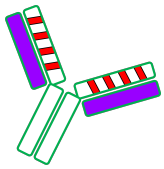
Baselining the Buzz

Trastuzumab-HER2 Affinity, and Beyond

Lewis Chinery[1,†], Alissa M. Hummer[1,†], Brij Bhushan Mehta[2,†], Rahmad Akbar[2], Puneet Rawat[2], Andrei Slabodkin[2], Khang Le Quy[2], Fridtjof Lund-Johansen[2], Victor Greiff[2], Jeliazko R. Jeliazkov[3], and Charlotte M. Deane[1,*]

[1] *University of Oxford,* [2] *University of Oslo,* [3] *Protein Design and Informatics, Computational Sciences, GSK R&D,*
[†] *Equal contribution,* [*] *To whom correspondence should be addressed.*

March 22, 2024



1 Abstract

There is currently considerable interest in the field of *de novo* antibody design, and deep learning techniques are now regularly applied to optimise antibody properties such as binding affinity. However, robust baselines within this field have not kept up with recent developments.

In this study, we generate a dataset of over 524,000 Trastuzumab variants and use this to show that standard computational methods such as BLOSUM, AbLang, ESM, and ProteinMPNN can be used to design diverse antibody libraries from just a single starting sequence. These novel libraries are predicted to be enriched in binding variants and experimental validation of 700 of these designs is ongoing. We also demonstrate that, even with only a very small number of experimental data points, simple machine learning classifiers can be trained in seconds to accurately pre-screen future designs. This pre-screening maintains library diversity and saves experimental time and money.

Availability: Source code is freely available at github.com/oxpig/Tz_her2_affinity_and_beyond. Data can be found at doi.org/10.5281/zenodo.10549114

Contact: deane@stats.ox.ac.uk

Supplementary information: Supplementary data are available at *bioRxiv* online.

2 Introduction

Antibodies are critical proteins of the adaptive immune system and a powerful class of therapeutic. They are used to treat a variety of diseases including viral infections and cancer and their importance as therapeutics has steadily increased over recent years[1, 2].

Developing antibody drugs however is expensive and time-consuming, often requiring over $1bn in investment and years of research and development, with no guarantee of success. Lead optimisation - where an initial drug candidate is engineered to improve its safety, efficacy, and developability - can dominate the pre-clinical stages of this pipeline[3].

Increasing the affinity of an antibody candidate for its target (antigen) is often a key aim of lead optimisation. Traditionally, Deep Mutational Scanning (DMS) has been favoured for this step[4]. DMS involves introducing point mutations within, or near, the binding site (paratope) of an antibody and measuring their impact on binding. Beneficial mutations are then weighted more highly when designing antibody libraries for further testing.

DMS has proved effective in designing libraries highly enriched in binding variants[5]. However, DMS adds additional time and costs to experiments, and can prematurely limit the subsequent sequence space explored. Recent developments in Machine Learning (ML) have sought to address some of these issues.

For example, Hie *et al.*[6] trained a Large Language Model (LLM) to suggest 'plausible' single-point mutations, 14-71% of which were shown to improve affinity, depending on the starting wild-type antibody and antigen. To obtain multi-point mutations, the top predicted single-point mutations first required experimental validation, similar to DMS. In this hybrid scenario, fewer single-point mutants needed testing, but the sequence space explored was limited.

Shanehsazzadeh *et al.*[7] computationally suggested immediate multi-point mutations for Trastuzumab binding Human Epidermal Growth Factor Receptor 2 (HER2), a protein overexpressed in certain breast cancers. This work focused on replacing the third Complementarity Determining Region of Trastuzumab's heavy chain (CDRH3) with novel designs using both structure-design and inverse-folding models. This complete CDRH3 infilling approach produced an estimated 10.6% binder-enriched library and resulted in some sequences unlikely to be made following standard DMS. However, the majority of suggested mutations either simply

restored germline residues, or were Glycine and Tyrosine substitutions.

LLMs have also been used to optimise other aspects of antibody design. Shuai *et al.*[8] used their Immunoglobulin Language Model (IgLM) to optimise antibody humanness and developability. IgLM was built upon Open AI's Generative Pre-trained Transformer 2 (GPT-2)[9] architecture and was used to infill entire CDR loops, similar to Shanehsazzadeh *et al.* In offering immediate multi-point mutations, IgLM could potentially explore sequence space that DMS would not. However, like Shanehsazzadeh *et al.*, IgLM risks regressing to the germline due to the nature of the data it was trained on[10–12]. Furthermore, this method has not yet been experimentally validated.

Finally, if experimental data against a target of interest is already available, antigen-specific models can be trained to generate binder-enriched antibody libraries. Akbar *et al.*[13] and Frey *et al.*[14] used RNN-LSTMs and Discrete Walk-Jump Sampling respectively to generate novel Trastuzumab variants against HER2. Each generative model was trained on ∼9k experimentally confirmed binding CDRH3 mutants[5]. Akbar *et al.* achieved a predicted enrichment of 68%, while wet lab validation confirmed 70% of Frey *et al.*'s designs expressed and maintained affinity to HER2. These enrichments are higher than those reached by the general-purpose library design methods described earlier but antigen-specific models are limited in their application by the large amount of experimental training data they require.

Most of the ML methods listed above focus on designing antibody libraries from a single initial lead. These ML-designed libraries should contain more high-affinity (or more human) antibody variants than random mutations would offer, while also exploring different areas of sequence space than DMS-guided designs. However, designed libraries can be optimised further once some designs are tested experimentally. In this second step of the design process, ML classifiers can be trained on experimentally confirmed binders/non-binders and then used to pre-screen subsequent designs saving time and costs.

Li *et al.*[15] used a pre-trained LLM, Gaussian Processes (GPs), and ensemble regression models to design and screen new high-affinity single-chain fragment variable antibodies (scFvs) against a conserved coronavirus peptide. They trained their models on experimental data (26.5k heavy and 26.2k light chain sequences) that involved up to three random CDR mutations from an initial candidate. Using this approach, Li *et al.* computationally explored a relatively large

portion of sequence space. However, the large amount of data used for training is often not practical.

Mason *et al.*[5] also used a large dataset, obtained by first performing a DMS screen of Trastuzumab CDRH3 mutants tested for binding to HER2. Using the DMS results, they generated a library and experimentally classified 36.4k mutants. They then trained a Convolutional Neural Network (CNN) classifier using this data. The CNN was able to relatively accurately predict binding, achieving areas under the receiver operating characteristic and precision-recall curves (ROC AUC & PR AUC) of 0.91 and 0.83 respectively. However, like Li *et al.*, the amount of data used for training the CNN is not practical for many applications.

In our work, we expand upon these previous developments in library design and affinity prediction, using Trastuzumab-HER2 as a model system. First, we describe a new, large dataset of over 524k mutations to Trastuzumab's CDRH3 for training and evaluating affinity classification tasks. We compare this new dataset to that collected by Mason *et al.*, examining how many variants are labelled as both binding and non-binding in each dataset. We then test different ML architectures for binding affinity classification, benchmark these methods, and test their accuracy when trained on small amounts of data. Finally, we use a variety of computational methods to design novel, diverse antibody CDRH3 libraries against HER2 and computationally screen these using our highest accuracy classifier. Experimental validation of 700 of our top designed sequences using Biolayer Interferometry (BLI) is ongoing.

Together, our developments represent both a baseline for comparison of novel computational advances in this area and a lightweight end-to-end pipeline that can be used to speed up, enrich, and diversify antibody library design in many research settings (Figure 1).

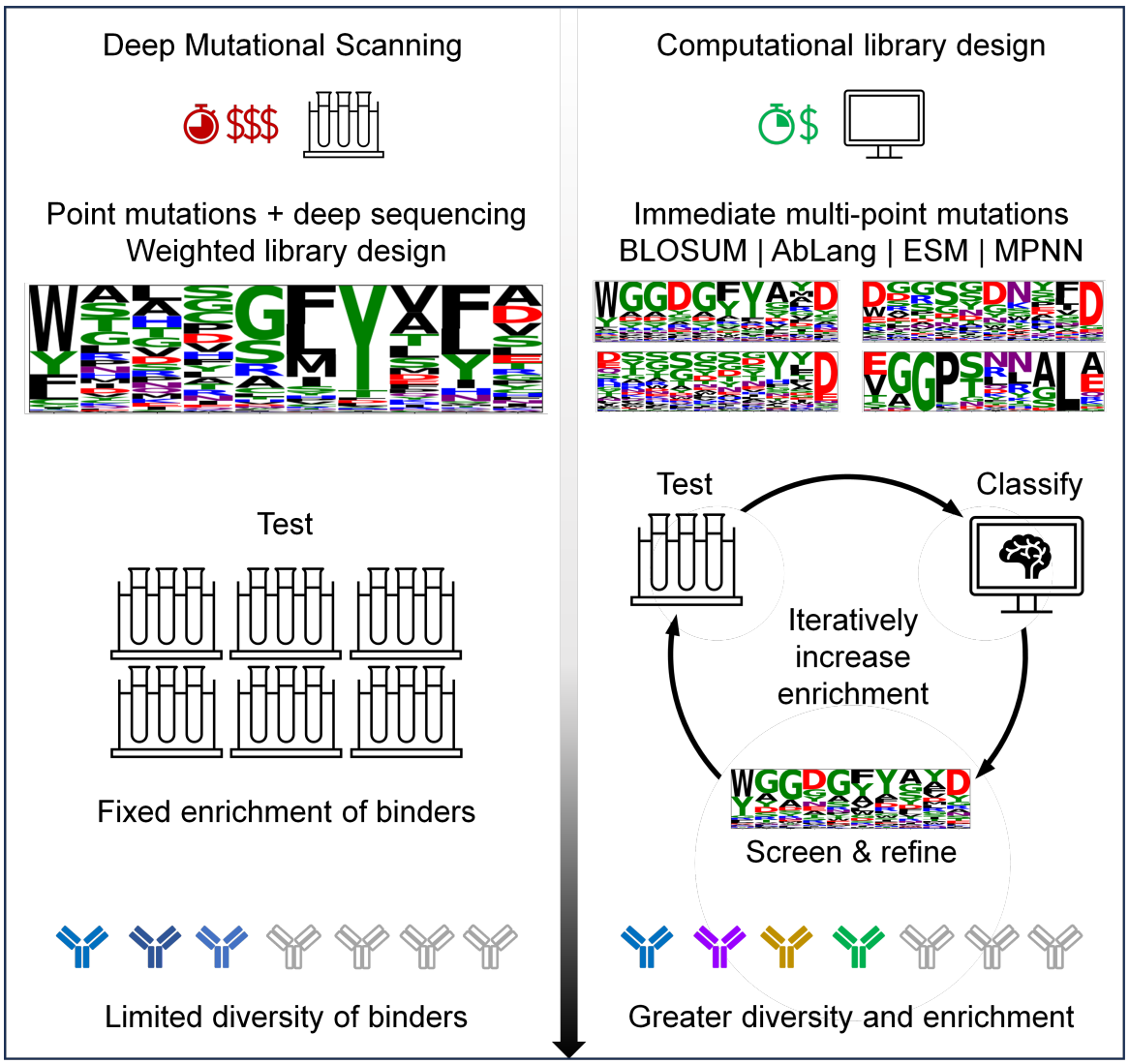


Figure 1: Overview of classical (left) and our proposed computationally-assisted (right) pipelines for discovering high-affinity variants of an initial antibody lead. The classical pipeline uses the results from Deep Mutational Scanning (DMS) to design an antibody library against a specific antigen. Thousands of sequences from this library are then experimentally tested, and enrichments of binding antibodies are usually found. Our proposed pipeline uses computational tools such as BLOSUM[16], AbLang[17], ESM[18], and ProteinMPNN[19], to design initial antibody libraries, without the need for DMS. These initial libraries are already enriched in binding antibodies above a random baseline, but this enrichment is increased further using continuous/active machine learning once experimental data starts to be collected. Our results demonstrate that machine learning classifiers can be trained on just a few hundred sequences to predict which antibodies are likely to bind. These classifiers are then used to screen untested sequences. This iterative process of experimental validation, retraining of the classifier, and refined screening of the antibody library results in highly enriched final libraries and less time and cost spent testing non-binding antibodies compared to standard experimental approaches. Furthermore, computational library design methods will produce sequences that DMS is unlikely to explore, resulting in a more diverse pool of binders.

3 Results

3.1 HER2-aff-large dataset

We use Trastuzumab-HER2 as a model system to explore improvements to computational antibody library design and iterative binder enrichment. We also evaluate our methods on a smaller dataset of broadly neutralizing anti-influenza antibodies[20], results of those tests are given in the SI.

Our HER2-aff-large dataset contains over half a million Trastuzumab variants split into 'high', 'medium', and 'low' classes based on their binding affinity to HER2. As in Mason *et al.*, a previous study of Trastuzumab-HER2 binding affinity (Table 1), mutations were limited to IMGT positions between 107 and 116 of the heavy chain only. HER2-aff-large's library design was built upon the DMS results from Mason *et al.* (see Methods) and contains 178,160, 196,392, and 171,732 high, medium, and low-affinity binders, respectively (Table 1).

A small percentage of overlap between affinity classes was observed in HER2-aff-large, with 1.1% of 'high' sequences also found in the 'medium' class, and 2.9% found in the 'low' class. This inter-class overlap is lower than that observed in Mason *et al.* (Table 1). Assigning overlapping sequences to the 'high' class only, and removing any redundancy between 'medium' and 'low' classes results in a total dataset of size 530,357 and a class imbalance of 33.6%. Removing any overlapping sequences completely results in a dataset of 524,346 sequences with a class imbalance of 32.8%. We use this latter approach throughout our work.

In our analyses, we observe that the 'medium' and 'low' classes of HER2-aff-large cluster with the negative binders from Mason *et al.* using tSNE visualisations (see SI). Furthermore, we observe that classification methods trained on data from Mason *et al.* offer low predicted binding probabilities for both 'medium' and 'low' classes, and high probabilities for the 'high' class (see SI). Due to these observations, we assign 'high' affinity sequences to be positive binders and group 'medium' and 'low' affinity sequences to be negative binders for binary classification (Table 1). This aligns broadly with the goal of selecting high-affinity antibodies during lead optimisation.

	label	# total	# unique	positive overlap	negative overlap
Mason *et al.*	positive	11,300	11,277	-	**20.76%**
	negative	27,539	27,456	8.53%	-

	label	# total	# unique	positive overlap	negative overlap
HER2-aff-lrg	positive	178,160	178,160	-	3.37%
	negative	368,124	362,549	1.66%	-

	Mason *et al.* label	# total	# unique	HER2-aff-large positive overlap	HER2-aff-large negative overlap
Inter	positive	11,300	11,277	110	**77**
	negative	27,539	27,456	**29**	46

Table 1: A breakdown of Mason *et al.* (top) and HER2-aff-large (middle) HER2 binding affinity datasets. Mason *et al.* split their dataset into two classes - positive and negative. Our HER2-aff-large dataset is split into three classes - high, medium, and low-affinity binders. In our analyses, we assign high-affinity binders to be our positive class and group the mid and low-binding affinity classes into a single negative class. This binary separation aligns with the goal of selecting high-affinity antibodies during lead optimisation. Some inter-class redundancy is expected within these datasets given the relatively low precision of the high throughput experimental methods used. More inter-set redundancy is observed in Mason *et al.*'s data than in HER2-aff-large. Little overlap exists between Mason *et al.*'s data and HER2-aff-large due to the large possible sequence space ($\sim 10^{13}$, 20 possible amino acids and 10 sequence positions). As this overlap is small, we show the absolute number of overlapping sequences here instead of percentages. We find 41% of 187 overlapping sequences labelled as binding in Mason *et al.*'s dataset are labelled negatively in our HER2-aff-large dataset. Overlaps with contrasting classes greater than 20% are shown in bold. We focus our analysis on the HER2-aff-large dataset due to its order-of-magnitude larger size and smaller inter-class overlap.

3.2 CNN classifies high-affinity antibodies with little training data

We present our results in reverse order to Figure 1 (affinity classification before computational library design) as our trained binding affinity classifier was used to computationally screen our newly designed libraries before experimental validation which is currently ongoing.

To be useful in most real-world settings, binder classification methods must perform well with limited data availability. This requirement is essential for continuous learning and iterative refinement of library design in subsequent rounds of experimentation.

We tested three classification methods - a Fast Library for Automated Machine Learning (FLAML)[21], a Convolutional Neural Network (CNN)[5], and an Equivariant Graph Neural Network (EGNN)[22] - using varying amounts of data for training and validation. Full architecture details can be found in our Methods and SI. Data was assigned to train, validation, and test sets randomly, but we ensured class imbalances remained consistent between each. Results are also presented with train, validation, and test sets split by clonotype (see Methods and SI).

The CNN outperformed both FLAML and the EGNN when trained on small amounts of data, up to 2,000 sequences (Figure 2a). The low training data requirement of the CNN (achieving a PR AUC of 0.71 when trained on only 170 sequences) offers the potential to iteratively increase experimental enrichments of binding antibodies through continuous learning as more data is collected. PR AUC is used as our main evaluation metric as it correlated with ROC AUC for all methods (see SI) but offered a greater range of discriminatory values. Baseline studies have also previously shown PR AUC to be well suited for evaluating class imbalanced problems[23]. All classifiers were optimised for ROC AUC or Accuracy to avoid potential training biases[24]. Further evaluation metrics and approximate training times can be found in the Methods and SI.

Beyond the low data regime, FLAML outperformed the CNN for intermediate train set sizes, but the performances converged when all available training data was used. Our subsequent analyses focus on the CNN due to its superior performance on small datasets, especially when the data is split by clonotype (Figure 2a).

One obvious way for a predictor to perform well in this setting is for it to favour sequences with shorter edit distances (fewer mutations) from Trastuzumab. In the training data 98% of antibodies with edit distance of one from Trastuzumab were labelled as binding, while only 17%

with an edit distance of ten bound. When the CNN was trained on all available HER2-aff-large data, using a 70-15-15 split, we observed near-perfect accuracy for edit distances one to nine (Figure 2b). Antibodies with an edit distance of nine have a similar class imbalance (19%) to those with an edit distance of ten (17%), but the PR AUC is substantially higher (0.99 vs 0.88). This disparity indicates some knowledge of the original sequence is of benefit to the trained network.

3.3 Classification accuracy varies when trained on data from different experiments

We investigated the performance of the CNN architecture when trained on different data sources. We trained our CNN on the same amount of data from both HER2-aff-large and Mason *et al.* datasets and evaluated the performance on the corresponding test sets (see SI).

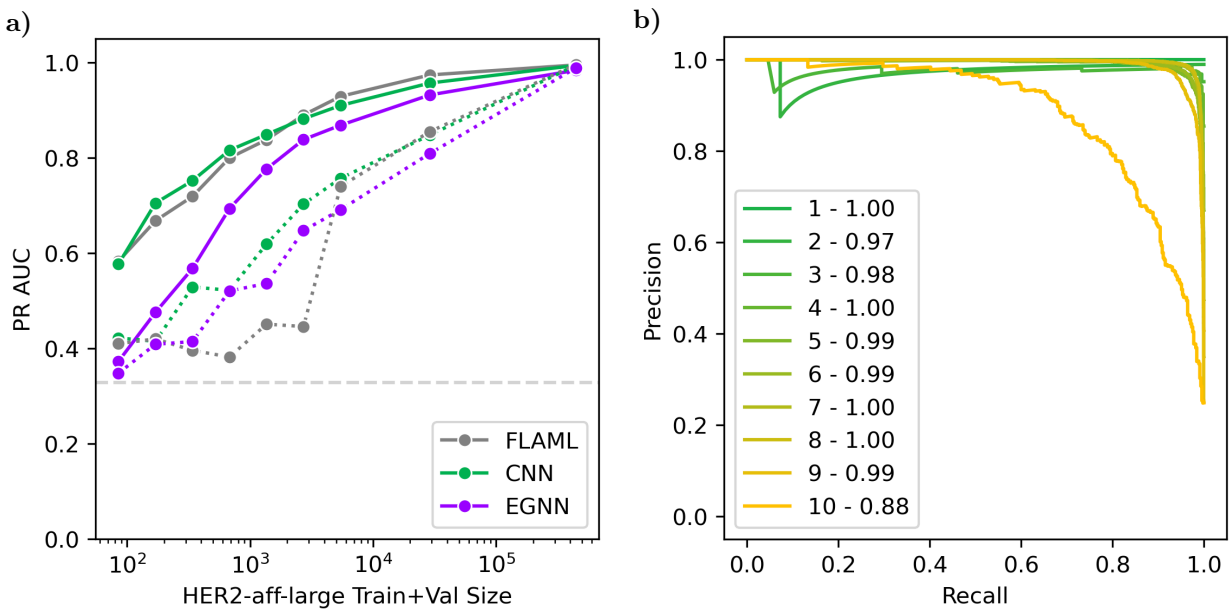


Figure 2: **a)**: Areas under the Precision-Recall curves (PR AUC) for all binder classification methods (FLAML, CNN, and EGNN). Results are shown on our HER2-aff-large dataset with overlapping sequences removed. Train, validation, and test sets are split randomly (solid) and by clonotype (dashed). Train and validation sets have a relative size ratio of 70:15. All data not assigned to the train or validation dataset is used as the test set, on which the results are presented. Random guessing would result in PR AUC values equal to HER2-aff-large's class imbalance of 0.33 (light grey dashed line). **b)**: Precision-Recall curves for variants grouped by edit distance from Trastuzumab for the CNN trained on all HER2-aff-large data, using a 70-15-15 random split (CNN-HER2-max). Results are given on the corresponding test set. The legend shows the edit distances and corresponding PR AUC values. The dips in precision at recall ∼0.1 for edit distances two and three occur due to rare misclassifications within the small test sets of these edit distances.

To allow a fair comparison, we randomly sub-sampled positive binding sequences from HER2-aff-large so that the class imbalances between the datasets were the same (26.2%). We also confirmed that the datasets contained similar edit distance distributions.

When trained (using 28.9k sequences) and tested on Mason *et al.*'s data, the CNN achieved a PR AUC of 0.84, compared to a PR AUC of 0.94 when trained and tested on HER2-aff-large (again, using 28.9k training sequences). Furthermore, the performance of the CNN fell sharply with increasing edit distance from Trastuzumab when trained and tested on data from Mason *et al.* This effect was less strong when the CNN was trained and tested on HER2-aff-large, despite this dataset containing proportionally fewer sequences with edit distances of nine and ten compared to the data from Mason *et al.*

3.4 Trained classifiers do not transfer well between experiments

To test the robustness of the CNN affinity classifier, we trained on data from HER2-aff-large and tested on data from Mason *et al.*, and vice versa. We also evaluated both trained CNNs on the 198 Trastuzumab-length-matched binding variants released by Shanehsazzadeh *et al.*[7].

When tested on data from an experiment different from the training data, the predictive power dropped compared to evaluation on data from the same experiment (see SI). This drop in performance was greater than the fall seen when splitting a dataset by clonotype (Figure 2a), indicating that the drop was not caused by exploring different areas of sequence space alone. Instead, differences in experimental set-ups and/or cut-offs used to separate positive and negative classes may contribute to the poor generalisability.

As HER2-aff-large offered the largest dataset of HER2 binding data available and also produced the best models (highest PR AUC, see SI), we used our CNN trained on all this data (CNN-HER2-max) to screen our novel computational library designs before experimental validation.

3.5 BLOSUM, AbLang, ESM, and ProteinMPNN generate antibody libraries with high proportions of predicted binders

We next explored the ability of computational methods to complement or replace DMS for library generation (Figure 1). Computational methods, if able to replace DMS, could design

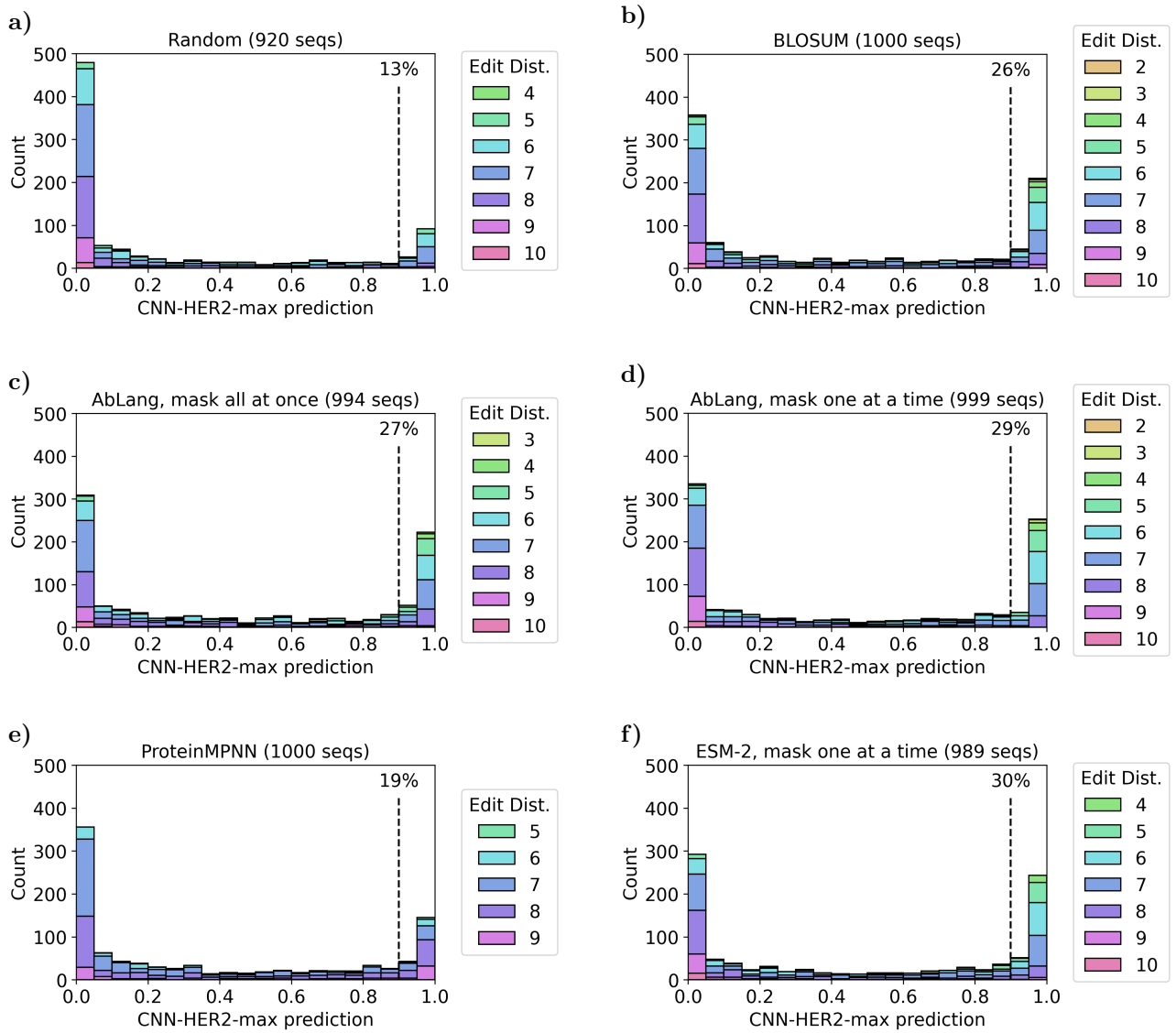


Figure 3: Distributions of HER2 binding predictions for ∼1k sequences generated using BLOSUM, AbLang, ESM, and ProteinMPNN, plus a random baseline. These sequences were subsampled from a maximum of 1m generated sequences to match the edit distance distribution observed in HER2-aff-large, where possible. Some methods failed to produce any sequences at certain edit distances - when an edit distance is not present in the legend, no sequences of this edit distance were generated in the 1m attempts, and the total count (shown in the figure titles) is slightly less than 1,000. Binding predictions (x-axis) are those given by our CNN trained on all HER2-aff-large data (CNN-HER2-max), using a 70-15-15 split. A value close to '1' indicates a high predicted binding probability, while a value close to '0' indicates a low probability of binding. Dashed lines are drawn at 90% CNN-HER2-max predicted binding probabilities and the percentage of sequences with binding probabilities above this cut-off are stated above the line. All methods succeeded in producing some sequences with high predicted binding probabilities to HER2 across a range of edit distances.

more diverse antibody libraries and save initial experimental time and money.

Using our CNN trained on all HER2-aff-large data (CNN-HER2-max), we predicted the binding probabilities of Trastuzumab CDRH3 variants generated using BLOSUM[16], AbLang[17], ESM[18], ProteinMPNN[19], and a random baseline. Briefly, BLOSUM matrices measure whether or not an amino acid substitution is conservative, AbLang and ESM are LLMs that output masked-residue likelihoods, and ProteinMPNN designs sequences for a given structure (see Methods and original papers for more details). These computational tools were used to design diverse, length-matched Trastuzumab CDRH3 libraries without relying on the results of any experiments, effectively zero-shot (AbLang - mask all at once) and one-shot (BLOSUM, ProteinMPNN, ESM-2, AbLang - mask one at a time) prediction of the CDRH3.

To allow fair comparisons to experimental data, for each library design method we aimed to sub-sample 1,000 sequences from a maximum of 1,000,000 generated sequences to match the edit distance distribution observed in HER2-aff-large (see SI). This sub-sampling is required as smaller edit distances contain proportionally more high-affinity variants than larger edit distances.

The exact libraries generated varied depending on the starting seed, but some methods consistently failed to explore small edit distances. This restriction meant the total counts of AbLang and ESM sub-sampled libraries were below 1,000 but above 985. Our Random library contained just 920 sequences. ProteinMPNN only sampled edit distances of five to nine, so no edit distance restrictions were included for this method. Full details of how these sequences were generated can be found in our Methods. Predicted enrichments when enforcing edit distance distributions to match those of ProteinMPNN can be found in the SI.

We found a large proportion of sequences generated using all methods to have high CNN-HER2-max predicted binding probabilities (Figure 3). For Random, BLOSUM, AbLang (mask all ten residues at once), AbLang (mask one residue at a time), ProteinMPNN, and ESM (mask one residue at a time), we found 13%, 26%, 27%, 29%, 19%, and 30% of sequences to have CNN-HER2-max binding probabilities above 90%, respectively. To determine how many sequences actually bind, Biolayer Interferometry experiments are being run for 140 sequences designed by each method. This preprint will be updated once these results become available.

3.6 BLOSUM, AbLang, ESM, and ProteinMPNN's predicted binders cover diverse areas of sequence space

The sequences predicted to bind HER2, designed using all methods, covered diverse areas of sequence space (Figure 4), with many large edit distances away from Trastuzumab's original sequence (`SRWGGDGFYAMDY`, IMGT positions 105 to 117; Figures 3, 4c, and SI).

Many of the designed sequences were also large edit distances from the closest germline (`AR-------YFDY`, IMGT positions 105 to 117; Figure 4c), as given by the V and J genes (IGHV3-66 and IGHJ4) assigned using ANARCI[25], with the corresponding sequence taken from imgt.org. The residue positions indicated by '-' are non-germline positions.

Finally, most designed sequences and confirmed binders occupied areas of sequence space far from our observed data (HER2-aff-large) as visualised using a tSNE plot (Figure 4a,b). PCA and UMAP visualisations showing similar distributions can be found in the SI. These newly confirmed binding sequences were unlikely to have been designed following DMS experiments due to DMS weighting the search space based on observations from single-point mutations. HER2-aff-large's comparatively small search area in Figure 4a (grey) highlights the restrictive nature of DMS.

Despite their diversity, some shared features were observed between most computationally designed antibody libraries, including a tendency towards Glycines at IMGT positions 108, 109, and 110, and Aspartic acid at position 116. These shared features are observed in Trastuzumab's original CDRH3, and some are similar to those in HER2-aff-large and Mason *et al.*'s data (Figure 4c), which suggests that they are beneficial for HER2 binding.

ProteinMPNN occupies a distinct area of the tSNE visualisation in Figure 4a, and its logo plot stands out from those of all other methods (Figure 4c and SI). This could be because ProteinMPNN is the only structure-based design method we include, or because it is a general-purpose protein sequence design method. However, ESM is also a general-purpose protein sequence design method and it designed sequences similar to those from AbLang (Figure 4 and SI).

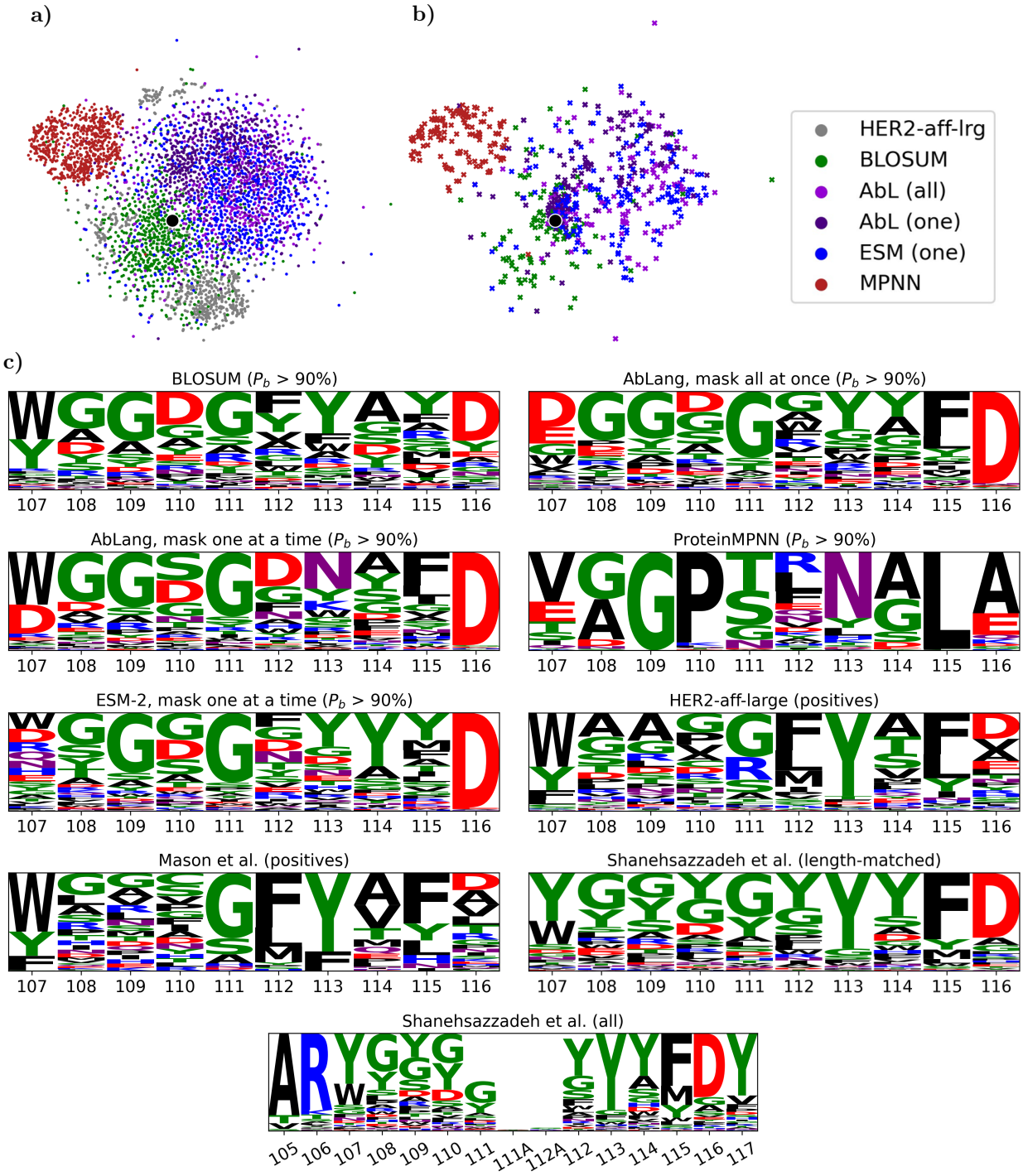


Figure 4: **a)** t-SNE plot comparing the sequence space explored by different methods when designing HER2 binders. ∼800 sequences designed by each method are shown and compared to HER2-aff-large. Our t-SNE takes as input the flattened one-hot encodings of the designed CDRH3 loops (residues 107 to 116). **b)** t-SNE of sequences shortlisted for experimental validation with CNN-HER2-max binding probabilities above 90%. Both t-SNEs use the same scale. Outlying data points are cropped from **a)** for plotting clarity. Trastuzumab is shown in black. **c)** Logo plots for BLOSUM, AbLang, ESM, and ProteinMPNN HER2 predicted binders. All positive experimental data are shown for comparison.

3.7 Simulations show rapid increases in enrichment are achievable without sacrificing library diversity through active learning and continuous iterative library refinement

Figure 5 demonstrates how we can combine both computational library design methods, such as BLOSUM, AbLang, ESM, and ProteinMPNN, and a simple ML classifier, such as our CNN, to design and iteratively improve diverse antibody libraries, as originally suggested in Figure 1.

Our previous results show how computational zero-shot design methods, such as AbLang (mask all ten residues at once), can generate libraries with high predicted binder enrichments, ∼25%. These initial libraries are diverse (Figure 4) and contain more binding variants than one would expect from random mutations, yet this enrichment can be increased further still.

Using a reduced form of HER2-aff-large with just 10% enrichment, we demonstrate how the iterative retraining of a CNN classifier can quickly increase the starting enrichment above that achieved by DMS (Figure 5). We chose 10% starting enrichment assuming that not all predicted binders will bind ($< \mu$M K_D, in agreement with Shanehsazzadeh *et al.*). In similar work, Porebski *et al.*[26] experimentally confirm that approximately half of their designs with $P_b > 90\%$ bound with greater Fluorescence Intensities (FI, shown to correlate with K_D) than a confirmed 320nM design. Should we observe similar results, this would correspond with true floor enrichments ∼10%. Biolayer Interferometry experiments (ongoing) will confirm our true success rate for each library design method.

Our simulations show that after testing just 540 sequences, subsequent rounds of experimentation contained libraries with enrichments above 30%. Enrichments continued to increase rapidly up to and above 80% after testing fewer than 2k sequences. This continuous retraining of the CNN and screening of new sequences did not limit the sequence search space but instead explored it in a more time and cost-efficient manner (Figure 5). All areas of the tSNE plot containing positive data continued to receive hits in the latter rounds of enrichment. Additionally, the CNN did not learn to favour shorter edit distances from Trastuzumab, with the distribution of edit distances seen in latter rounds closely matching that of HER2-aff-large's positive data.

Crucially, this cycle of continuous learning and iterative library refinement is independent of

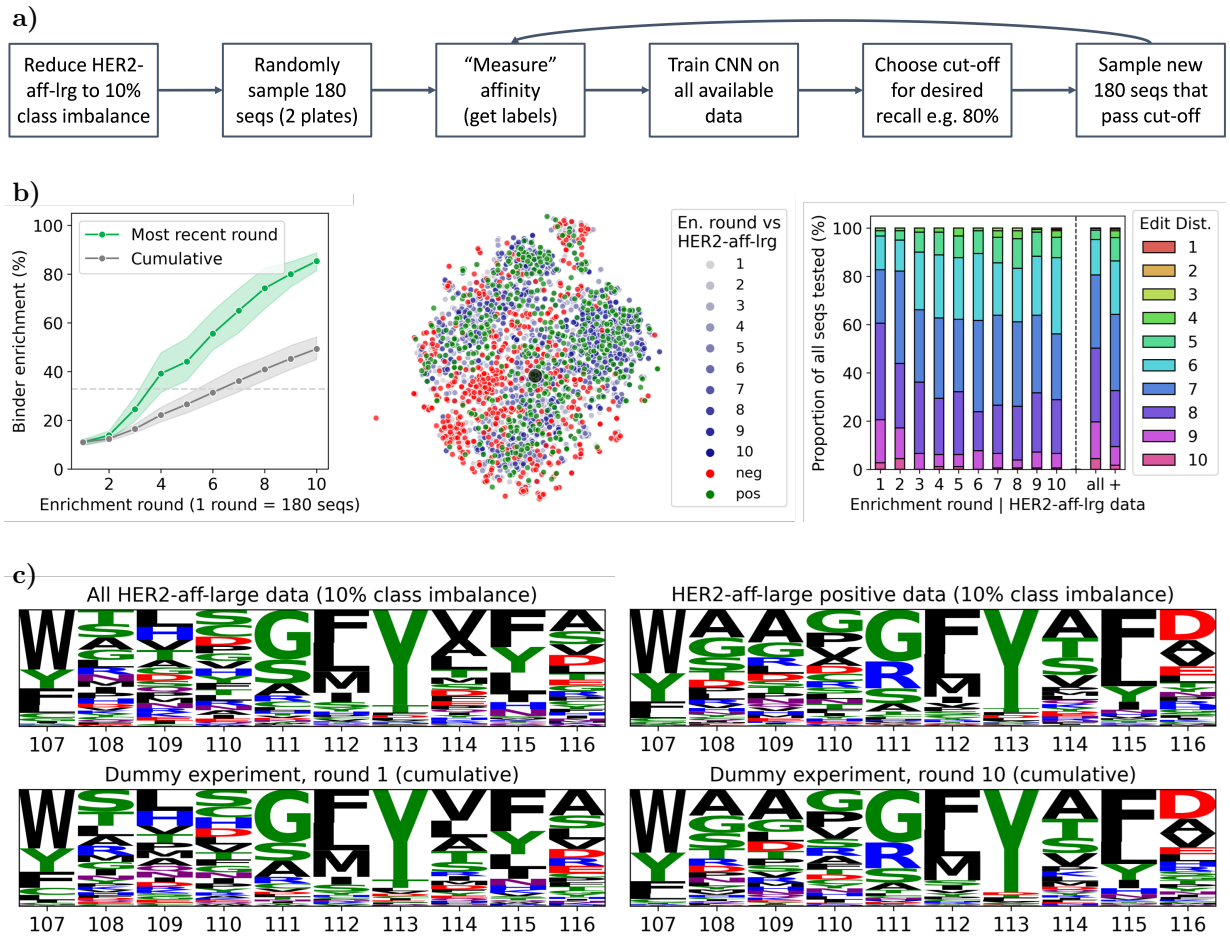


Figure 5: Results from a simulated experimental setup showing the iterative increase in binder enrichment achieved as more data is collected. Flow chart **a)** illustrates the main steps of this pipeline. First, 180 antibodies (∼two 96-well plates, allowing for controls) were randomly selected from our HER2-aff-large dataset. We used a reduced form of HER2-aff-large for this experiment with a class imbalance of 10% (HER2-aff-large-10%) to simulate the lower enrichments one might achieve when using a computational library design method, such as BLOSUM, in place of DMS. After sampling our initial sequences, we 'measured' their affinity i.e. we unblinded their labels in our simulation. With this newly labelled data, we trained a CNN to classify binding from non-binding sequences, using an 80-20 train-validation data split. The trained CNN was then used to screen newly drawn data from HER2-aff-large-10%, keeping only sequences that passed an optimised cut-off and that had not been tested in the previous round. The cut-off was designed to maintain 80% recall on our validation set to ensure we did not prematurely limit our sequence search space. With 180 new sequences available, we 'measured' their affinity and retrained the CNN on all 360 labelled sequences. This process was repeated for ten rounds of enrichment so that our final library contained 1,800 unique antibodies. Figure **b), left** shows the increasing affinity achieved after each round of experimentation and retraining of the CNN (ten repeats with 95% confidence shaded confidence intervals). The dashed line shows the 32.8% enrichment achieved following DMS. Figure **b), middle** shows that in addition to the rapid increase in enrichment, we continue to explore the full HER2-aff-large-10% positive sequence space thoroughly in latter rounds of enrichment. Figure **b), right** further highlights that the CNN does not learn to simply favour shorter edit distances and that the edit distance distribution of latter rounds closely matches that of HER2-aff-large-10% positive data. Finally, Figure **c)** supports the above, showing that, as expected, the logo plot of our antibody library following the first round of experimentation matches that of HER2-aff-large-10%. However, after ten rounds of enrichment, the logo plot of our 1,800-member antibody library closely matched that of the positive subset of HER2-aff-large-10%.

the initial library design method. Such an approach could therefore allow faster and/or cheaper exploration of high-affinity antibody variants across diverse sequence space.

4 Discussion

Machine learning shows great promise for accelerating therapeutic antibody design and optimisation. The field, however, lacks robust baselines quantifying expected binder enrichments and computational strategies that can be seamlessly integrated with experimental workflows.

To support baseline studies, we generated the largest publicly available dataset of antibody variants with antigen binding affinity labels (>524,000 Trastuzumab CDRH3 variants). ML models trained on this dataset were shown to accurately distinguish high- from medium- and low-affinity HER2-binders. A CNN architecture struck the balance between simplicity and complexity: it outperformed a more complex, structure-based EGNN model when trained on small datasets and was more robust to strict train-test splits than simpler, tree-based models. ML classifiers with low training data requirements, such as our CNN, are useful when obtaining large amounts of labelled data is too costly or impractical. Additionally, these classifiers could be used to screen and prioritize novel designs in an iterative development cycle.

We also explored the effectiveness of computational methods (BLOSUM, AbLang, ESM, and ProteinMPNN) for novel library design. A large proportion of sequences generated using each method were predicted to bind with high affinity. The predicted enrichments of 19-30% are close to those achieved by traditional DMS approaches, of ∼33%. Experimental validation of our predictions is ongoing. Predicted binders exhibited broad sequence diversity, although not always within a method. Benefits could therefore be found by employing multiple methods in the initial stages of library design if maximising diversity is a top priority.

Finally, we combined the above concepts in a simulated iterative development cycle. Starting with just 180 affinity-labelled sequences with an assumed 10% minimum true enrichment of our computationally designed libraries, we simulated iteratively training the CNN, scoring sequences, and experimentally validating designs. Binder enrichment ratios were found to increase rapidly, surpassing traditional DMS approaches after just three rounds of training and testing. After ten rounds (fewer than 2,000 sequences) enrichments exceeded 80%. The resulting library closely resembled the high-affinity variants in the HER2-aff-large dataset in

terms of sequence identity, edit distances, and diversity. These similarities indicate that our approach continues to fully explore the entire high-affinity sequence space, but does so more efficiently by avoiding testing low-affinity designs. This iterative computationally-aided cycle is fast and independent of the target or library design method.

Recent breakthroughs in deep learning, such as generative AI, promise revolutionary changes to the world of drug discovery. We demonstrate, however, that comparatively simple methods may achieve similar performances. Effectively integrating lightweight *in silico* and experimental methods into a continuous learning cycle will allow high-affinity antibody variants to be explored at low costs and timescales in most research settings.

5 Methods

5.1 HER2-aff-large collection methods

The Trastuzumab single chain variable fragment (scFv) CDRH3 dataset used in this study, HER2-aff-large, was guided by the site-specific Deep Mutational Scanning results generated previously by Mason *et al.*[5].

Briefly, a Trastuzumab scFv antibody library was cloned in a pSYD yeast display vector, a variant of the pDNL6 yeast display vector (pSYD uses N-terminal fusion for scFv-aga2 display, while pDNL6 uses a C-terminal fusion of aga2-scFv). The Trastuzumab scFv antibody library cloned in pSYD vector was transformed in EBY100 yeast cells (ATCC #MYA-4941DQ) selected on SD + CAA plates (2% dextrose, 0.67% yeast nitrogen base, and 0.5% casamino acids yeast selection media) at 30°C for 48-72 hours. Yeast display analysis of the Trastuzumab scFv library was performed as described previously by Ferrara *et al.*[27] and Chao *et al.*[28].

The next day, the cell pellet was resuspended in SG + CAA (containing 2% galactose and 0.1% dextrose) at 0.5 OD/ml and incubated at 20°C with shaking for one to two doublings, as determined by OD. The cells were washed with the wash buffer and processed for staining to check HER2 binding. Around 1-10$\times 10^7$ cells were labelled with 100μg/ml anti-V5 tag antibody followed by addition of 100nM HER2 and incubated for 30 minutes on ice. The cells were then washed twice more with wash buffer and labelled with a 1:200 dilution of secondary reagents (goat anti-mouse - Alexa 488 and streptavidin-PE).

Finally, the cells were incubated for 30 minutes on ice, washed twice with a wash buffer, and resuspended in 1ml of sorting buffer. To determine their affinity, the cells were sorted for the brightest V5 FITC positive (scFv expression) antigen binding population (PE positive) and labelled as high-affinity binders, as shown in the SI. The cells were further sorted for the brightest V5 FITC positive medium and low-affinity antigen binding populations. The populations were sorted into tubes containing YPD media and grown in SD + CAA liquid media at 30°C with shaking overnight as described previously[27].

Plasmid DNA was isolated using a yeast plasmid isolation kit (Zymoprep Yeast Plasmid Miniprep I #D2100) following user protocol. The variable heavy (VH) gene containing the CDRH3 sequence for each population was PCR amplified using in-house NGS-specific primers.

The amplicons were PCR-cleaned and prepared for NGS. The DNA libraries were sequenced on Illumina using NovaSeq 6000 S2 Reagent Kit v1.5 (300 cycles) and the raw data has been deposited on Zenodo - doi.org/10.5281/zenodo.10549114. The primers used for generating variable heavy amplicons were:

`NGSVH Fwd: 5´CACCCGTTATGCCGACAG3´`

`NGSVH Rev: 5´GGGATTGGTTTGCCGCTAG3´`

The raw paired NGS reads were merged using PEAR (v0.9.6). The subsequent dataset consisted of 618,585, 799,368, and 663,397 high, medium and low-affinity unique CDRH3 sequences, respectively. Singleton (count=1) sequences were removed from the dataset to improve the quality of the data. The final Trastuzumab variant dataset comprised 178,160, 196,392, and 171,732 sequences in 'high', 'medium', and 'low' affinity binder classes, respectively. The heavy and light chain sequences (from *1n8z*) were numbered according to the IMGT scheme. Heavy chain insertion start and stop positions are 107 and 116, respectively (spliced positions are shown in bold).

Heavy chain sequence:

`EVQLVESGGGLVQPGGSLRLSCAASGFNIKDTYIHWVRQAPGKGLEWVARIYPTNGYTRYADSVKG`
`RFTISADTSKNTAYLQMNSLRAEDTAVYYCSR`**`WGGDGFYAMD`**`YWGQGTLVTVSSA`

Light chain sequence:

`DIQMTQSPSSLSASVGDRVTITCRASQDVNTAVAWYQQKPGKAPKLLIYSASFLYSGVPSRFSGSR`
`SGTDFTLTISSLQPEDFATYYCQQHYTTPPTFGQGTKVEIKR`

5.2 Train, validation, and test set creation

For all tests, a train-validation dataset size ratio of 70-15 was used as in Mason *et al.* For each train-validation dataset, all remaining data was assigned to the test set. For each train-validation dataset size, we split the data both randomly and by clonotype. When splitting by clonotype, sequences were clustered according to their V and J genes (as annotated by ANARCI[25]), and by 70% sequence identity across the CDRH3. All HER2-aff-large sequences share the same V-gene (IGHV3-66) and one of two J-genes (IGHJ4 or IGHJ1). All members of a clonotype were added to the same train, validation, or test set. In all cases, we ensured train, validation, and test sets have the same class imbalances (i.e. ratio of binders to non-binders).

When comparing models, such as FLAML and CNN, we used identical train, validation, and test sets. All datasets can be found at doi.org/10.5281/zenodo.10549114.

5.3 FLAML Auto-ML

FLAML - a Fast Library for Automated Machine Learning[21] was used to provide a simple ML baseline. FLAML's default settings were used, allowing it to test multiple tree-based architectures, such as LightGBM, XGBoost, and random forest. Hyperparameter optimisation was also performed automatically.

For HER2 binders, FLAML took as input a one-hot encoding of the 10 CDRH3 residues between IMGT positions 107 and 116. When the mutated residues were dispersed throughout the antibody, as was the case for the Influenza data, a one-hot encoding of the entire sequence was used. FLAML requires 1D, not 2D, input; therefore, the one-hot encodings were flattened by concatenating the individual residue encodings e.g. $20 \times 10 \rightarrow 200 \times 1$.

FLAML was allowed up to six hours to train and the best model was selected for testing in each instance. XGBClassifier performed best for all HER2-aff-large train and validation set sizes except the smallest (85) and largest (445,694) when the LGBMClassifier was selected. Early stopping was enabled. For training dataset sizes up to 3,000 sequences, less than five minutes was required to complete training on a CPU.

5.4 Convolutional Neural Network

We used a Convolutional Neural Network (CNN), adapted from Mason *et al.*[5], to classify binding and non-binding antibodies (see SI for details).

Similar to FLAML, for HER2 binders the CNN took as input a one-hot encoding of the 10 CDRH3 residues between IMGT positions 107 and 116. When the mutated residues were dispersed throughout the antibody, as was the case for the Influenza data, a one-hot encoding of the entire sequence was used. The CNN input was left as 2D (10×20 for HER2 binders).

The CNN was trained for a maximum of 100 epochs and training was stopped if no decrease in loss was observed for five successive epochs. CPU training times ranged from less than a minute when trained on fewer than 3,000 sequences, to three hours when the entire HER2-aff-large dataset was used. Hyperparameters were optimised by Mason *et al.*[5] and we tried no

further optimisation.

5.5 Equivariant Graph Neural Network

In order to test the ability of a more complex method that takes into account protein structure, we applied an Equivariant Graph Neural Network (EGNN) architecture adapted from Hummer *et al.* [29]. The model takes as input a 3D structure of the Trastuzumab-HER2 complex and generates residue-level graphs. The graphs include the C_α atoms of the 10 mutated CDRH3 residues and surrounding neighborhood (antibody C_α atoms within 10 Å of CDRH3 C_α atoms (antibody neighborhood), antigen C_α atoms within 10 Å of the antibody neighborhood and antigen C_α atoms within 10 Å of these antigen atoms). The node features are a one-hot encoded vector describing the residue type and chain type (antibody or antigen). The edge features are a one-hot encoded vector describing whether the edge is intra-binding partner, i.e. between atoms on the same binding partner, or inter-binding partner, i.e. between atoms on different binding partners.

The graphs are fed through a network composed of three E(n) Equivariant Graph Convolutional (EGC) layers [22] with a hidden dimension of 128. The models were trained with Binary Cross-Entropy with Logits loss. The architecture was implemented using PyTorch and PyTorch Geometric.

To generate the structural inputs for the EGNN, we used FoldX BuildModel [30] to introduce mutations to the Trastuzumab CDRH3, starting from a FoldX-'repaired' structure in complex with HER2 (PDB 1N8Z [31]). As FoldX does not model changes to the backbone [32], the true structural effects of the mutations are unlikely to be represented. However, this approach has the advantages of speed (and therefore compatibility with high-throughput datasets) and avoiding the need for docking by starting from the structure of a bound complex.

The EGNN was trained for 100 epochs. GPU training times ranged from less than 3.6 hours, on datasets composed of fewer than 3,000 sequences, to ca. 20 days on the full HER2-aff-large dataset.

5.6 Computational library design methods

We used BLOSUM[16], AbLang[17], ESM-2[18], and ProteinMPNN[19] to generate antibody libraries against HER2, using Trastuzumab as an initial lead. For each method, we generated 1,000,000 sequences and aimed to retain 1,000 sequences that matched the edit distance distribution of HER2-aff-large. This sub-sampling allowed a fairer comparison between methods (Figure 3), as smaller edit distances from Trastuzumab contain proportionally more high-affinity variants than large edit distances.

Some methods tended to generate sequences with larger edit distances from Trastuzumab due to the nature of their sampling distributions (SI). In these instances, the true number of sequences sampled at shorter edit distances was sometimes below the target number. The true number of sequences sampled for each method is shown in Figure 3 and the SI.

5.7 Random library design

As a baseline, we randomly mutated each of the ten Trastuzumab residues between positions 107 and 116 to each of the 20 standard amino acids with equal probability. The distribution of amino acids to sample from for each of the ten sequence positions was the same (see SI).

5.8 BLOSUM library design

BLOSUM matrices (BLOcks SUbstitution Matrices) provide information on which amino acid substitutions are most likely to be observed[16]. These matrices can be used to obtain the frequencies with which we expect to observe each amino acid type replaced with any of the standard 20 amino acids[33].

For our BLOSUM library design, we used the BLOSUM-45 matrix and, as background, the amino acid frequencies observed in CRDH3s in SAbDab (the Structural Antibody Database)[34, 35] in our reverse calculation (see SI). These background frequencies describe how often each amino acid is found in a CDRH3, not how often each is likely to be replaced by another.

Once our final BLOSUM frequencies had been obtained, we used these to weight the sampling of each amino acid type based on the original Trastuzumab residue between positions IMGT 107 and 116. The distribution of amino acids sampled from was identical for matching amino acid types in Trastuzumab e.g. the Glycine residues at positions 108, 109, and 111 (see

SI).

5.9 AbLang library design

AbLang is an antibody language model designed to restore missing residues[17]. AbLang was trained on over 14m sequences from the Observed Antibody Space database, OAS[10, 11]. This set was dominated by germline sequences, so the restored residues often reflect germline observations.

We used AbLang to obtain amino acid likelihoods at each sequence position, independent of the original Trastuzumab residues. We tested two methods - masking the entire ten residues between IMGT positions 107 and 116 at once and masking just one residue at a time. In both instances, we limited AbLang to predicting CDRH3s of the same length as Trastuzumab.

The likelihoods returned by AbLang were used to weight the sampling of each amino acid type between positions IMGT 107 and 116. AbLang's likelihoods are position-specific, meaning the sampling weights were unique for each sequence position, unlike BLOSUM (see SI).

5.10 ESM-2 library design

Evolutionary Scale Modeling (ESM) is a general-purpose protein language model (not fine-tuned for studying antibodies) from Meta's Fundamental AI Research (FAIR) Protein Team[18]. Recent methods have found success in using ESM to suggest affinity-improving single-point mutations[6]; here we test its efficacy for multi-site mutations.

Like AbLang, ESM-2 can be used to obtain amino acid likelihoods at masked sequence positions. We masked each residue between IMGT positions 107 and 116 one at a time and used the 33-layer, 650m parameter implementation of ESM-2 (esm2_t33_650M_UR50D) to suggest residue likelihoods. We also tested masking the entire ten residues (IMGT positions 107 to 116) at once, limiting ESM-2 to predicting CDRH3s of the same length as Trastuzumab. However, predicted enrichments when masking all residues at once were lower than when masking residues one at a time (see SI), so we only experimentally tested the latter.

The likelihoods returned by ESM-2 were used to weight the sampling of each amino acid type between positions IMGT 107 and 116. ESM-2's likelihoods are position-specific, like AbLang (see SI).

5.11 ProteinMPNN library design

ProteinMPNN[19] is a deep-learning method for predicting protein sequences for a given structure. We used the ABodyBuilder2[36] predicted structure of Trastuzumab as the base structure as any structure-based method must work with a predicted structure as crystal structures are not readily available for most antibodies. CDRH3 residues between IMGT positions 107 and 116 were then masked, and ProteinMPNN was tasked with generating sequences predicted to fit the modelled CDRH3 conformation.

We used a high sampling temperature of 0.3 for ProteinMPNN to produce diverse sequences. Unlike the previous methods described which were used to generate likelihoods to sample from, we used the exact sequences predicted by ProteinMPNN for our generated library. Due to speed limitations and an observed lack of diversity, we generated only 3,000 sequences with ProteinMPNN, which resulted in 2,331 non-redundant outputs.

Considering the constrained distribution of sequences created by ProteinMPNN, it was not possible to sub-sample from these to match the edit distance distribution of HER2-aff-large. Instead, 1k sequences were randomly sampled from the 2,331 designed by ProteinMPNN for comparison against other methods. Further details can be found in the SI.

6 Data Availability

6.1 HER2 affinity data

Our classification methods were trained and evaluated on data for two distinct targets. The first and most varied of these datasets involves Trastuzumab variants tested for binding against HER2, a protein overexpressed in certain breast cancers.

The HER2 binding data used in this paper was collected from three sources - Mason *et al.*[5], Shanehsazzadeh *et al.*[7], and our own dataset - HER2-aff-large. We focused most of our analysis on HER2-aff-large as this dataset is larger than the others (Table 1). To allow direct comparisons against previous analyses[5], we also show results on Mason *et al.*

6.1.1 Mason *et al.* data

The Mason *et al.* dataset contains 38,733 mutated Trastuzumab sequences classified as positive (11,277) and negative (27,456) HER2 binders. Ten residues between IMGT[37] positions 107 and 116 were mutated in their analysis, while all other positions were fixed to match Trastuzumab. Binding sequences were classified using mammalian display and Fluorescence Activated Cell Sorting (FACS). The ∼30% enrichment of binding sequences observed in their experiments is a result of single-site deep mutational scanning (DMS) being used to guide the library design.

Overlap is observed between the two reported classes, with 20.8% of positive-labelled sequences also found in the negative class. Mason *et al.*[5] assign all overlapping sequences to the positive class in their analysis, resulting in 36,391 non-redundant sequences with a class imbalance of 31.0%. Choosing instead to remove any overlapping sequences completely results in a dataset of 34,049 sequences with a class imbalance of 26.2%. Throughout the paper, we present our results using the latter approach, given the ambiguity of the overlapping sequences.

6.1.2 Shanehsazzadeh *et al.* data

Shanehsazzadeh *et al.* recently published a dataset of 421 HER2 binding antibody sequences, validated using Surface Plasmon Resonance (SPR)[7]. No negative sequences were published. This dataset could therefore not be used for training and was only used as a test set.

Shanehsazzadeh *et al.* allow mutations between IMGT positions 105 and 117. Their designs also frequently delete one residue from Trastuzumab's CDRH3 and occasionally insert up to two additional residues. We limited our analysis to only those sequences that matched Trastuzumab in length (198 sequences). We did not enforce the matching of IMGT positions 105, 106, and 117 to Trastuzumab as this reduced the usable dataset size to only four sequences.

6.1.3 HER2-aff-large data

Our HER2-aff-large dataset contains over half a million CDRH3 sequences split into 'high', 'medium', and 'low' classes based on their binding affinity to HER2. Similar to Mason *et al.*, mutations are limited to IMGT positions between 107 and 116. The HER2-aff-large dataset contains 178,160, 196,392, and 171,732 high, medium, and low-affinity binders, respectively.

In our analyses, we observe that the 'medium' and 'low' classes of our HER2-aff-large dataset cluster with the negative binders from Mason *et al.* using tSNE visualisations (see SI). Furthermore, we observe that classification methods trained on data from Mason *et al.* offer low predicted binding probabilities for both 'medium' and 'low' classes, and high probabilities for the 'high' class (see SI). Due to these observations, we assign 'high' affinity sequences to be positive binders and group 'medium' and 'low' affinity sequences to be negative binders in our methods.

Overlap between classes was also observed in our HER2-aff-large dataset, with 1.1% of 'high' sequences also found in the 'medium' class, and 2.9% found in the 'low' class. Assigning overlapping sequences to the positive (high) class only, and removing any redundancy between medium and low classes results in a total dataset of size 530,357 and a class imbalance of 33.6%. Removing any overlapping sequences completely results in a dataset of 524,346 sequences with a class imbalance of 32.8%. We use this latter dataset in our analyses.

6.2 Influenza affinity data

In addition to anti-HER2 data, we test our methods on an anti-influenza dataset (see SI). The influenza data we used was obtained from Phillips *et al.*[20] This dataset consists of variants of two broadly-neutralising anti-influenza antibodies (CR9114 and CR6261), with binding affinities measured against four subtypes of the influenza surface protein hemagglutinin (HA).

Mutations were made throughout the entire heavy chain but were limited to mutations back to the germline only.

The CR9114 sub-dataset contains 65,536 variants (16 positions with all wildtype-germline mutation combinations). The CR6261 sub-dataset contains 2,048 variants (11 wildtype-germline combinations). The imbalance of binding vs. non-binding variants for each sub-dataset varied greatly, from 0.3% to 97%, when their affinity was measured against each of the four HA subtypes (see SI).

We trained our CNN classifier on all available CR9114 antibody-HA-subtype combinations and compared our results to those from Bachas *et al.*[38], achieving higher classification accuracy in most instances (see SI).

Author contributions

L.C. prepared the training data, investigated the library design methods, and evaluated the FLAML and CNN affinity classification approaches. A.M.H. performed the EGNN analysis. B.B.M. led the HER2-aff-large data collection. R.A., P.R., A.S., K.L.Q., and F.L.J. supported the data collection and pre-processing. L.C., A.M.H., and B.B.M. wrote the manuscript. C.M.D., J.R.J., and V.G. supervised the work. All authors read and approved the final version of the manuscript.

Funding

This work was supported by the Biotechnology and Biological Sciences Research Council (BBSRC) [#BB/V509681/1 awarded to L.C.], the Medical Research Council [#MR/N013468/1 awarded to A.M.H], the Leona M. and Harry B. Helmsley Charitable Trust [#2019PG-T1D011 awarded to V.G.], UiO World-Leading Research Community [awarded to V.G.], UiO: Life-Science Convergence Environment Immunolingo [awarded to V.G. and F.L.J.], the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement [#801133 awarded to P.R.], EU Horizon 2020 iReceptorplus [#825821 awarded to V.G.], a Norwegian Cancer Society Grant [#215817 awarded to V.G.], Research Council of Norway projects [#300740 awarded to V.G., and #331890 awarded to V.G. and F.L.J.], a Research Council of Norway IKTPLUSS project [#311341 awarded to V.G.], GlaxoSmithKline (GSK), the Innovative Medicines Initiative 2 Joint Undertaking (Inno4Vac, supported by the European Union's Horizon 2020 research and innovation programme and EFPIA) [#101007799], and the European Union (ERC, AB-AG-INTERACT) [#101125630].

Ethics declarations

Competing interests: V.G. declares advisory board positions in aiNET GmbH, Enpicom B.V, Absci, Omniscope, and Diagonal Therapeutics. V.G. is a consultant for Adaptive Biosystems, Specifica Inc, Roche/Genentech, immunai, LabGenius, and FairJourney Biologics. J.R.J. is employed by GlaxoSmithKline plc. The remaining authors declare no competing interests.

References

1. Urquhart, L. Top drugs and companies by sales in 2018. *Nature Reviews Drug Discovery.* ISSN: 1474-1776 (Mar. 2019).
2. Lu, R. M. *et al.* Development of therapeutic antibodies for the treatment of diseases. *Journal of Biomedical Science 2020 27:1* **27,** 1–30. ISSN: 1423-0127 (1 Jan. 2020).
3. Paul, S. M. *et al.* How to improve RD productivity: the pharmaceutical industry's grand challenge. *Nature Reviews Drug Discovery 2010 9:3* **9,** 203–214. ISSN: 1474-1784 (3 Feb. 2010).
4. Whitehead, T. A. *et al.* Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. *Nature Biotechnology 2012 30:6* **30,** 543–548. ISSN: 1546-1696 (6 May 2012).
5. Mason, D. M. *et al.* Optimization of therapeutic antibodies by predicting antigen specificity from antibody sequence via deep learning. *Nature Biomedical Engineering 2021 5:6* **5,** 600–612. ISSN: 2157-846X (6 Apr. 2021).
6. Hie, B. L. *et al.* Efficient evolution of human antibodies from general protein language models. *Nature Biotechnology 2023,* 1–9. ISSN: 1546-1696 (Apr. 2023).
7. Shanehsazzadeh, A. *et al.* Unlocking de novo antibody design with generative artificial intelligence. *bioRxiv,* 2023.01.08.523187 (Mar. 2023).
8. Shuai, R. W. *et al.* Generative language modeling for antibody design. *bioRxiv,* 2021–12 (2021).
9. Solaiman, I. *et al.* OpenAI Report Release Strategies and the Social Impacts of Language Models (2019).
10. Kovaltsuk, A. *et al.* Observed Antibody Space: A Resource for Data Mining Next-Generation Sequencing of Antibody Repertoires. *The Journal of Immunology* **201,** 2502–2509. ISSN: 0022-1767 (8 Oct. 2018).
11. Olsen, T. H. *et al.* Observed Antibody Space: A diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. *Protein Science : A Publication of the Protein Society* **31,** 141. ISSN: 1469896X (1 Jan. 2022).

12. Olsen, T. H. *et al.* Addressing the antibody germline bias and its effect on language models for improved antibody design. *bioRxiv,* 2024.02.02.578678 (Feb. 2024).
13. Akbar, R. *et al.* In silico proof of principle of machine learning-based antibody design at unconstrained scale. *mAbs* **14.** ISSN: 19420870 (1 Dec. 2022).
14. Frey, N. C. *et al.* Protein Discovery with Discrete Walk-Jump Sampling (June 2023).
15. Li, L. *et al.* Machine learning optimization of candidate antibody yields highly diverse sub-nanomolar affinity antibody libraries. *Nature Communications 2023 14:1* **14,** 1–12. ISSN: 2041-1723 (1 June 2023).
16. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America* **89,** 10915. ISSN: 00278424 (22 Nov. 1992).
17. Olsen, T. H. *et al.* AbLang: an antibody language model for completing antibody sequences. *Bioinformatics Advances* **2.** ISSN: 26350041 (1 Jan. 2022).
18. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379.** Earlier versions as preprint: bioRxiv 2022.07.20.500902, 1123–1130 (2023).
19. Dauparas, J. *et al.* Robust deep learning–based protein sequence design using ProteinMPNN. *Science* **378,** 49–56 (2022).
20. Phillips, A. M. *et al.* Binding affinity landscapes constrain the evolution of broadly neutralizing anti-influenza antibodies. *eLife* **10.** ISSN: 2050084X (Sept. 2021).
21. Wang, C. *et al.* FLAML: A FAST AND LIGHTWEIGHT AUTOML LIBRARY (2020).
22. Satorras, V. G. *et al.* E(n) Equivariant Graph Neural Networks (2021).
23. Chinery, L. *et al.* Paragraph—antibody paratope prediction using graph neural networks with minimal feature vectors. *Bioinformatics* **39.** ISSN: 1367-4803 (1 Jan. 2023).
24. McDermott, M. B. A. *et al.* A Closer Look at AUROC and AUPRC under Class Imbalance (Jan. 2024).
25. Dunbar, J. & Deane, C. M. ANARCI: antigen receptor numbering and receptor classification. *Bioinformatics* **32,** 298. ISSN: 14602059 (2 Jan. 2016).

26. Porebski, B. T. *et al.* Rapid discovery of high-affinity antibodies via massively parallel sequencing, ribosome display and affinity screening. *Nature Biomedical Engineering 2023,* 1–19. ISSN: 2157-846X (Oct. 2023).

27. Ferrara, F. *et al.* Using Phage and Yeast Display to Select Hundreds of Monoclonal Antibodies: Application to Antigen 85, a Tuberculosis Biomarker. *PLOS ONE* **7,** e49535. ISSN: 1932-6203 (11 Nov. 2012).

28. Chao, G. *et al.* Isolating and engineering human antibodies using yeast surface display. *Nature Protocols 2006 1:2* **1,** 755–768. ISSN: 1750-2799 (2 July 2006).

29. Hummer, A. M. *et al.* Investigating the Volume and Diversity of Data Needed for Generalizable Antibody-Antigen G Prediction. *bioRxiv* (2023).

30. Schymkowitz, J. *et al.* The FoldX web server: An online force field. *Nucleic Acids Research* **33,** 382–388. ISSN: 03051048 (2005).

31. Cho, H.-S. *et al.* Structure of the extracellular region of HER2 alone and in complex with the Herceptin Fab. *Nature* **421,** 756–760 (2003).

32. Van Durme, J. *et al.* A graphical interface for the FoldX forcefield. *Bioinformatics* **27,** 1711–1712. ISSN: 1367-4803 (Apr. 2011).

33. Eddy, S. R. Where did the BLOSUM62 alignment score matrix come from? *Nature Biotechnology 2004 22:8* **22,** 1035–1036. ISSN: 1546-1696 (8 Aug. 2004).

34. Dunbar, J. *et al.* SAbDab: the structural antibody database.

35. Schneider, C. *et al.* SAbDab in the age of biotherapeutics: updates including SAbDab-nano, the nanobody structure tracker. *Nucleic Acids Research* **50,** D1368–D1372. ISSN: 0305-1048 (D1 Jan. 2022).

36. Abanades, B. *et al.* ImmuneBuilder: Deep-Learning models for predicting the structures of immune proteins. *bioRxiv,* 2022.11.04.514231 (Nov. 2022).

37. Lefranc, M. P. *et al.* IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Developmental and Comparative Immunology* **27,** 55–77. ISSN: 0145305X (1 Jan. 2003).

38. Bachas, S. *et al.* Antibody optimization enabled by artificial intelligence predictions of binding affinity and naturalness. *bioRxiv,* 2022.08.16.504181 (Aug. 2022).