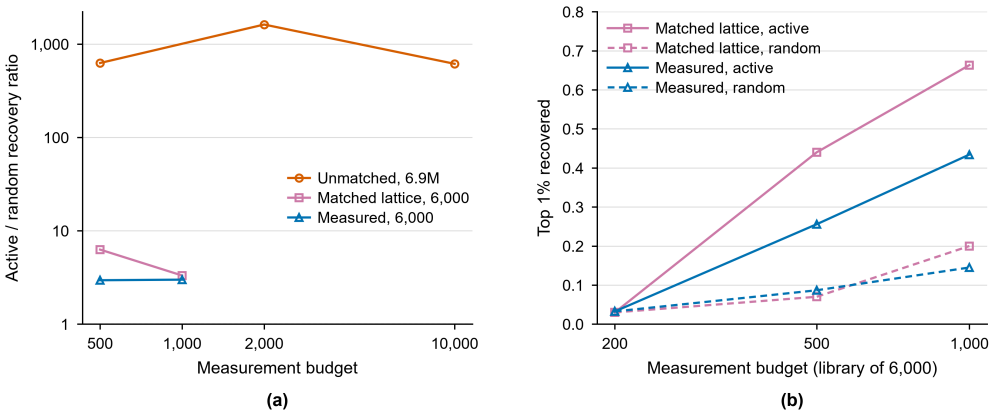


Fig. 3 The magnitude of the search advantage is set by library size. **(a)**, the ratio of top-fraction recovery for guided over random search, before and after matching; the ordinate is logarithmic. **(b)**, recovery curves in the matched setting. Absolute levels differ between the two regimes while the advantage over random sampling agrees.

matched library, while the language model was ahead at three of the five sizes after Holm correction (Fig. 4; Supplementary Table S2). Pooled over splits and sizes the mean Spearman difference was −0.013 (95% CI −0.016 to −0.009), and the language model was ahead in 80 of the 100 individual splits. The advantage is largest when labels are scarce and narrows as they accumulate: −0.037 at 200 labels, −0.018 at 500, −0.007 at 1,000, and indistinguishable from zero by 2,000 (Holm-adjusted $p = 0.17$). Since the oracle, the scoring function and the model code were unchanged, the reversal follows the alignability of the library rather than the synthetic labels. Because it rests on a single matched antigen, we report it as a demonstration that the earlier attribution does not hold, rather than as a general result.

The architecture crossover was not a difference between regimes but an artefact of the range examined. Extending the measured arm from 5,000 to 20,000 training labels using the AlphaBind fine-tuning splits [4], over twelve splits for each of three targets, located the crossover at 2,000 labels. Ridge regression was ahead below that point (mean Spearman difference +0.026, +0.031 and +0.015 at 200, 500 and 1,000 labels) and the convolutional network above it (−0.018, −0.026 and −0.027 at 5,000, 10,000 and 20,000), with every one of those six differences surviving Holm correction; 2,000 labels was the only size at which the two were indistinguishable (−0.002, Holm-adjusted $p = 0.51$). The same ordering, ridge first and the convolutional network later, holds on the synthetic library, so what had looked like a disagreement about architecture was a disagreement about where each study had stopped looking.

The learnability decomposition proved not to be measurable with these data. Repeating it on the matched library changed the split from 54/46 to 38/62, and the matched lattice library varies at 15 positions against 59–89 for the measured single-chain variable fragments, so the two arms remain dimensionally unmatched. We therefore report this decomposition as undetermined rather than quoting either value.

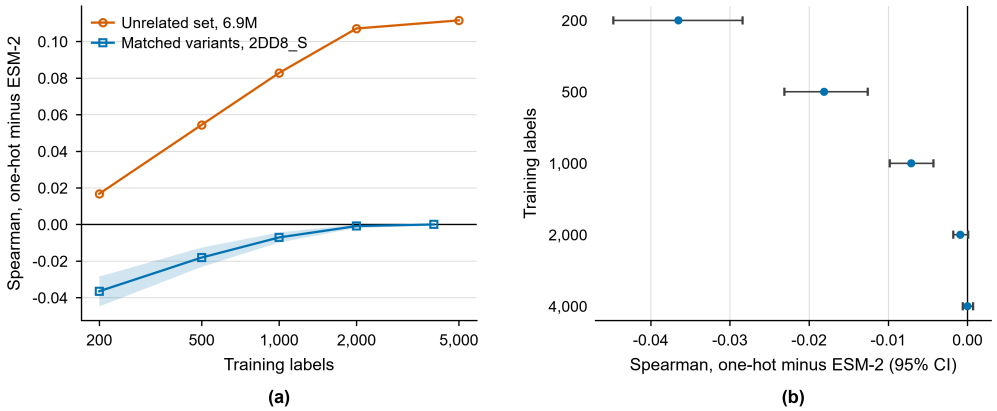


Fig. 4 Representation preference tracks library structure rather than label origin. **(a)**, the difference in Spearman correlation between a one-hot encoding and ESM-2 embeddings; positive values favour one-hot. The same Absolut! oracle scores both libraries, so only the library composition changes, and the preference reverses. Bands are percentile bootstrap intervals over twenty independent splits. **(b)**, the same effect per training size with 95% intervals. The advantage of the language model narrows as labels accumulate and is indistinguishable from zero beyond 2,000 labels.

Effects that survive matching

Several findings were unaffected.

Guided search retains a single-digit advantage. At matched size and structure, active learning recovered 3.0 times as much of the top 1% as random sampling across four measured targets, and 3.3 times on the matched lattice, at a budget of 1,000. At a budget of 200 out of 6,000 neither strategy improved on random in either regime, indicating a floor below which the surrogate carries no information.

Measurement noise alone costs 0.105 in Spearman. Adding Gaussian noise matched to the measured noise-to-signal ratio (0.424) to the matched lattice labels reduced Spearman by 0.095 to 0.116 across training-set sizes. Because this comparison is made within one dataset with structure held fixed, it is not subject to the confound above.

Random splits overstate prospective performance 8.1-fold (Fig. 5). On the AlphaSeq coronavirus panel, a model evaluated by random split reached Spearman 0.695 and recovered 40.3% of the true top 100 within its top 500. The same model evaluated prospectively, trained on one design campaign and tested on sequences designed later, reached 0.413 and 5.0%. Batch effects do not explain this: wild-type replicates measured in both campaigns shifted by +0.037 on average with inconsistent sign, and Spearman is invariant to monotone shifts. The cause is visible in the sequences themselves. The first campaign stayed close to the parent antibody, at a median of 4 mutations with a 90th percentile of 5, whereas the second explored much further, at a median of 9 and a 90th percentile of 32 (Fig. 5c). The two campaigns separate along the first principal component of the encoded sequences (Kolmogorov–Smirnov $D = 0.75$), so a model fitted to the first is asked to extrapolate rather than interpolate. Random splits never pose that question.

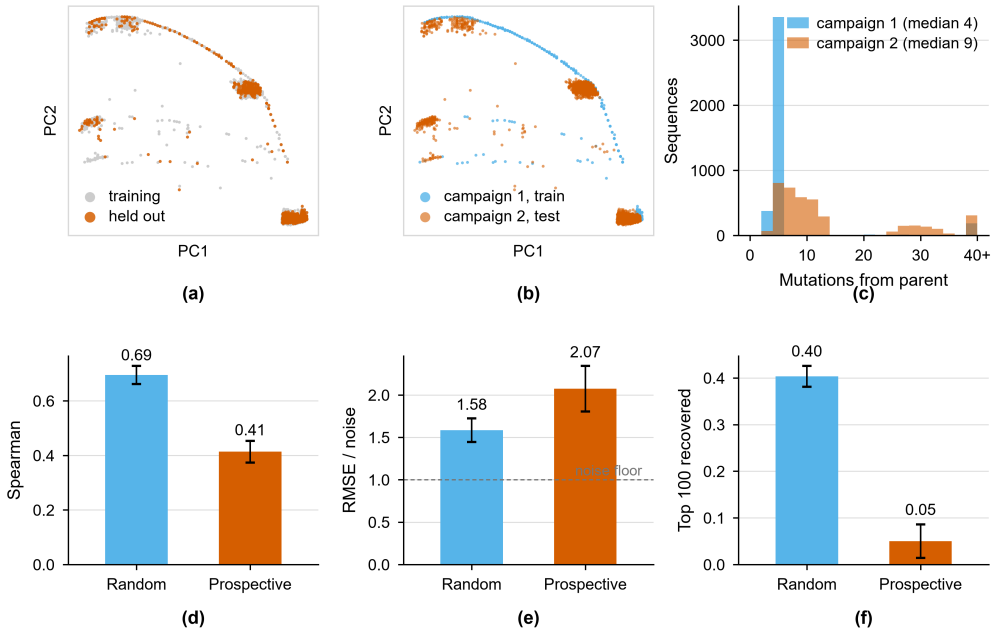


Fig. 5 Random splits overstate prospective performance, and why. **(a)**, under a random split the held-out sequences lie among the training sequences. **(b)**, under a prospective split the test set is a displaced cloud; coordinates are the first two principal components of the one-hot encoded sequences. **(c)**, the source of the displacement: the second campaign designed sequences far further from the parent antibody than the first. **(d)**–**(f)**, the cost. The same model recovers 8.1-fold less of the true top 100 when evaluated prospectively. The measurement noise floor was estimated from 100 wild-type replicates.

Read-count models do not beat a trivial baseline on panning data (Supplementary Fig. S1). Re-analysing a five-round c-Met panning campaign (BioProject PRJNA607865) [6], a random forest predicting final-round abundance from early rounds achieved AUROC 0.992, while simply ranking by the last available round's read count achieved 0.997. Sequence alone, in contrast, identified the top 1% of enriched clonotypes at AUROC 0.811 against a permutation null of 0.482 ± 0.056 ($z = 5.9$, $p < 0.001$), indicating that the value of learning lies in scoring sequences that have not been measured rather than in re-reading counts.

Combined representations rank highest among seven models. The four measured targets do not all extend to the same training-set size (HER2 ends at 2,000 labels and PD-1 at 10,000), so we compare models at 2,000, the largest size for which all four are present, rather than averaging over a set that thins out as the curve advances. There, ridge on ESM-2 embeddings combined with one-hot features reached Spearman 0.648 (95% CI 0.584–0.716 by bootstrap over targets), ahead of ESM alone (0.639), one-hot alone (0.626), the convolutional network (0.576), random forest (0.533), boosted trees (0.526) and a k-nearest-neighbour control on Hamming distance (0.328) (Fig. 6). The ordering of the top three is unchanged at 5,000 and 10,000 labels on the three targets that reach them. The clear separation of the trivial control confirms that the models are learning rather than exploiting dataset structure.

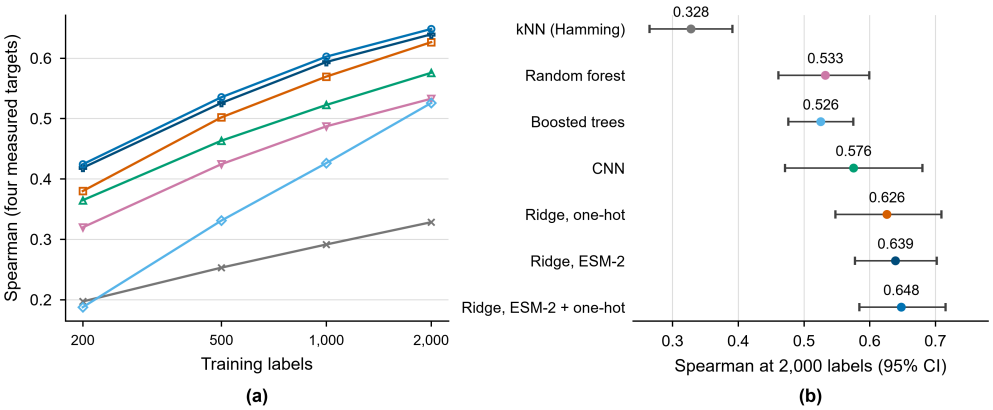


Fig. 6 Seven models across the four measured targets. Both panels are restricted to training-set sizes at which all four targets are present, since the available sizes differ between targets and averaging over the full range would mix a trend with a change of composition. **(a)**, learning curves, averaged over targets; each model keeps the colour it is given in **(b)**, which names them, so no key is repeated here. **(b)**, performance at 2,000 training labels with 95% intervals from a bootstrap that resamples targets rather than predictions, since nested subsets of one dataset are not independent observations. The nearest-neighbour control separates clearly from the remaining models, which is the minimal evidence that the comparison is not dominated by dataset structure.

An independent prospective campaign agrees

A-Alpha Bio released the candidates proposed during the AlphaBind campaign together with the AlphaSeq measurement of each [4]. Because the generators differ only in featurisation (one-hot against ESM-2nv embeddings), these 119,828 measurements constitute a prospective, wet-lab test of the representation question by an independent group.

The language model produced better candidates for all three parental antibodies (Fig. 7): 76% against 47% of candidates improved on the parent for anti-TIGIT, 92% against 52% for pembrolizumab, and 47% against 28% for VHH72, with median improvements differing by -0.275, -1.374 and -0.466 in $\log_{10} K_D$ ($p < 10^{-249}$ in all six comparisons). These candidates are mutational libraries, so the direction agrees with our matched-library result.

We note that the correlation between each model's own predicted affinity and the subsequent measurement is strongly attenuated in this dataset, because candidates were selected to have similar predicted values; predicted values have 0.16 times the standard deviation of the measurements. We therefore do not interpret that correlation.

Discussion

None of the four differences we examined was a property of the synthetic labels. Three were removed by changing library size and structure while holding the oracle fixed, and the fourth turned out not to be measurable once the two arms were made comparable. Read positively, this is evidence that Absolut! is a useful proxy: at matched size and

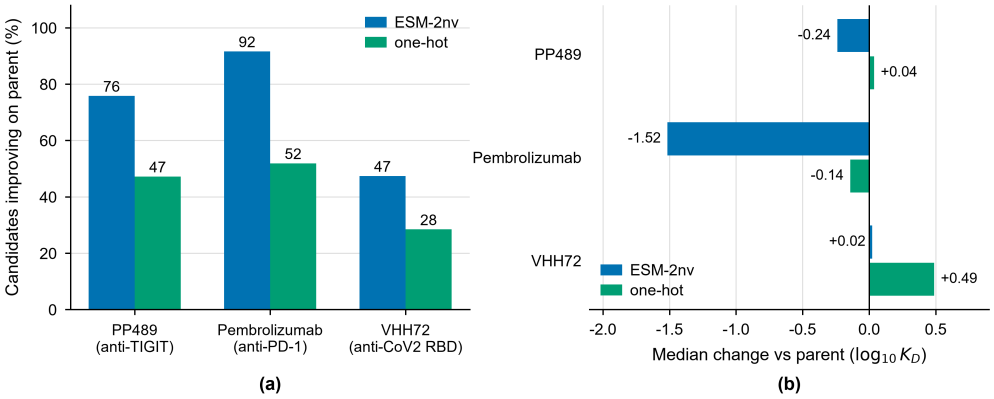


Fig. 7 An independent prospective campaign agrees. Candidates from generators differing only in featurisation were all measured by AlphaSeq. **(a)**, the fraction of candidates that improved on the parent antibody. **(b)**, the median change in affinity, where negative values are improvements. The language model is ahead on both measures for all three parental antibodies.

structure it gave the same answers as measured data for the search advantage, the representation choice and the architecture crossover. The difficulty lies in comparisons that vary several factors at once.

This has a practical consequence for how synthetic benchmarks are used. Before selecting a method on a synthetic benchmark, we suggest matching five properties to the intended application: library size; library structure (mutational neighbourhood or unrelated sequences); the number of varying positions; measurement noise, injected at the noise-to-signal ratio of the intended assay; and the range of training-set sizes, checked to confirm that the crossover lies inside it. Reporting these five alongside benchmark results would make cross-study comparison interpretable.

The effects that survived matching are worth separating from those that did not, because they are the ones a practitioner can rely on. Guided search retained a threefold advantage over random sampling in both regimes, which is modest next to the figure a large library suggests but is stable across libraries, oracles and four measured targets. Measurement noise alone cost 0.105 in Spearman when injected into otherwise identical data, a comparison made within one dataset and therefore not subject to the confound examined here; this sets a floor on what any model can achieve on a noisy assay and argues for reporting error relative to assay noise rather than in absolute units. A combined representation, concatenating language-model embeddings with one-hot features, ranked highest across the four measured targets, and the nearest-neighbour control separated clearly from every model, which is the minimal evidence that a comparison is measuring learning rather than dataset composition [22].

Two of the surviving results concern evaluation protocol rather than models, and both point the same way. Random splits overstated prospective recovery 8.1-fold, and the reason is visible in the sequences: the second design campaign explored much further from the parent antibody than the first, so a model fitted to the first is asked to extrapolate. Retrospective accuracy on a mutational neighbourhood therefore says little about performance on the next round of designs, a gap that iterative and

uncertainty-aware methods are built to address [27, 28]. Similarly, on panning data a random forest predicting final-round abundance from early rounds did not improve on ranking by the last available round. Read counts largely predict read counts. The value of learning appeared instead where sequence alone identified enriched clonotypes, which is the setting in which a model can score sequences that were never measured.

Our findings are consistent with the transferability originally reported for Absolut! [1] rather than in tension with it. That study compared classical classifiers on a binary task and found their ranking preserved; we compared representations and architectures on continuous affinity ranking and found agreement once conditions were matched. The two results delimit the same conclusion from different directions. They also sit alongside a broader finding in protein engineering, that the design of the training set (how far it spreads from the parent, and how it is sampled) often matters more than the choice of model [24–26]. What we add is that this applies to the benchmark as much as to the campaign: the library a method is tested on shapes the conclusion as strongly as the method does.

Several limitations bound what we can claim. We performed no wet-lab experiments; all measured data are third-party, and the prospective evidence we cite was generated by other groups for their own purposes. The matched library comprises 6,000 sequences for a single antigen and varies at 15 positions against 59–89 in the measured single-chain variable fragments, so the dimensional mismatch is reduced but not eliminated, and one antigen cannot speak for the 159 in the Absolut! collection. A second matched library was built and then withheld when it failed the consistency check described in Methods, so the reversal of representation preference rests on one antigen and should be read as a demonstration that the earlier attribution fails rather than as a general claim about matched lattices. Our confidence intervals resample train/test splits, which captures variability due to the split and to model initialisation but not variability due to which sequences the library happens to contain; the latter would require independently constructed libraries. We could not test on measured data whether enlarging a library at fixed budget continues to yield better binders: across the 15-fold range available (2,000 to 30,000), only one of four targets improved monotonically, and the range is far shorter than the 692-fold sweep possible on the synthetic library. Finally, the lattice oracle is a coarse physical model, and our results speak to whether conclusions transfer between library designs, not to whether lattice energies are an accurate account of binding.

Two directions follow. The first is to repeat the matched construction across many more antigens, which is inexpensive because the oracle is a binary that can be run on any sequence, and would establish how far the agreement we observe generalises. The second is prospective: the protocol above makes predictions about which methods should transfer, and those predictions are testable in a campaign that designs, measures and reports a held-out round. Neither requires a new model, which we take as the useful implication of this work: a substantial part of the apparent disagreement between synthetic and measured evidence is a property of how the comparison was set up, and is therefore fixable by design rather than by better methods.

Methods

Data

Absolut! binding energies for six antigens (2DD8_S, 1NCA_N, 1ZTX_E, 1OSP_O, 2Q8A_A, 3VRL_C) were obtained from the project archive and reduced to one energy per CDRH3 by retaining the best-scoring 11-mer slice, giving 6,917,820 sequences common to all six. Measured affinities, obtained by the yeast-mating AlphaSeq assay [29], were taken from the open-alphaseq release [3]; we retained pairs with a reported affinity, sufficient replicate observations and above-background binding, and averaged replicates. The AlphaBind campaign data and its fine-tuning and validation splits were obtained from the accompanying repository [4]. Panning reads for BioProject PRJNA607865 were downloaded from the European Nucleotide Archive; CDR3 sequences were extracted by translating both mates in three frames and matching the conserved framework anchors, recovering CDR3 from 86–92% of reads. One run (VL round 1) was excluded because only 5.1% of its reads carried a recognisable CDR3.

Matched libraries

The Absolut! source was compiled from the public repository. On Windows the binary linked but crashed while processing structure files; all scoring was therefore performed under Linux. Precomputed structures for each antigen were downloaded and the interaction profiles regenerated locally. Matched libraries were built by selecting a parent CDRH3 of length 17 with the lowest energy for that antigen, enumerating all 286 single mutants at non-terminal positions, and adding random double (60%) and triple (40%) mutants until 6,000 unique sequences were reached. Sequences were scored in `repertoire` mode and the best slice per sequence retained.

Each library was then checked against the released one. The parent appears in both, and because the same binary, antigen and scoring function produced them, its energy must agree. For SARS-CoV spike RBD it agreed exactly, as did every other sequence common to the two. A library built for influenza N9 neuraminidase failed this check, its parent differing by 13.15 energy units, and re-scoring did not complete; it is withheld and is reported nowhere in this article. The matched-library analysis therefore uses one antigen.

Models and features

Sequences were aligned to the parent and one-hot encoded at positions that vary within the library. Seven models were compared: ridge regression on one-hot features; histogram gradient boosting and random forest on ordinal tokens; k-nearest neighbours ($k = 10$) on Hamming distance as a triviality control; a convolutional network with an embedding layer, two convolutional layers and global max pooling, whose pooling mirrors the best-slice rule of the lattice oracle; ridge on per-position ESM-2 (35M) embeddings reduced to 48 principal components; and ridge on the concatenation of the last two. Language-model embeddings were mean-pooled over the whole sequence in an initial analysis and taken per varying position thereafter, as whole-sequence pooling averages away the few positions that distinguish members of a mutational library.

Ridge regularisation was set to $\alpha = 1$ for one-hot features and $\alpha = 50$ for the higher-dimensional embedding features, values chosen once on a held-out fold of a single target and then held fixed for every experiment reported here, so that no comparison benefits from per-arm tuning. The convolutional network used a 32-dimensional embedding, 96 channels, and a fixed budget of approximately 4,000 gradient steps at batch size 128, with targets standardised to zero mean and unit variance; this budget rather than a fixed epoch count keeps optimisation effort comparable across training-set sizes. Boosted trees and random forests used the library defaults with 200 estimators and a minimum leaf size of two.

Evaluation

Training-set sizes of 200 to 20,000 were drawn from a held-out permutation with 1,500–5,000 sequences reserved for testing, repeated over three to five seeds. We report Spearman rank correlation and the fraction of the true top 100 recovered within a model's top 500. Confidence intervals were obtained by resampling targets rather than individual predictions, because nested subsets of one dataset are not independent observations; intervals are percentile bootstrap over 10,000 resamples.

Search experiments

Search proceeded in two stages: a sparse ridge model scored the entire library, and the convolutional network reranked its top 200,000. Three strategies were compared at equal budget: uniform random sampling; a single-shot strategy that spends an initial quarter of the budget at random, fits once and selects the remainder; and active learning, which spends the remaining budget over three rounds with 20% held for exploration. We report recovery of the top 0.01% or top 1% depending on library size, the best value found, and the gap to the library optimum.

Noise injection

Measurement noise was estimated from 100 replicate AlphaSeq measurements of the wild-type parent in each campaign ($\sigma = 0.27$ in $\log_{10} K_D$), giving a mean noise-to-signal ratio of 0.424 across the four targets. Gaussian noise at that ratio of the label standard deviation was added to matched lattice labels.

Prospective evaluation

The AlphaSeq coronavirus panel contains two design campaigns. The retrospective task splits one campaign at random; the prospective task trains on the first campaign and tests on the second, after removing from the test set every sequence that also appears in the first, so that re-measured designs cannot inflate the result. To characterise the shift between campaigns we encoded both sets one-hot against the most frequent sequence and took the first two components of a truncated singular value decomposition, and separately counted the number of substitutions of each sequence from that reference. Batch effects were assessed using the wild-type replicates present in both campaigns.

Panning re-analysis

Clonotype abundances were computed per round after rarefying every library to the smallest sequencing depth in the series and discarding clonotypes seen fewer than five times, since diversity statistics and enrichment ratios are both sensitive to depth. Three tasks were compared for identifying the top 1% of final-round clonotypes: a random forest on early-round read counts; a baseline that ranks by the last available early round alone; and a model given only the CDR3 sequence, with no abundance information. A permutation null was obtained by shuffling labels over twenty repetitions.

Software and reproducibility

Analyses were run in Python with scikit-learn, PyTorch, pandas and SciPy; ESM-2 embeddings were computed with the released 35M-parameter checkpoint. The Absolut! binary was built from its public repository under Linux. Every quantity reported in this article is regenerated by a single script from the result tables provided as Supplementary Data, and Supplementary Table S1 maps each number to the table and computation that produced it.

Statistics

The two comparisons that the text states as directional claims (one-hot against the language model on the matched library, and ridge against the convolutional network on measured targets) were each repeated over twenty and twelve independent train/test splits respectively, rather than the three to five used for the exploratory sweeps. At each training size the paired difference in Spearman correlation was tested against zero by a one-sample t-test across splits, with the sign confirmed by a Wilcoxon signed-rank test, and summarised by a percentile bootstrap over 10,000 resamples of the splits.

Because each comparison is swept over training sizes, the sweep forms a family of tests and the smallest p-value in it would overstate the evidence if quoted alone. Each family was therefore corrected by the Holm–Bonferroni procedure [30], controlling the family-wise error rate, and separately by the Benjamini–Hochberg procedure [31], controlling the false discovery rate; both adjusted values are reported in Supplementary Table S2. Families were corrected separately rather than pooled, since they address distinct questions and pooling would penalise each for the breadth of the others.

Splits are independent partitions of one library, so these intervals describe variability due to the train/test split and to model initialisation. They do not describe variability due to which sequences the library contains, which would require independently constructed libraries; the matched-library results therefore rest on a single antigen and should be read as such.

Supplementary information. Supplementary Table S1 lists every quantity reported in this article together with the source file and the computation that produced it. Supplementary Table S2 gives the raw, Holm–Bonferroni-adjusted and Benjamini–Hochberg-adjusted p-values for every family of tests, with the effect size beside each.

Supplementary Figure S1 reports the re-analysis of the panning campaign. Supplementary Data comprise the matched library generated here and the per-experiment result tables underlying every figure.

Acknowledgements. This work was supported by the Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2024 (RS-2024-00398310). This work was supported by internal fund/grant of Electronics and Telecommunications Research Institute (25YT1400). We thank the authors of Absolut!, open-alphaseq and AlphaBind for releasing the data and code that made this re-analysis possible.

Declarations

- **Funding.** This work was supported by the Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2024 (RS-2024-00398310). This work was supported by internal fund/grant of Electronics and Telecommunications Research Institute (25YT1400).
- **Competing interests.** The authors declare no competing interests.
- **Author contributions.** D.K. conceived and designed the study, performed the analyses, wrote the manuscript, administered the project and acquired funding. M.B. contributed to the methodology, to the validation of the results and to data curation, and revised the manuscript. Both authors read and approved the final manuscript.
- **Data availability.** Every primary dataset analysed here is third-party and already public: Absolut! lattice binding energies for the six antigens (https://greifflab.org/absolut/); open-alphaseq AlphaSeq affinity measurements (https://huggingface.co/datasets/aalphabio/open-alphaseq); the AlphaBind campaign measurements (https://github.com/A-Alpha-Bio/alphabind); the c-Met panning reads deposited under BioProject PRJNA607865. The data generated in this work are provided in full as Supplementary Data: the matched library (6,000 CDRH3 sequences for 2DD8_S with their Absolut! energies) and the per-experiment result tables from which every figure and every quantity in the text is computed. The second matched library, for 1NCA_N, is withheld because it failed the consistency check described in Methods.
- **Code availability.** Custom code implementing the matched-library construction, the models, the search experiments and the noise injection is available from the corresponding author on reasonable request.
- **Ethics approval.** Not applicable.

References

[1] Robert, P.A., Akbar, R., Frank, R., Pavlović, M., Widrich, M., Snapkov, I., et al.: Unconstrained generation of synthetic antibody–antigen structures to guide machine learning methodology for antibody specificity prediction. Nature Computational Science **2**, 845–865 (2022)

[2] Mason, D.M., Friedensohn, S., Weber, C.R., Jordi, C., Wagner, B., Meng, S.M., et al.: Optimization of therapeutic antibodies by predicting antigen specificity from antibody sequence via deep learning. Nature Biomedical Engineering **5**, 600–612 (2021)

[3] Engelhart, E., Emerson, R., Shing, L., Lennartz, C., Guion, D., Kelley, M., et al.: A dataset comprised of binding interactions for 104,972 antibodies against a SARS-CoV-2 peptide. Scientific Data **9**, 653 (2022)

[4] Agarwal, A.A., Harrang, J., Noble, D., McGowan, K.L., Lange, A.W., Engelhart, E., et al.: AlphaBind, a domain-specific model to predict and optimize antibody–antigen binding affinity. mAbs **17**, 2534626 (2025)

[5] Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., et al.: Evolutionary-scale prediction of atomic-level protein structure with a language model. Science **379**, 1123–1130 (2023)

[6] Yoo, D.K., Lee, S.R., Jung, Y., Han, H., Lee, H.K., Han, J., et al.: Machine learning-guided prediction of antigen-reactive in silico clonotypes based on changes in clonal abundance through bio-panning. Biomolecules **10**, 421 (2020)

[7] Khan, A., Cowen-Rivers, A.I., Grosnit, A., Deik, D.G., Robert, P.A., Greiff, V., et al.: Toward real-world automated antibody design with combinatorial Bayesian optimization. Cell Reports Methods **3**, 100374 (2023)

[8] 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 Science **31**, 141–146 (2022)

[9] Hie, B.L., Shanker, V.R., Xu, D., Bruun, T.U.J., Weidenbacher, P.A., Tang, S., et al.: Efficient evolution of human antibodies from general protein language models. Nature Biotechnology **42**, 275–283 (2024)

[10] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al.: Highly accurate protein structure prediction with AlphaFold. Nature **596**, 583–589 (2021)

[11] Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., et al.: Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proceedings of the National Academy of Sciences **118**, e2016239118 (2021)

[12] Ruffolo, J.A., Chu, L.-S., Mahajan, S.P., Gray, J.J.: Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies. Nature Communications **14**, 2389 (2023)

[13] Akbar, R., Robert, P.A., Weber, C.R., Widrich, M., Frank, R., Pavlović, M., et al.: In silico proof of principle of machine learning-based antibody design at unconstrained scale. mAbs **14**, 2031482 (2022)

[14] Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, X., Canny, J., et al.: Evaluating protein transfer learning with TAPE. In: Advances in Neural Information Processing Systems, vol. 32 (2019)

[15] Dallago, C., Mou, J., Johnston, K.E., Wittmann, B.J., Bhattacharya, N., Goldman, S., et al.: FLIP: benchmark tasks in fitness landscape inference for proteins. In: Advances in Neural Information Processing Systems, Datasets and Benchmarks Track (2021)

[16] Notin, P., Kollasch, A.W., Ritter, D., van Niekerk, L., Paul, S., Spinner, H., et al.: ProteinGym: large-scale benchmarks for protein fitness prediction and design. In: Advances in Neural Information Processing Systems, vol. 36, pp. 64331–64379 (2023)

[17] Wallach, I., Heifets, A.: Most ligand-based classification benchmarks reward memorization rather than generalization. Journal of Chemical Information and Modeling **58**, 916–932 (2018)

[18] Sieg, J., Flachsenberg, F., Rarey, M.: In need of bias control: evaluating chemical data for machine learning in structure-based virtual screening. Journal of Chemical Information and Modeling **59**, 947–961 (2019)

[19] Whalen, S., Schreiber, J., Noble, W.S., Pollard, K.S.: Navigating the pitfalls of applying machine learning in genomics. Nature Reviews Genetics **23**, 169–181 (2022)

[20] Kapoor, S., Narayanan, A.: Leakage and the reproducibility crisis in machine-learning-based science. Patterns **4**, 100804 (2023)

[21] Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Machine Intelligence **2**, 665–673 (2020)

[22] Chuang, K.V., Keiser, M.J.: Adversarial controls for scientific machine learning. ACS Chemical Biology **13**, 2819–2821 (2018)

[23] Walsh, I., Fishman, D., Garcia-Gasulla, D., Titma, T., Pollastri, G., et al.: DOME: recommendations for supervised machine learning validation in biology. Nature Methods **18**, 1122–1127 (2021)

[24] Yang, K.K., Wu, Z., Arnold, F.H.: Machine-learning-guided directed evolution for protein engineering. Nature Methods **16**, 687–694 (2019)

[25] Biswas, S., Khimulya, G., Alley, E.C., Esvelt, K.M., Church, G.M.: Low-N protein engineering with data-efficient deep learning. Nature Methods **18**, 389–396 (2021)

[26] Wittmann, B.J., Yue, Y., Arnold, F.H.: Informed training set design enables efficient machine learning-assisted directed evolution. Cell Systems **12**, 1026–1045 (2021)

[27] Hie, B., Bryson, B.D., Berger, B.: Leveraging uncertainty in machine learning accelerates biological discovery and design. Cell Systems **11**, 461–477 (2020)

[28] Settles, B.: Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison (2009)

[29] Younger, D., Berger, S., Baker, D., Klavins, E.: High-throughput characterization of protein–protein interactions by reprogramming yeast mating. Proceedings of the National Academy of Sciences **114**, 12166–12171 (2017)

[30] Holm, S.: A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics **6**, 65–70 (1979)

[31] Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B **57**, 289–300 (1995)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- 2supplementaryinformation.zip