Deep learning enables therapeutic antibody optimization in mammalian cells

Derek M Mason[1], Simon Friedensohn[1], Cédric R Weber[1], Christian Jordi[1], Bastian Wagner[1], Simon Meng[1] and Sai T Reddy[1]*

[1]Department of Biosystems Science and Engineering, ETH Zurich, 4058, Basel, Switzerland
Correspondence: *sai.reddy@ethz.ch

ABSTRACT

Therapeutic antibody optimization is time and resource intensive, largely because it requires low-throughput screening (10^3 variants) of full-length IgG in mammalian cells, typically resulting in only a few optimized leads. Here, we use deep learning to interrogate and predict antigen-specificity from a massive diversity of antibody sequence space. Using a mammalian display platform and the therapeutic antibody trastuzumab, rationally designed site-directed mutagenesis libraries are introduced by CRISPR/Cas9-mediated homology-directed repair (HDR). Screening and deep sequencing of relatively small libraries (10^4) produced high quality data capable of training deep neural networks that accurately predict antigen-binding based on antibody sequence (~85% precision). Deep learning is then used to predict millions of antigen binders from an *in silico* library of ~10^8 variants. Finally, these variants are subjected to multiple developability filters, resulting in tens of thousands of optimized lead candidates, which when a small subset of 30 are expressed, all 30 are antigen-specific. With its scalability and capacity to interrogate a vast protein sequence space, deep learning offers great potential for antibody engineering and optimization.

INTRODUCTION

In antibody drug discovery, the 'target-to-hit' stage is a well-established process, as screening hybridomas, phage or yeast display libraries typically result in a number of potential lead candidates. However, the time and costs associated with lead candidate optimization often take up the majority of the preclinical discovery and development cycle[1]. This is largely due to the fact that lead optimization of antibody molecules consists of addressing multiple parameters in parallel, including expression level, viscosity, pharmacokinetics, solubility, and immunogenicity[2,3]. Once a lead candidate is discovered, additional engineering is often required; phage and yeast display offer a powerful method for high-throughput screening of large mutagenesis libraries (>10^9), however they are primarily only used for increasing affinity or specificity to the target antigen[4]. The fact that nearly all therapeutic antibodies require expression in mammalian cells as full-length IgG means that the remaining development and optimization steps must occur in this context. Since mammalian cells lack the capability to stably replicate plasmids, this last stage of development is done at very low-throughput, as elaborate cloning, transfection and purification strategies must be implemented to screen libraries in the max range of 10^3, meaning only minor changes (e.g., point mutations) are screened[5]. Interrogating such a small fraction

of protein sequence space also implies that addressing one development issue will frequently cause rise of another or even diminish antigen binding altogether, making multi-parameter optimization very challenging.

Machine learning applied to biological sequence data offers a powerful approach to construct models capable of making predictions of genotype-phenotype relationships[6,7]. This is due to the capability of models to extrapolate complex relationships between sequence and function. One of the principle challenges in constructing accurate machine learning models is the collection of appropriate high-quality training data. Directed evolution platforms are well-suited for this as they rely on the linking of biological sequence data (DNA, RNA, protein) to a phenotypic output[8]. In fact, it has long been proposed to use machine learning models trained on data generated by mutagenesis libraries as a means to guide protein engineering[9,10]. Recently, Gaussian processes, a Bayesian learning model, was used to engineer cytochrome enzymes, enabling navigation through a vast protein sequence space to discover highly thermostable variants[11]. Similarly, the design and screening of a structure-guided library of channel rhodopsin membrane proteins was used to train Gaussian process and regression models, which were able to accurately predict variants that could express and localize on mammalian cell membranes[12].

In recent years, access to deep sequencing and parallel computing has enabled the construction of deep learning models capable of predicting molecular phenotype from sequence data[13,14]. For example, deep learning has been used to learn the sequence specificities of RNA- and DNA-binding proteins[15], regulatory grammar of protein expression in yeast[16], and HLA-neoantigen presentation on tumor cells[17]. In most cases deep (artificial) neural networks represent the class of algorithm utilized. While the complexity of neural networks has changed drastically since their conception, the fundamental concept remains the same: mimicking the connections of biological neurons to learn complex relationships between variables[18]. As an extension of a single-layer neural network, or perceptron[19], deep learning incorporates multiple hidden layers to deconvolute relationships buried in large, high-dimensional data sets, such as the millions of reads gathered from a single deep sequencing experiment. Well trained models can then be used to make predictions on completely unseen and novel variants. This application of model extrapolation lends itself perfectly to protein engineering because it provides a way to interrogate a much larger sequence space than what is physically possible. For example, even for a short stretch of just 10 amino acids, the combinatorial sequence diversity explodes to 10^{13}, a size which is nearly impossible to interrogate experimentally.

Here, we leverage the power of deep learning to perform multi-parameter optimization of therapeutic antibodies (full-length IgG) directly in mammalian cells (Figure 1). Starting with a mammalian display cell line[20] expressing the therapeutic antibody trastuzumab (Herceptin), we use CRISPR-Cas9-mediated homology-directed repair (HDR) to introduce site-directed mutagenesis libraries in the variable heavy chain complementarity determining region 3 (CDRH3)[21]. In order to generate information rich training data, single-site deep mutational scanning (DMS) is first performed[22], which is then used to guide the design of combinatorial mutagenesis libraries. An experimental (physical) library size of 5 x 10^4 variants was then screened for specificity to the antigen HER2. All binding and non-binding variant

sequences were then used to train recurrent and convolutional deep neural networks, which when fully-trained and optimized were able with high accuracy and precision to predict antigen-specificity based on antibody sequence. Neural networks are then used to predict antigen-specificity on a subset of sequence variants from the DMS-based combinatorial mutagenesis library (~10^8 sequences), resulting in >3.0 x 10^6 variants predicted to have a high probability of being antigen-specific. These variants are then subjected to several sequence-based in silico filtering steps to optimize for developability parameters such as viscosity, solubility and immunogenicity, resulting in over 40,000 optimized antibody sequence variants. Finally, a random selection of variants were recombinantly expressed and tested, resulting in 30 out of 30 showing antigen-specific binding.

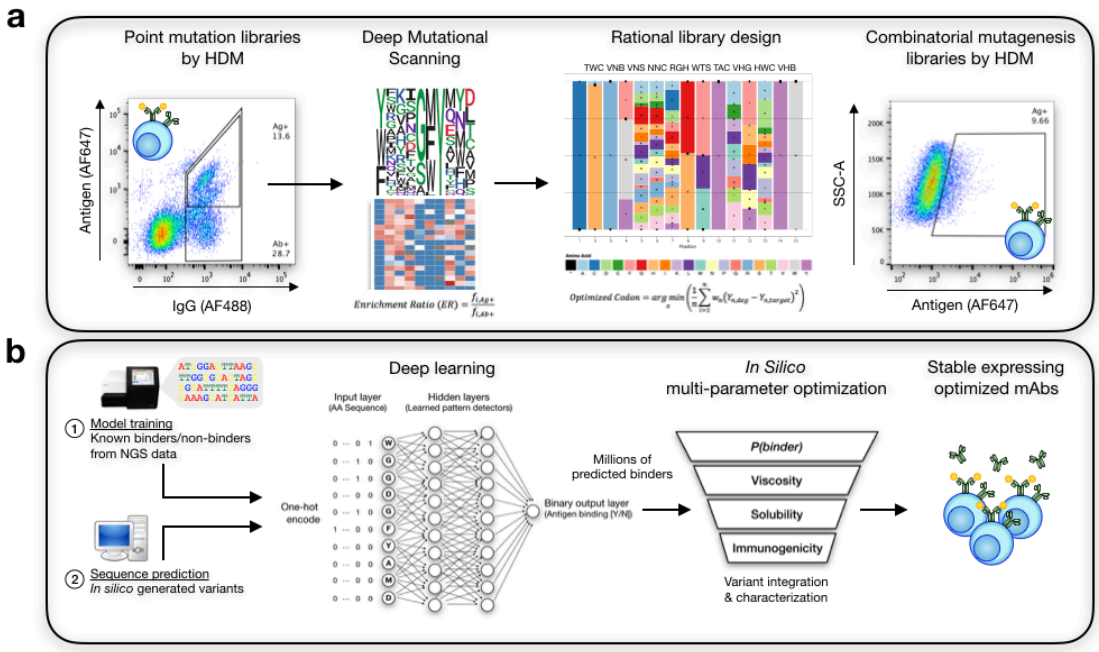


Figure 1: Implementing deep learning to predict antibody target specificity

(a) Generating quality data capable of training accurate models. First, deep mutational scanning assesses the impact mutations have on protein function across many different positions. These insights can then be applied to combinatorial mutagenesis strategies to guide protein library design capable of producing thousands of binding variants. **(b)** Sequence information for binders and non-binders can then be used to train deep neural networks to accurately predict antigen specificity of unknown antibody variants, producing millions of predicted binders. These binders can then be subjected to any available in silico methods for predicted various developability attributes.

RESULTS

Deep mutational scanning determines antigen-specific sequence landscapes and guides rational antibody library design

As the amino acid sequence of an antibody's CDRH3 is a key determinant of antigen specificity, we performed DMS on this region to resolve the specificity determining residues. To start, a hybridoma cell-line was used that expressed a trastuzumab variant that could not bind HER2 antigen (mutated

CDRH3 sequence) (Supplementary Fig. 1). Libraries were generated by CRISPR-Cas9-mediated homology-directed mutagenesis (HDM)[21], which utilized guide RNA (gRNA) for Cas9 targeting of CDRH3 and a pool of homology templates in the form of single-stranded oligonucleotides (ssODNs) containing NNK degenerate codons at single-sites tiled across CDRH3 (Figure 2a, Supplementary Fig. 2). Libraries were then screened by fluorescence activated cell sorting (FACS), and populations expressing surface IgG which either were binding or not binding to antigen were isolated and subjected to deep sequencing (Illumina MiSeq) (Supplementary Table 1). Deep sequencing data was then used to calculate enrichment scores of the 10 positions investigated, which revealed six positions that were sufficiently amenable to a wide-range of mutations and an additional three positions that were marginally accepting to defined mutations (Figure 2b). Although residues 102D, 103G, 104F, and 105Y appear to be contacting amino acids of the CDRH3 loop with HER2[23,24], 105Y is the only residue completely fixed (Figure 2c).

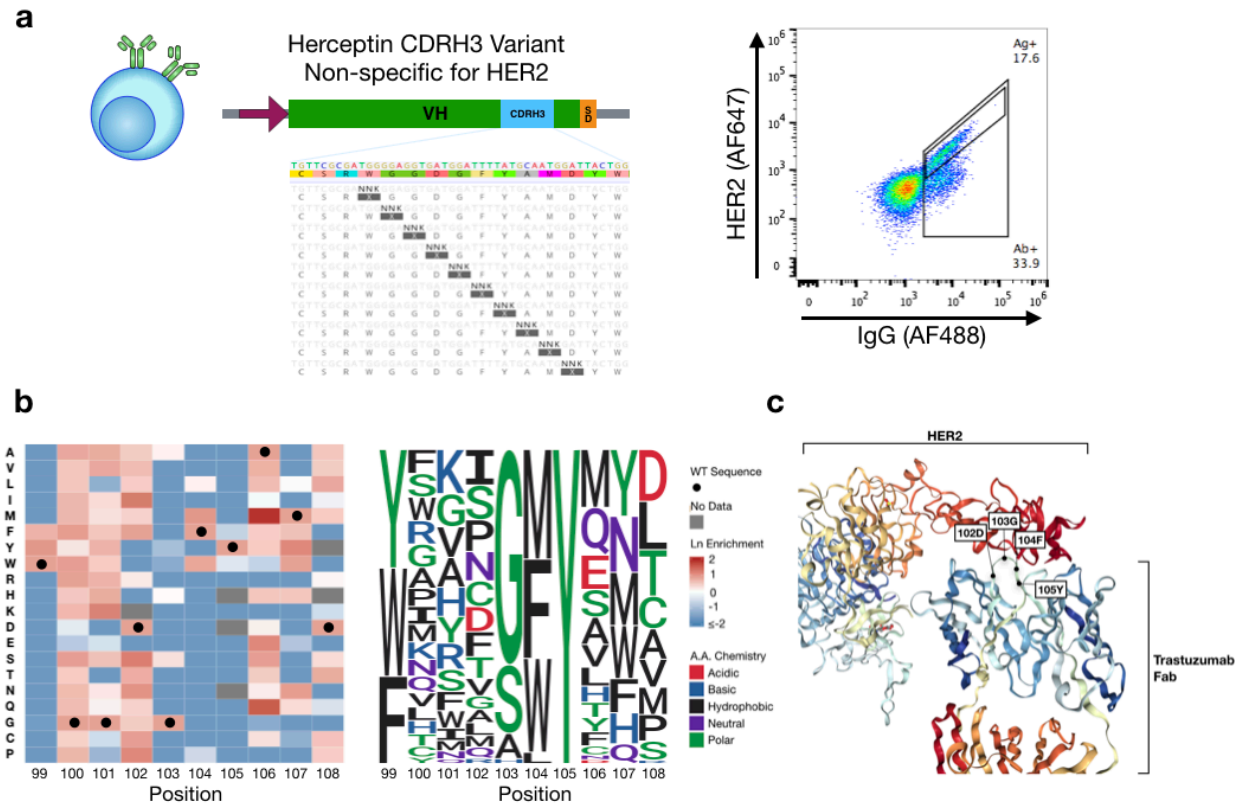


Figure 2: Deep mutational scanning reveals specificity determining residues

(a) Flow cytometry profile following integration of tiled mutations by homology-directed mutagenesis. Antigen specific variants underwent 3 rounds of enrichment (Supplementary Fig. 2) **(b)** Corresponding heatmap (left) following sequencing analysis of the pre-sorted (Ab+) and post-sorted (Ag+) populations (Supplementary Table 1). Wild type amino acids are marked by black circles. The resulting sequence logo plot (right) generated by positively enriched mutations per position. **(c)** 3D protein structure of trastuzumab in complex with its target antigen, HER2[23,24]. Locations of surface exposed residues: 102D, 103G, 104F, and 105Y are given[40].

Heatmaps and their corresponding sequence logo plots generated by DMS were used to guide the rational design of combinatorial mutagenesis libraries, which consisted of degenerate codons across all positions (except 105Y) (Supplementary Fig. 3, Supplementary Table 6). Degenerate codons were selected per position based on their amino acid frequencies which most closely resembled the degree of enrichment found in the DMS data following 1, 2, and 3 rounds of antigen-specific enrichment (Supplementary Fig. 2, Equation 2). This combinatorial library possesses a theoretical protein sequence space of 7.17 x 10^8, far greater than the single-site DMS library diversity of 200. Libraries containing CDRH3 variants were again generated in hybridoma cells through CRISPR-Cas9-mediated HDM in the same non-binding trastuzumab clone described previously (Figure 3a). Antigen binding cells were isolated by two rounds of enrichment by FACS (Figure 3b, Supplementary Fig. 3) and the binding/non-binding populations were subjected to deep sequencing. Sequencing data identified 11,300 and 27,539 unique binders and non-binders, respectively (Supplementary Table 2). These sequence variants represented only a miniscule 0.0054% of the theoretical protein sequence space of the combinatorial mutagenesis library. Amino acid usage per position was comparatively similar between antigen binding and non-binding populations (Figure 3c), thus making it difficult to develop any sort of heuristic rules or observable patterns to identify binding sequences.

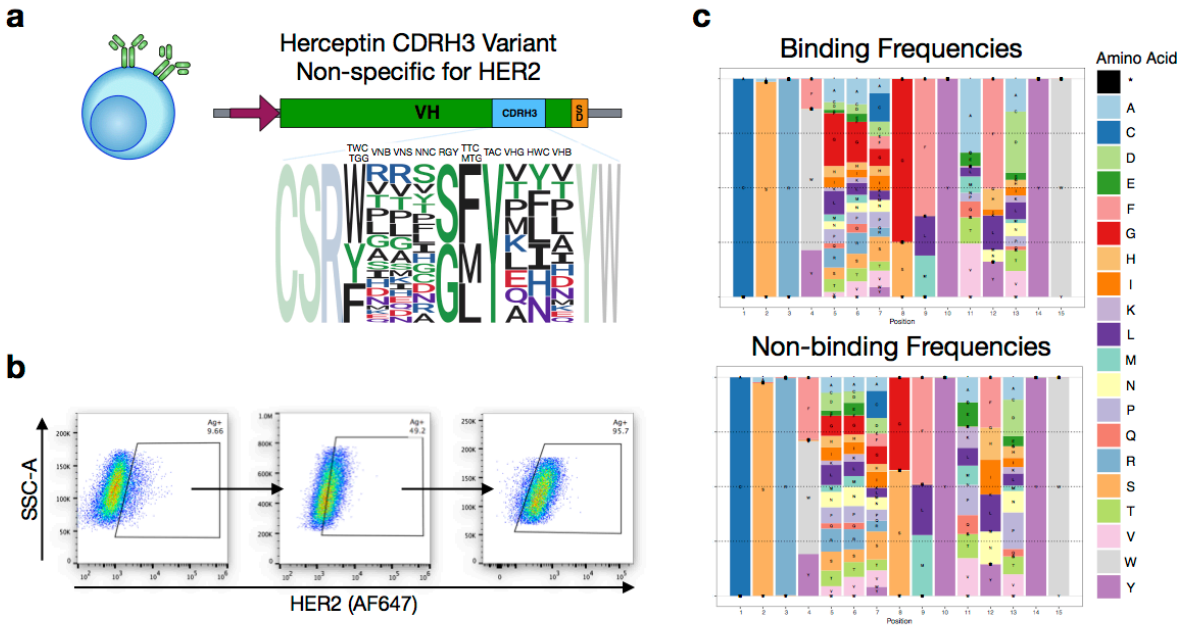


Figure 3: Combinatorial mutagenesis libraries generate data enriched with binding variants

(a) Combinatorial mutagenesis libraries are designed from enrichment ratios observed in DMS data and integrated into the trastuzumab variant by homology-directed mutagenesis. **(b)** Flow cytometry plots resulting from transfection of a rationally designed library. Two rounds of enrichment were performed to produce a library of antigen specific variants. Deep sequencing was performed on the library (Ab+), non-binding variants (Ag-), and binding variants after 1 and 2 rounds of enrichment (Ag+1, Ag+2) (Supplementary Fig. 3, Supplementary Table 2). **(c)** Amino acid frequency plots of antigen binding variants and non-binding variants reveals nearly indistinguishable amino acid usages across all positions.

Training deep neural networks to classify antigen-specificity based on antibody sequence

After having compiled deep sequencing data on binding and non-binding CDRH3 variants, we set out to develop and train deep learning models capable of predicting specificity towards the target antigen HER2. Amino acid sequences were converted to an input matrix by one-hot encoding, an approach where each column of the matrix represents a specific residue and each row corresponds to the position in the sequence, thus a 10 amino acid CDRH3 sequence as here results in a 10 x 20 matrix. Each row will contain a single '1' in the column corresponding to the residue at that position, whereby all other columns/rows receive a '0'. We utilized long short-term memory recurrent neural networks (LSTM-RNN) and convolutional neural networks (CNN), which represent two of the main classes of deep learning models used for biological sequence data[14]. LSTM-RNNs and CNNs both stem from standard neural networks, where information is passed along neurons that contain learnable weights and biases, however, there are fundamental differences in how the information is processed. LSTM-RNN layers contain loops, enabling information to be retained from one step to the next, allowing models to efficiently correlate a sequential order with a given output; CNNs, on the other hand, apply learnable filters to the input data, allowing it to efficiently recognize spatial dependencies associated with a given output. Model architecture and hyperparameters (Figures 4a, c) were selected by performing a grid search across various parameters (LSTM-RNN: nodes per layer, batch size, number epochs and optimizing function; CNN: number of filters, kernel size, dropout rate and dense layer nodes) using a k-fold cross-validation of the data set. All models were built to assess their accuracy and precision of classifying binders and non-binders from the available sequencing data. 70% of the original data set was used to train the models and the remaining 30% was split into two test data sets used for model evaluation: one test data set contained the same class split of sequences used to train the model and the other contained a class split of approximately 10/90 binders/non-binders to resemble physiological frequencies (Figure 3b). Performance of the LSTM-RNN and CNN were assessed by constructing receiver operating characteristic (ROC) curves and precision-recall (PR) curves derived from predictions on the unseen testing data sets (Figure 4b, d). Based on conventional approaches to training classification models, the data set was adjusted to allow for a 50/50 split of binders and non-binders during training. Under these training conditions, the LSTM-RNN and CNN were both able to accurately classify unseen test data (ROC curve AUC: 0.9 ± 0.0, average precision: 0.9 ± 0.0, Supplementary Fig. 6).

Next, we used the trained LSTM-RNN and CNN models to classify a random sample of 1 x 10^5 sequences from the potential sequence space. We observed, however, an unexpectedly high occurrence of positive classifications (25,318 ± 1,643 sequences or 25.3 ± 1.6%, Supplementary Table 3b). With the knowledge that the physiological frequency of binders should be approximately 10-15%, we sought to adjust the classification split of the training data with the hypothesis that models were being subject to some unknown classification bias. Additional models were then trained on classification splits of both 20/80, and 10/90 binders/non-binders, as well as a classification split with all available data (approximately 30/70 binders/non-binders). Unbalancing the sequence classification led to a significant reduction in the percentage of sequences classified as binders, but also led to a reduction in the model performance on the unseen test data (Supplementary Fig. 4-7, Supplementary Tables 3a,

b). Through our analysis, we concluded that the optimal data set for training the models was the set inclusive of all known CDRH3 sequences for the following reasons: 1) the percentage of sequences predicted as binders reflects this physiological frequency, 2) this data set maximizes the information the model sees, and 3) model performance on both test data sets. Final model architecture, parameters, and evaluation are shown in Figure 4. As a final measure of model validation, neural networks were trained with a data set containing randomly shuffled binding and non-binding class labels. Model performance of these networks revealed indiscriminate sequence classification on unseen test data (Supplementary Fig. 8), signifying the identification of learned patterns for networks trained with properly classified data.

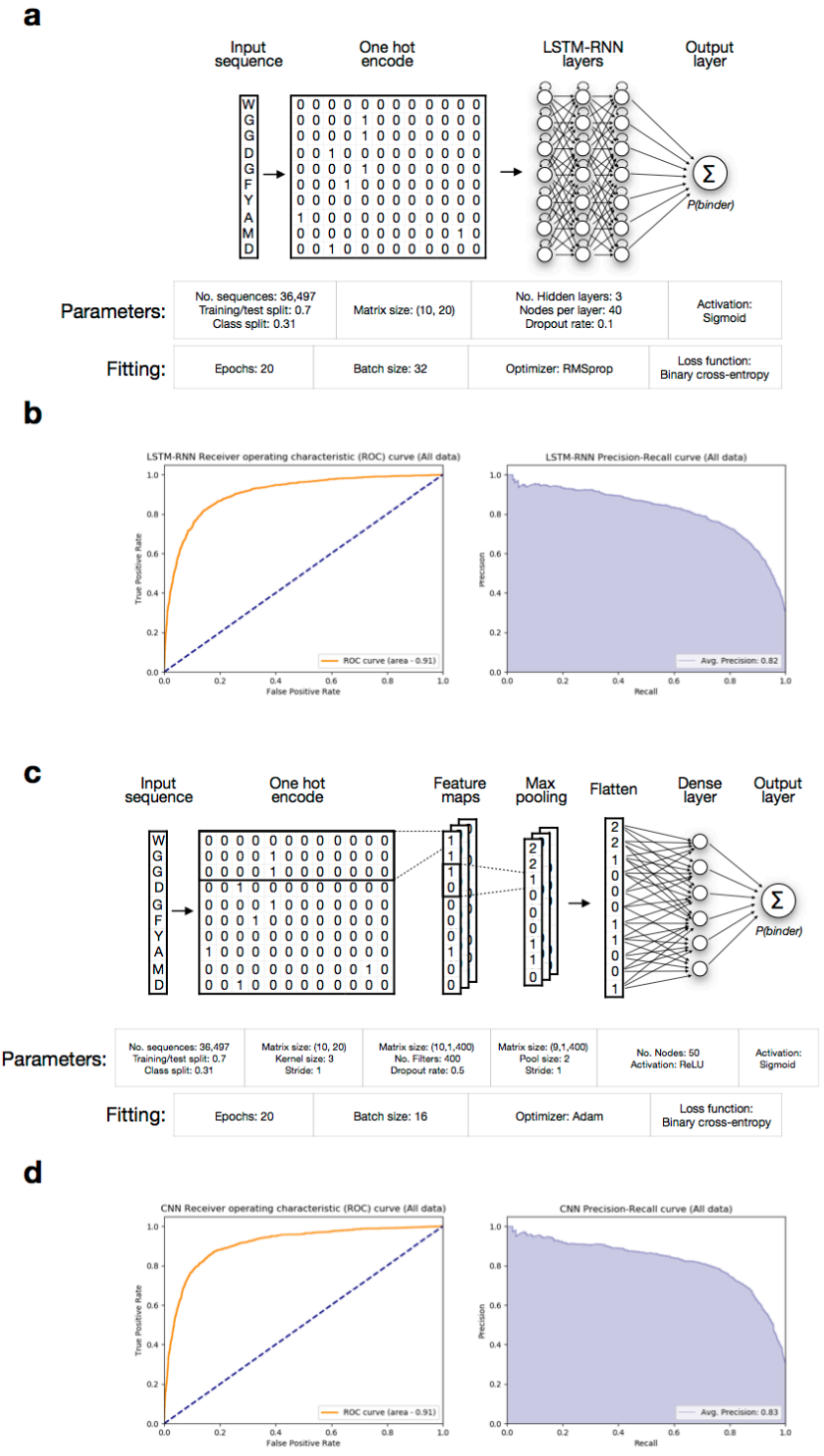


Figure 4: Deep learning models accurately predict antigen specificity

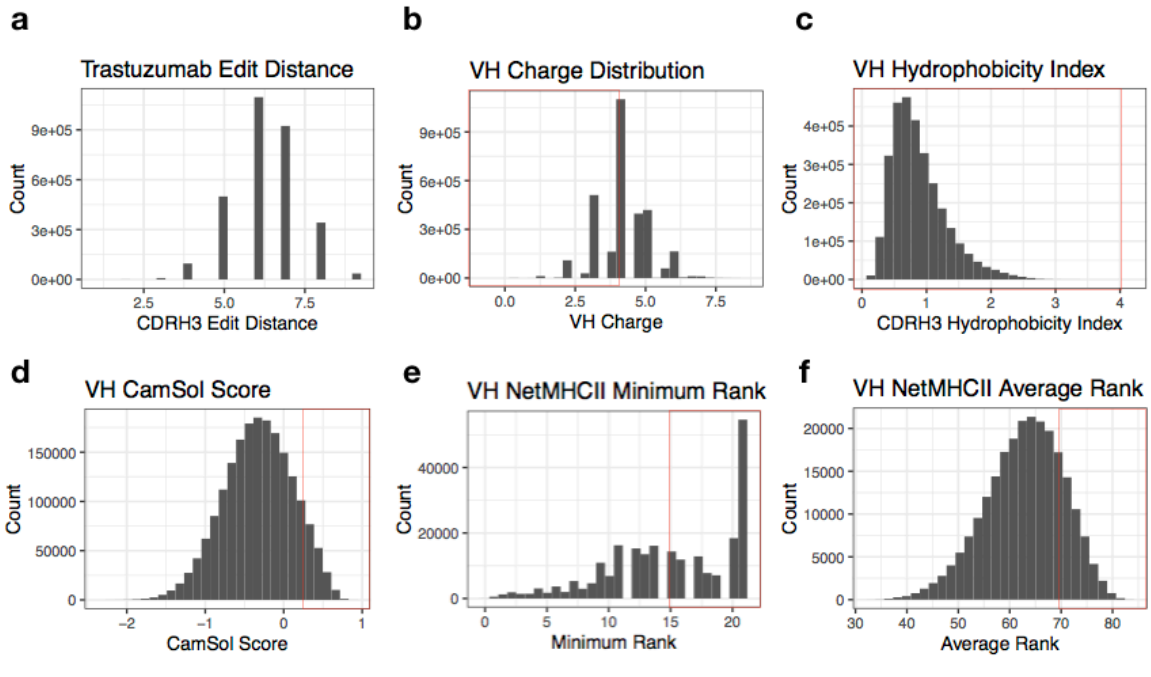
The selected network architectures and their model performance curves for classification of binding and non-binding sequences. Model training was performed on 70% of the data and testing was performed by withholding the remaining 30% and then comparing the model's classification of test sequences with the known classification. In lieu of adjusting the data set to a defined class split of binding/non-binding sequences, all known information was utilized to train and test the networks (approx. class split of 31%). (**a**) LSTM-RNN architecture and parameters used for model fitting. (**b**) ROC (receiver operating character) curve and PR (precision-recall) curve observed on the classification of sequences in the test set by the LSTM-RNN. (**c**) CNN architecture and parameters used for model fitting. (**d**) ROC curve and PR curve observed on the classification of sequences in the test set by the CNN. The high values observed for the ROC area under curve (AUC) and average precision of both networks represent robust measures of model accuracy and precision.

175

Multi-parameter optimization for developability by in silico screening of antibody sequence space

Using our DMS-based combinatorial mutagenesis library as a guide (Figure 3), 7.2 x10^7 possible sequence variants were generated in silico. The fully-trained LSTM-RNN and CNN models were used to classify all 7.2 x 10^7 sequence variants as either antigen binders or non-binders based on a probability score (*P*), resulting in a prediction of 8.55 x 10^6 (LSTM-RNN) and 9.52 x 10^6 (CNN) potential binders (*P > 0.50)*. This represented a reasonable fraction (11-13%) of antigen-specific variants based on experimental screening (Figure 3b). To increase confidence, we increased the prediction threshold for binder classification to *P* > 0.75 and took the consensus binders between the LSTM-RNN and CNN. This reduced the antigen-specific sequence space down to 3.0 x 10^6 variants.

Next, we characterized the 3.0 x 10^6 predicted antigen-specific sequences on a number of parameters. As a first metric, we investigated their sequence similarity to the original trastuzumab sequence by calculating the Levenshtein distance (*LD*). The majority of sequences showed an edit distance of *LD* > 4 (Figure 5a). The first step in filtering was to calculate the net charge and hydrophobicity index in order to estimate the molecule's viscosity and clearance[2]. According to Sharma et al., viscosity decreases with increasing variable fragment (Fv) net charge and increasing Fv charge symmetry parameter (FvCSP); however, the optimal Fv net charge in terms of drug clearance is between 0 and 6.2 with a CDRL1+CDRL3+CDRH3 hydrophobicity index sum < 4.0. Based on the wide range of values for these parameters in the 3.0 x 10^6 predicted variants (Figure 5b, c), we filtered any sequences out that had a Fv net charge > 4.2 and a CDRH3 hydrophobicity index > 4.0, which further reduced the sequence space down to 1.93 x 10^6 variants. We next padded the CDRH3 sequences with 10 amino acids on the 5' and 3' ends and then ran these sequences through CamSol, a protein solubility predictor developed by Sormanni et al.[25], which estimates and ranks sequence variants based on their theoretical solubility. The remaining variants produced a wide-range of protein solubility scores (Figure 5d) and sequences with a score < 0.2 were filtered out, leaving 2.36 x 10^5 candidates for further analysis. As a last step in our *in silico* screening process, we aimed at reducing immunogenicity by predicting the peptide binding affinity of the variant sequences to MHC Class II molecules by utilizing NetMHCIIpan, a model previously developed by Jensen et al.[26]. All possible 15-mers from the padded CDRH3 sequences were run through NetMHCIIpan. One output from the model is a given peptide's % Rank of predicted affinity compared to a set of 200,000 random natural peptides. Typically, molecules with a % Rank < 2 are considered strong binders and those with a % Rank < 10 are considered weak binders to the MHC Class II molecules scanned. After predicting affinity for HLA alleles DRB1*0101, DRB3*0101, DRB4*0101, DRB5*0101, sequences were filtered out if any of the 15-mers contained a % Rank < 15 (Figure 5e). The average % Rank across all 15-mers for the remaining sequences was then calculated and those with an average % Rank < 70 were also filtered out (Figure 5f). Based on these criteria, there were 40,588 multi-parameter optimized variants (Figure 5g).



g

Filtering Parameters	No. Sequences	Filtering Strategy
RNN *P*(binder) > 0.5 CNN *P*(binder) > 0.5	8,552,268 9,519,642	RNN & CNN Output
RNN *P*(binder) > 0.75 CNN *P*(binder) > 0.75	4,315,323 5,218,706	*P*(binder)
RNN, CNN consensus	3,002,476	RNN & CNN Consensus
Net charge < 4.2 Hydrophobicity Index < 4.0	1,927,942	Net Charge & Hyd. Index
Solubility score > 0.2	236,066	CamSol Score
Minimum % Rank > 15 Average % Rank > 70 (Across all 15-mers)	40,588	NetMHCII

Figure 5: In silico screening of predicted binders produces multi-parameter optimized variants

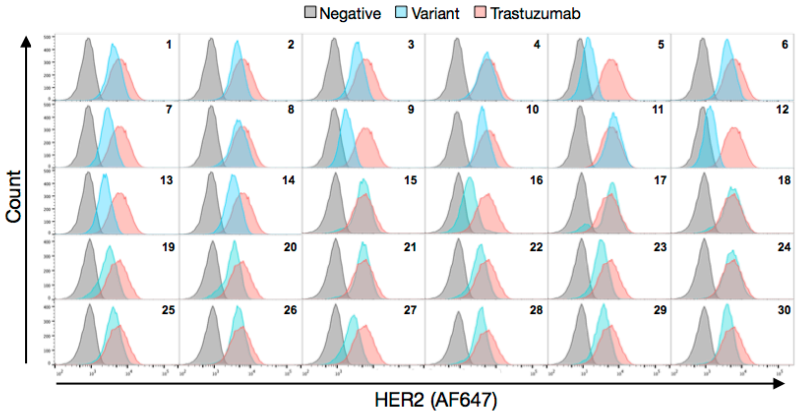
Antigen specific predictions yield variants with a wide range In silico calculated parameters for developability. The following histograms show the parameters distributions of all predicted variants at the different stages of filtering. Red boxes indicate filtering cut-offs. **(a)** Levenshtein distance from wild-type trastuzumab. **(b)** Net charge of the VH domain. **(c)** CDRH3 hydrophobicity index. **(d)** CamSol intrinsic solubility score. **(e)** Minimum NetMHCIIpan % Rank across all possible 15-mers. **(f)** Average NetMHCIIpan % Rank across all possible 15-mers. **(g)** Filtering parameters and the number of sequences at the corresponding stage of filtering.

Optimal antibody sequences are recombinantly expressed and antigen-specific

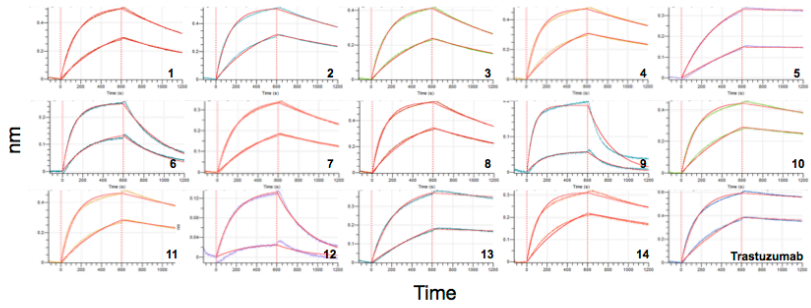
To validate the precision of our fully trained LSTM-RNN and CNN models, we randomly selected a subset of 30 CDRH3 sequences predicted to be antigen-specific and optimized across the multiple developability parameters. To further demonstrate the capacity of deep learning to identify novel sequence variants, we also added the criteria that the selected variants must have a minimum *LD* of 5 from the original CDRH3 sequence of trastuzumab, resulting in a library of 32,725 sequences to select from. CRISPR-Cas9-mediated HDR was used to generate mammalian display cell lines expressing the 30 different sequence variants. Flow cytometry was performed and revealed that 30 of the 30 variants

(100%) were antigen-specific (Figure 6a). Further analysis was performed on 14 of the antigen-binding variants to more precisely quantify the binding kinetics via biolayer interferometry (BLI, FortéBio Octet RED96e) (Figure 6b). The original trastuzumab sequence was measured to have an affinity towards HER2 of 4.0 x 10^{-10} M (equilibrium dissociation constant, K_D); and although the majority of variants tested had a slight decrease in affinity, 71% (10/14) were still in the single-digit nanomolar range, 21% (3/14) remained sub-nanomolar, and one variant (7%) showed a near 3-fold increase in affinity compared to trastuzumab (K_D = 1.4 x 10^{-10} M). We also investigated any correlations between flow cytometry fluorescence intensity and BLI measured affinity (Supplementary Fig. 9), as well as model prediction values and measured affinities (Supplementary Fig. 10). While there appears to be an overall increasing trend between fluorescence intensity and binding affinity, there also exists outlying points with low fluorescence signals, but high affinity values. Conversely, no observable trend is present when comparing model prediction values to binding affinities, however, the highest affinity variants do tend to have higher prediction values. Figure 6c displays the 30 tested sequence variants along with their associated developability and affinity metrics.

a



b



c

Variant	CDRH3 Sequence	RNN *P*(Binder)	CNN *P*(Binder)	LD	Net Charge	FvCSP	HI Sum	CamSol Score	NetMHCII Minimum	NetMHCII Average	KD (nM)	ka (1/Ms)	kd (1/s)
1	WQEAAFYAHD	0.833	0.930	5	6.02	9.06	1.07	0.45	21	63.4	3.31	2.22E+05	7.34E-04
2	YGGRGMYQLE	0.906	0.973	6	7.22	12.71	0.81	0.49	21	68.1	1.74	2.90E+05	5.02E-04
3	WNQLGMYVND	0.874	0.888	6	6.19	9.58	1.02	0.37	11	54.9	3.36	2.23E+05	7.47E-04
4	YGMRGLYAYE	0.942	0.948	6	7.22	12.71	0.92	0.48	13	50.5	1.47	3.10E+05	4.55E-04
5	WMETGFYTHD	0.895	0.966	5	6.02	9.06	0.97	0.38	20	66.9	0.14	1.06E+05	1.45E-05
6	WNAPSFYAND	0.885	0.977	5	6.19	9.58	1.02	0.57	17	64.4	7.43	3.02E+05	2.25E-03
7	WQGIGLYELD	0.942	0.889	5	5.25	6.70	1.28	0.46	12	60.9	3.20	1.95E+05	6.24E-04
8	WAMLGMYAHD	0.847	0.779	5	6.97	11.94	1.33	0.24	10	44.1	2.32	3.01E+05	6.96E-04
9	YQANGLYAYE	0.959	0.886	7	6.22	9.67	0.98	0.42	14	55.8	9.77	4.34E+05	4.24E-03
10	YRAVGFYTND	0.783	0.871	6	7.19	12.62	0.95	0.43	9	51.2	0.94	2.56E+05	2.41E-04
11	WAPYGLYAHD	0.948	0.959	5	6.97	11.94	1.19	0.41	19	59.0	1.28	2.92E+05	3.72E-04
12	WDGPAFYELD	0.860	0.936	5	4.27	3.73	1.13	0.53	13	70.2	31.1	1.03E+05	3.21E-03
13	WGIHSFYEHD	0.844	0.869	5	6.80	11.42	1.09	0.42	20	64.2	0.54	1.79E+05	9.63E-05
14	YGEYGMYVNK	0.888	0.870	7	7.22	12.71	0.85	0.47	21	65.0	1.24	3.23E+05	4.01E-04
15	WRDRGFYEYD	0.858	0.974	5	6.27	9.81	0.78	0.66	14	67.4	n.d.	n.d.	n.d.
16	WEEYGLYVHD	0.933	0.992	6	5.07	6.18	1.04	0.37	12	68.3	n.d.	n.d.	n.d.
17	YASAGMYTHD	0.927	0.883	7	6.97	11.94	0.89	0.54	19	61.7	n.d.	n.d.	n.d.
18	YGDAGMYALK	0.973	0.995	6	7.19	12.62	1.04	0.55	19	63.4	n.d.	n.d.	n.d.
19	WQLGGMYTHD	0.919	0.941	6	6.97	11.94	0.98	0.41	21	67.4	n.d.	n.d.	n.d.
20	WNSDGLYAYE	0.864	0.961	5	5.24	6.70	0.98	0.51	3	67.2	n.d.	n.d.	n.d.
21	WQRGGFYVND	0.956	0.993	5	7.19	12.62	0.94	0.47	9	64.1	n.d.	n.d.	n.d.
22	YGARGFYQND	0.892	0.789	5	7.19	12.61	0.82	0.69	13	65.2	n.d.	n.d.	n.d.
23	YAGPGMYTNQ	0.870	0.830	7	7.17	12.55	0.80	0.46	20	74.1	n.d.	n.d.	n.d.
24	WNPHGLYVND	0.939	0.974	6	6.97	11.93	1.01	0.49	13	65.9	n.d.	n.d.	n.d.
25	YGSNGLYANQ	0.914	0.908	6	7.17	12.55	0.90	0.51	21	70.4	n.d.	n.d.	n.d.
26	WPKVGLYTND	0.853	0.865	6	7.19	12.62	0.99	0.62	15	64.5	n.d.	n.d.	n.d.
27	WGIVSFYEND	0.871	0.873	5	5.24	6.70	1.24	0.20	17	59.6	n.d.	n.d.	n.d.
28	YSMPGMYTNA	0.848	0.938	8	7.17	12.55	0.85	0.37	21	62.8	n.d.	n.d.	n.d.
29	WAEAGMYEFD	0.880	0.915	6	4.30	3.83	1.12	0.48	17	65.6	n.d.	n.d.	n.d.
30	WPMCGMYTHD	0.835	0.855	6	6.97	11.93	0.88	0.30	21	69.5	n.d.	n.d.	n.d.
Trastuzumab	WGGDGFYAMD	0.962	0.938	0	5.21	6.61	0.92	0.49	11	68.3	0.40	2.51E+05	1.01E-04

Figure 6: Neural network predicted sequences are experimentally validated to be antigen-specific

30 variants were randomly selected and integrated into individual hybridoma cells lines by separately transfecting ssODN donor sequences with gRNA. **(a)** Out of the 30 sequences selected and integrated, all 30 bound the target antigen indicated by flow cytometry. **(b)** Affinities for 14 of the 30 variant sequences were determined by biolayer interferometry (BLI). The majority of sequences measured exude affinities in the single nano molar or sub-nanomolar range **(c)** A final table of the 30 variants randomly selected with their developability parameters. Values are shaded green to red according to their measure of developability. (n.d., not determined).

235

DISCUSSION

Addressing the limitation of antibody optimization in mammalian cells, we have developed an approach based on deep learning that enables us to identify antigen-specific sequences with high precision. Using the clinically approved antibody trastuzumab, we performed single-site DMS followed by combinatorial mutagenesis to determine the antigen-binding landscape of CDRH3. This DMS-based mutagenesis strategy is crucial for attaining high quality training data that is enriched with antigen-binding variants, in this case nearly 10% of our library (Figure 3b). In contrast, if a completely randomized combinatorial mutagenesis strategy was employed (i.e., NNK degenerate codons), it would be unlikely to produce any significant fraction of antigen-binding variants. In the future, other approaches to mutagenesis that generate enriched training data[27], such as shotgun scanning mutagenesis[28], binary substitution[29] and recombination[30,12] may also be explored for training deep neural networks.

A remarkable finding in this study was that experimental screening of a library of only 5×10^4 variants, which reflected a tiny fraction (0.0054%) of the total sequence diversity of the DMS-based combinatorial mutagenesis library (7.17×10^8), was capable of training accurate neural networks. This suggests that physical library size limitations of mammalian expression systems (or other expression platforms such as phage and yeast) and deep sequencing read depth will not serve as a limitation in deep learning-guided protein engineering. Another important result was that deep sequencing of antigen-binding and non-binding populations showed nearly no observable difference in their positional amino acid usage (Figure 3c), suggesting that neural networks are effectively capturing non-linear patterns/interactions.

In the current study, we selected LSTM-RNNs and CNNs as the basis of our classification models, as they represent two state-of-the-art approaches in deep learning. Other machine learning approaches such as k-nearest neighbors, random forests, and support vector machines are also well-suited at identifying complex patterns from input data, but as data set sizes continue to grow, as is realizable with biological sequence data, deep neural networks tend to outperform these classical techniques[15]. Furthermore, deep generative modeling methods such as variational autoencoders may also be used to explore the mutagenesis sequence space from directed evolution[31].

We *in silico* generated approximately 7.2×10^7 CDRH3 variants from DMS-based combinatorial diversity and used fully trained LSTM-RNN and CNN models to classify each sequence as a binder or non-binder. The 7.2×10^7 sequence variants comprise only a subset of the potential sequence space and was chosen to minimize the computational effort, however, it still represents a library size several orders of magnitude greater than what is experimentally achievable in mammalian cells. We easily envision extending the screening capacity through script optimization and employing parallel computing on high performance clusters. Out of all variants classified, the LSTM-RNN and CNN predicted approximately 11-13% to bind the target antigen, showing exceptional agreement with the experimentally observed frequencies by flow cytometry (Figure 3b). With the exception of critical residues determined by DMS, the majority of predicted binders were substantially distant from the original trastuzumab sequence with 80% of sequences having an edit distance of at least 6 residues. This high degree of sequence variability indicated the potential for a wide range of biomolecular properties.

Once an antibody's affinity for its target antigen is within a desirable range for efficacious biological modification, addressing other biomolecular properties becomes the focus of antibody development. With recent advances in computational predictions[32,33], a number of these properties, including viscosity, clearance, stability[2], specificity[34], solubility[25] and immunogenicity[26] can be approximated from sequence information alone. With the aim of selecting antibodies with improved characteristics, we subjected the library of predicted binders to a number of these in silico approaches in order to provide a ranking structure and filtering strategy for developability (Figure 5). After implementing these methods to remove variants with a high likelihood of having poor viscosity, clearance or solubility, as well as those with high immunogenic potential, over 40,000 multi-parameter optimized antibody variants remained. It is interesting to note that a considerable number of sequences scored even better than the original trastuzumab sequence. Future work to apply more stringent or additional filters which address other developability parameters (e.g. stability, specificity, humanization) could also be implemented to further reduce the sequence space down to highly developable therapeutic candidates. For instance, previous studies have investigated the likeness of therapeutic antibodies to the human antibody repertoire[35].

Lastly, to experimentally validate the precision of neural networks to predict antigen specificity, we randomly selected and expressed 30 variants from the library of optimized sequences with a minimum edit distance of 5 from trastuzumab. The precision of the LSTM-RNN and CNN models were each estimated to be ~85% (at $P > 0.75$) according to predictions made on the test data sets (Figure 4b, d). By taking the consensus between models, however, we experimentally validated that all randomly selected (30/30) of the antigen-predicted (and developability filtered) sequences were indeed binders, and several of which were high affinity. While we anticipate false positives would be discovered by increasing the sample size tested, validation of this subset strongly infers that potentially thousands of optimized lead candidates maintain a binding affinity in the range of therapeutic relevance, while also containing substantial sequence variability from the starting trastuzumab sequence. Future work to increase the stringency of selection during screening or a more detailed investigation of correlations between prediction probability and affinity could prove insightful towards retaining high target affinities. We also envision this approach to enable the optimization of other functional properties of therapeutic antibodies, such as pH-dependent antibody recycling[36] or affinity/avidity tuning[37,38]. Additionally, extending this approach to other regions across the variable light and heavy chain genes, namely other CDRs, may yield deep neural networks that are able to capture long-range, complex relationships between an antibody and its target antigen. To understand these patterns in greater depth, it may also prove useful to compare neural network predictions with protein structural modeling predictions[39].

METHODS

Mammalian cell culture and transfection

Hybridoma cells were cultured and maintained according to the protocols described by Mason et al.[21]. Hybridoma cells were electroporated with the 4D-Nucleofector™System (Lonza) using the SF Cell Line 4D-Nucleofector® X Kit L or X Kit S (Lonza, V4XC-2024, V4XC-2032) with the program CQ-104. Cells

were prepared as follows: cells were isolated and centrifuged at 125 x G for 10 minutes, washed with Opti-MEM® I Reduced Serum Medium (Thermo, 31985-062), and centrifuged again with the same parameters. The cells were resuspended in SF buffer (per kit manufacturer guidelines), after which Alt-R gRNA (IDT) and ssODN donor (IDT) were added. All experiments performed utilize constitutive expression of Cas9 from *Streptococcus pyogenes* (SpCas9). Transfections of $1x10^6$ and $1x10^7$ cells were performed in 100 μl, single Nucleocuvettes™ with 0.575 or 2.88 nmol Alt-R gRNA and 0.5 or 2.5 nmol ssODN donor respectively. Transfections of $2x10^5$ cells were performed in 16-well, 20 μl Nucleocuvette™ strips with 115 pmol Alt-R gRNA and 100 pmol ssODN donor.

Flow cytometry analysis and sorting

Flow cytometry-based analysis and cell isolation were performed using the BD LSR Fortessa™ (BD Biosciences) and Sony SH800S (Sony), respectively. When labeling with fluorescently conjugated antigen or anti-IgG antibodies, cells were first washed with PBS, incubated with the labeling antibody and/or antigen for 30 minutes on ice, protected from light, washed again with PBS and then analyzed or sorted. The labeling reagents and working concentrations are described in Supplementary Table 4**.** For cell numbers different from 10^6, the antibody/antigen amount and incubation volume were adjusted proportionally.

Sample preparation for deep sequencing

Sample preparation for deep sequencing was performed similar to the antibody library generation protocol of the primer extension method described previously[41]. Genomic DNA was extracted from 1-$5x10^6$ cells using the Purelink™ Genomic DNA Mini Kit (Thermo, K182001). Extracted genomic DNA was subjected to a first PCR step. Amplification was performed using a forward primer binding to the beginning of the VH framework region and a reverse primer specific to the intronic region immediately 3' of the J segment. PCRs were performed with Q5® High-Fidelity DNA polymerase (NEB, M0491L) in parallel reaction volumes of 50 ml with the following cycle conditions: 98˚C for 30 seconds; 16 cycles of 98˚C for 10 sec, 70˚C for 20 sec, 72˚C for 30 sec; final extension 72˚C for 1 min; 4˚C storage. PCR products were concentrated using DNA Clean and Concentrator (Zymo, D4013) followed by 0.8X SPRIselect (Beckman Coulter, B22318) left-sided size selection. Total PCR1 product was amplified in a PCR2 step, which added extension-specific full-length Illumina adapter sequences to the amplicon library. Individual samples were Illumina-indexed by choosing from 20 different index reverse primers. Cycle conditions were as follows: 98°C for 30 sec; 2 cycles of 98°C for 10 sec, 40°C for 20 sec, 72°C for 1 min; 6 cycles of 98°C for 10 sec, 65°C for 20 sec, 72°C for 1 min; 72°C for 5 min; 4°C storage. PCR2 products were concentrated again with DNA Clean and Concentrator and run on a 1% agarose gel. Bands of appropriate size (~550bp) were gel-purified using the Zymoclean™ Gel DNA Recovery kit (Zymo, D4008). Concentration of purified libraries were determined by a Nanodrop 2000c spectrophotometer and pooled at concentrations aimed at optimal read return. The quality of the final sequencing pool was verified on a fragment analyzer (Advanced Analytical Technologies) using DNF-473 Standard Sensitivity NGS fragment analysis kit. All samples passing quality control were sequenced. Antibody library pools were sequenced on the Illumina MiSeq platform using the reagent

kit v3 (2x300 cycles, paired-end) with 10% PhiX control library. Base call quality of all samples was in the range of a mean Phred score of 34.

Bioinformatics analysis and graphics

The MiXCR v2.0.3 program was used to perform data pre-processing of raw FASTQ files[42]. Sequences were aligned to a custom germline gene reference database containing the known sequence information of the V- and J-gene regions for the variable heavy chain of the trastuzumab antibody gene. Clonotype formation by CDRH3 and error correction were performed as described by Bolotin et al[42]. Functional clonotypes were discarded if: 1) a duplicate CDRH3 amino acid sequence arising from MiXCR uncorrected PCR errors, or 2) a clone count equal to one. Downstream analysis was performed using R v3.2.2[43] and Python v3.6.5[44]. Graphics were generated using the R packages ggplot2[45], RColorBrewer[46], and ggseqlogo[47].

Calculation of enrichment ratios (ERs) in DMS

The ERs of a given variant was calculated according to previous methods[48]. Clonal frequencies of variants enriched for antigen specificity by FACS, $f_{i,Ag+}$, were divided by the clonal frequencies of the variants present in the original library, $f_{i,Ab+}$, according to Equation 1.

$$ER = \frac{f_{i,Ag+}}{f_{i,Ab+}}$$

(Eq. 1)

A minimum value of -2 was designated to variants with log[ER] values less than or equal -2 and variants not present in the dataset were disregarded in the calculation. A clone was defined based on the exact amino acid sequence of the CDRH3.

Codon selection for rational library design

Codon selection for rational library design was based off the equation provided by Mason et al.[21], (Equation 2), where $Y_{n,deg}$ represents the amino acid frequency for a given degenerate codon scheme, $Y_{n,target}$ is the target amino acid frequency, and n is the number of amino acids, 20. Residues identified in DMS analysis to have a positive enrichment (ER > 1, or log[ER] > 0) were normalized according to their enrichment ratios and were converted to theoretical frequencies and taken as the target amino acid frequencies. Degenerate codon schemes were then selected which most closely reflect these frequencies as calculated by the mean squared error between the degenerate codon and the target frequencies.

$$Optimal\ Codon = arg_x min(\frac{1}{n}\sum_{i=1}^{n}(Y_{n,deg} - Y_{n,target})^2)$$

(Eq. 2)

In certain instances, if the selected degenerate codon did not represent desirable amino acid frequencies or contained undesirable amino acids, a mixture of degenerate codons were selected and pooled together to achieve better coverage of the functional sequence space.

Deep learning model construction

Deep learning models were built in Python v3.6.5. LSTM-RNNs, and CNNs were built using the Keras[49] v2.1.6 Sequential model as a wrapper for TensorFlow[50] v1.8.0. Model architecture and hyperparameters were optimized by performing a grid search of relevant variables for a given model. These variables include nodes per layer, activation function(s), optimizer, loss function, dropout rate, batch size, number of epochs, number of filters, kernel size, stride length, and pool size. Grid searches were performed by implementing a k-fold cross validation of the data set.

Deep learning model training and testing

Data sets for antibody expressing, non-binding, and binding sequences (Sequencing statistics: Supplementary Tables 1, 2) were aggregated to form a single, binding/non-binding data set where antibody expressing sequences were classified as non-binders, unless also identified among the binding sequences. Sequences from one round of antigen enrichment were excluded from the training data set. The complete, aggregated data set was then randomly arranged and appropriate class labeled sequences were removed to achieve the desired classification ratio of binders to non-binders (50/50, 20/80, 10/90, and non-adjusted). The class adjusted data set was further split into a training set (70%), and two testing sets (15% each), where one test set reflected the classification ratio observed for training and the other reflected a classification ratio of approximately 10/90 to resemble the physiological expected frequency of binders.

In silico sequence classification and sequence parameters

All possible combinations of amino acids present in the DMS-based combinatorial mutagenesis libraries were used to calculate the total theoretical sequence space of 7.17×10^8. 7.2×10^7 sequence variants were generated *in silico* by taking all possible combinations of the amino acids used per position in the combinatorial mutagenesis library designed from the DMS data following three rounds of enrichment for antigen binding variants (Supplementary Fig. 2c, 3c); Alanine was also selected to be included at position 103. All *in silico* sequences were then classified as a binder or non-binder by the trained LSTM-RNN and CNN models. Sequences were selected for further analysis if they were classified in both models with a prediction probability (*P*) of more than 0.75.

The Fv net charge and Fv charge symmetry parameter (FvCSP) were calculated as described by Sharma et al. Briefly, the net charge was determined by first solving the Henderson-Hasselbalch equation for each residue at a specified pH (here 5.5) with known amino acid pKas[51]. The sum across all residues was then calculated as the Fv net charge. The FvCSP was calculated by taking the product of the VL and VH charges. The hydrophobicity index (HI) was also calculated as described by Sharma et al., according to the following equation: $HI = -(\sum n_i E_i / \sum n_j E_j)$. E represents the Eisenberg value of an amino acid, n is the number of an amino acid, and i and j are hydrophobic and hydrophilic residues respectively.

The protein solubility score was determined for each, full-length CDRH3 sequence (15 a.a.) padded with 10 amino acids on both the 5' and 3' ends (35 a.a.) by the CamSol method[25] at pH 7.0.

The binding affinities for HLA alleles DRB1*0101, DRB3*0101, DRB4*0101, DRB5*0101 were determined for each 15-mer contained within the 10 amino acid padded CDRH3 sequence (35 a.a.) by NetMHCIIpan 3.2[26]. The output provides for each 15-mer a predicted affinity in nM and the % Rank which reflects the 15-mer's affinity compared to a set of random natural peptides. The % Rank measure is unaffected by the bias of certain molecules against stronger or weaker affinities and is used to classify peptides as weak or strong binders towards the specified MHC Class II allele.

Affinity measurements by biolayer interferometry

Monoclonal populations of the individual variants were isolated by performing a single-cell sort. Following expansion, supernatant for all variants was collected and filtered through a 0.20 μm filter (Sartorius, 16534-K). Affinity measurements were then performed on an Octet RED96e (FortéBio) with the following parameters: anti-human capture sensors (FortéBio, 18-5060) were hydrated in conditioned media diluted 1 in 2 with kinetics buffer (FortéBio, 18-1105) for at least 10 minutes before conditioning through 4 cycles of regeneration consisting of 10 seconds incubation in 10 mM glycine, pH 1.52 and 10 seconds in kinetics buffer. Conditioned sensors were then loaded with 0 μg/mL (reference sensor), 10 μg/mL trastuzumab (reference sample), or hybridoma supernatant (approximately 20 μg/mL) diluted 1 in 2 with kinetics buffer followed by blocking with mouse IgG (Rockland, 010-0102) at 50 μg/mL in kinetics buffer. After blocking, loaded sensors were equilibrated in kinetics buffer and incubated with either 5 nM or 25 nM HER2 protein (Sigma-aldrich, SRP6405-50UG). Lastly, sensors were incubated in kinetics buffer to allow antigen dissociation. Kinetics analysis was performed in analysis software Data Analysis HT v11.0.0.50.

ACKNOWLEDGEMENTS

We acknowledge the ETH Zurich D-BSSE Single Cell Unit and the Genomics Facility Basel for support, in particular, M. Di Tacchio, A. Gumienny, E. Burcklen, and C. Beisel. We also thank the Vendruscolo Lab (Cambridge, UK), in particular P. Sormanni, for assistance with implementing the CamSol method on large libraries, as well as the group of Prof. Morten Nielson (DTU, Denmark) for providing an easy-to-use package for MHC Class II affinity predictions. Funding was provided by the National Competence Center for Research on Molecular Systems Engineering.

AUTHOR CONTRIBUTIONS

D.M.M., S.F., C.R.W. and S.T.R. developed the methodology; D.M.M. and S.T.R. designed the experiments and wrote the manuscript; D.M.M., C.R.W. and S.F. analyzed sequencing data and performed deep learning analysis; C.J. generated in silico libraries; D.M.M. performed experiments; B.W., and S.M.M. performed cell line development.

458 COMPETING INTERESTS

459 ETH Zurich has filed for patent protection on the technology described herein, and D.M.M., S.F.,
460 C.R.W., and S.T.R. are named as co-inventors on this patent (United States Patent and Trademark
461 Office Provisional Application: 62/831,663).

462

463 **REFERENCES**

464 1. Paul, S. M. *et al.* How to improve R&D productivity: the pharmaceutical industry's grand
465 challenge. *Nature Reviews Drug Discovery* **9**, 203–214 (2010).
466 2. Sharma, V. K. *et al.* In silico selection of therapeutic antibodies for development: viscosity,
467 clearance, and chemical stability. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 18601–18606 (2014).
468 3. Jain, T. *et al.* Biophysical properties of the clinical-stage antibody landscape. *Proc. Natl. Acad.*
469 *Sci. U. S. A.* **114**, 944–949 (2017).
470 4. Hu, D. *et al.* Effective Optimization of Antibody Affinity by Phage Display Integrated with High-
471 Throughput DNA Synthesis and Sequencing Technologies. *PLoS One* **10**, e0129125 (2015).
472 5. Bos, A. B. *et al.* Development of a semi-automated high throughput transient transfection system.
473 *J. Biotechnol.* **180**, 10–16 (2014).
474 6. Greiff, V. *et al.* Learning the High-Dimensional Immunogenomic Features That Predict Public and
475 Private Antibody Repertoires. *J. Immunol.* **199**, 2985–2997 (2017).
476 7. Christensen, T., Frandsen, A., Glazier, S., Humpherys, J. & Kartchner, D. Machine Learning
477 Methods for Disease Prediction with Claims Data. *2018 IEEE International Conference on*
478 *Healthcare Informatics (ICHI)* (2018). doi:10.1109/ichi.2018.00108
479 8. Packer, M. S. & Liu, D. R. Methods for the directed evolution of proteins. *Nat. Rev. Genet.* **16**,
480 379–394 (2015).
481 9. Fox, R. *et al.* Optimizing the search algorithm for protein engineering by directed evolution.
482 *Protein Eng.* **16**, 589–597 (2003).
483 10. Fox, R. Directed molecular evolution by machine learning and the influence of nonlinear
484 interactions. *J. Theor. Biol.* **234**, 187–199 (2005).
485 11. Romero, P. A., Krause, A. & Arnold, F. H. Navigating the protein fitness landscape with Gaussian
486 processes. *Proc. Natl. Acad. Sci. U. S. A.* **110**, E193–201 (2013).
487 12. Bedbrook, C. N., Yang, K. K., Rice, A. J., Gradinaru, V. & Arnold, F. H. Machine learning to
488 design integral membrane channelrhodopsins for efficient eukaryotic expression and plasma

membrane localization. *PLoS Comput. Biol.* **13**, e1005786 (2017).

13. Angermueller, C., Pärnamaa, T., Parts, L. & Stegle, O. Deep learning for computational biology. *Mol. Syst. Biol.* **12**, 878 (2016).
14. Wainberg, M., Merico, D., Delong, A. & Frey, B. J. Deep learning in biomedicine. *Nat. Biotechnol.* **36**, 829–838 (2018).
15. Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* **33**, 831–838 (2015).
16. Cuperus, J. T. *et al.* Deep learning of the regulatory grammar of yeast 5' untranslated regions from 500,000 random sequences. *Genome Res.* **27**, 2015–2024 (2017).
17. Bulik-Sullivan, B. *et al.* Deep learning using tumor HLA peptide mass spectrometry datasets improves neoantigen identification. *Nat. Biotechnol.* (2018). doi:10.1038/nbt.4313
18. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
19. Rosenblatt, F. *The Perceptron, a Perceiving and Recognizing Automaton Project Para*. (1957).
20. Pogson, M., Parola, C., Kelton, W. J., Heuberger, P. & Reddy, S. T. Immunogenomic engineering of a plug-and-(dis)play hybridoma platform. *Nat. Commun.* **7**, 12535 (2016).
21. Mason, D. M. *et al.* High-throughput antibody engineering in mammalian cells by CRISPR/Cas9-mediated homology-directed mutagenesis. *Nucleic Acids Res.* **46**, 7436–7449 (2018).
22. Whitehead, T. A. *et al.* Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. *Nat. Biotechnol.* **30**, 543–548 (2012).
23. PDB ID: 1N8Z
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).
24. Rose, A. S. *et al.* NGL viewer: web-based molecular graphics for large complexes. *Bioinformatics* **34**, 3755–3758 (2018).
25. Sormanni, P., Aprile, F. A. & Vendruscolo, M. The CamSol method of rational design of protein mutants with enhanced solubility. *J. Mol. Biol.* **427**, 478–490 (2015).
26. Jensen, K. K. *et al.* Improved methods for predicting peptide binding affinity to MHC class II molecules. *Immunology* **154**, 394–406 (2018).
27. Wu, Z., Kan, S. B. J., Lewis, R. D., Wittmann, B. J. & Arnold, F. H. Machine learning-assisted directed protein evolution with combinatorial libraries. *Proc. Natl. Acad. Sci. U. S. A.* (2019).

519 doi:10.1073/pnas.1901979116

520 28. Vajdos, F. F. *et al.* Comprehensive functional maps of the antigen-binding site of an anti-ErbB2
521 antibody obtained with shotgun scanning mutagenesis. *J. Mol. Biol.* **320**, 415–428 (2002).

522 29. Townsend, S. *et al.* Augmented Binary Substitution: Single-pass CDR germ-lining and
523 stabilization of therapeutic antibodies. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 15354–15359 (2015).

524 30. Trudeau, D. L., Smith, M. A. & Arnold, F. H. Innovation by homologous recombination. *Curr.*
525 *Opin. Chem. Biol.* **17**, 902–909 (2013).

526 31. Riesselman, A. J., Ingraham, J. B. & Marks, D. S. Deep generative models of genetic variation
527 capture the effects of mutations. *Nat. Methods* **15**, 816–822 (2018).

528 32. Sormanni, P., Aprile, F. A. & Vendruscolo, M. Third generation antibody discovery methods: in
529 silico rational design. *Chem. Soc. Rev.* **47**, 9137–9157 (2018).

530 33. Raybould, M. I. J. *et al.* Five computational developability guidelines for therapeutic antibody
531 profiling. *Proc. Natl. Acad. Sci. U. S. A.* (2019). doi:10.1073/pnas.1810576116

532 34. Rabia, L. A., Zhang, Y., Ludwig, S. D., Julian, M. C. & Tessier, P. M. Net charge of antibody
533 complementarity-determining regions is a key predictor of specificity. *Protein Eng. Des. Sel.*
534 (2019). doi:10.1093/protein/gzz002

535 35. Abhinandan, K. R. & Martin, A. C. R. Analyzing the 'degree of humanness' of antibody
536 sequences. *J. Mol. Biol.* **369**, 852–862 (2007).

537 36. Igawa, T. *et al.* Antibody recycling by engineered pH-dependent antigen binding improves the
538 duration of antigen neutralization. *Nat. Biotechnol.* **28**, 1203–1207 (2010).

539 37. Kang, J. C. *et al.* Engineering a HER2-specific antibody–drug conjugate to increase lysosomal
540 delivery and therapeutic efficacy. *Nature Biotechnology* (2019). doi:10.1038/s41587-019-0073-7

541 38. Slaga, D. *et al.* Avidity-based binding to HER2 results in selective killing of HER2-overexpressing
542 cells by anti-HER2/CD3. *Sci. Transl. Med.* **10**, (2018).

543 39. Dunbar, J. *et al.* SAbPred: a structure-based antibody prediction server. *Nucleic Acids Res.* **44**,
544 W474–8 (2016).

545 40. An, Y., Zhang, Y., Mueller, H.-M., Shameem, M. & Chen, X. A new tool for monoclonal antibody
546 analysis. *mAbs* **6**, 879–893 (2014).

547 41. Menzel, U. *et al.* Comprehensive evaluation and optimization of amplicon library preparation
548 methods for high-throughput antibody sequencing. *PLoS One* **9**, e96727 (2014).

549 42. Bolotin, D. A. *et al.* MiXCR: software for comprehensive adaptive immunity profiling. *Nat.*
550 *Methods* **12**, 380–381 (2015).
551 43. R Development Core Team (2008) R: A Language and Environment for Statistical Computing. R
552 Foundation for Statistical Computing, Vienna, Austria.
553 44. Van Rossum, G. & Jr. Drake, F. L. *The Python Language Reference Manual*. (Network Theory.,
554 2011).
555 45. Wilkinson, L. ggplot2: Elegant Graphics for Data Analysis by WICKHAM, H. *Biometrics* **67**, 678–
556 679 (2011).
557 46. Brewer, C. A., Hatchard, G. W. & Harrower, M. A. ColorBrewer in Print: A Catalog of Color
558 Schemes for Maps. *Cartography and Geographic Information Science* **30**, 5–32 (2003).
559 47. Wagih, O. ggseqlogo: a versatile R package for drawing sequence logos. *Bioinformatics* **33**,
560 3645–3647 (2017).
561 48. Fowler, D. M. *et al.* High-resolution mapping of protein sequence-function relationships. *Nat.*
562 *Methods* **7**, 741–746 (2010).
563 49. Website. Available at: Chollet, F. (2015) keras, GitHub. https://github.com/fchollet/keras.
564 (Accessed: 24th April 2019)
565 50. Abadi, M. *et al.* TensorFlow: A System for Large-Scale Machine Learning. in *12th {USENIX}*
566 *Symposium on Operating Systems Design and Implementation ({OSDI} 16)* 265–283 (2016).
567 51. Peggy, K. Amino Acid pKa Values and Side Chain Identities. Available at:
568 http://homepage.smc.edu/kline_peggy/Organic/Amino_Acid_pKa.pdf. (Accessed: 24th April
569 2019)

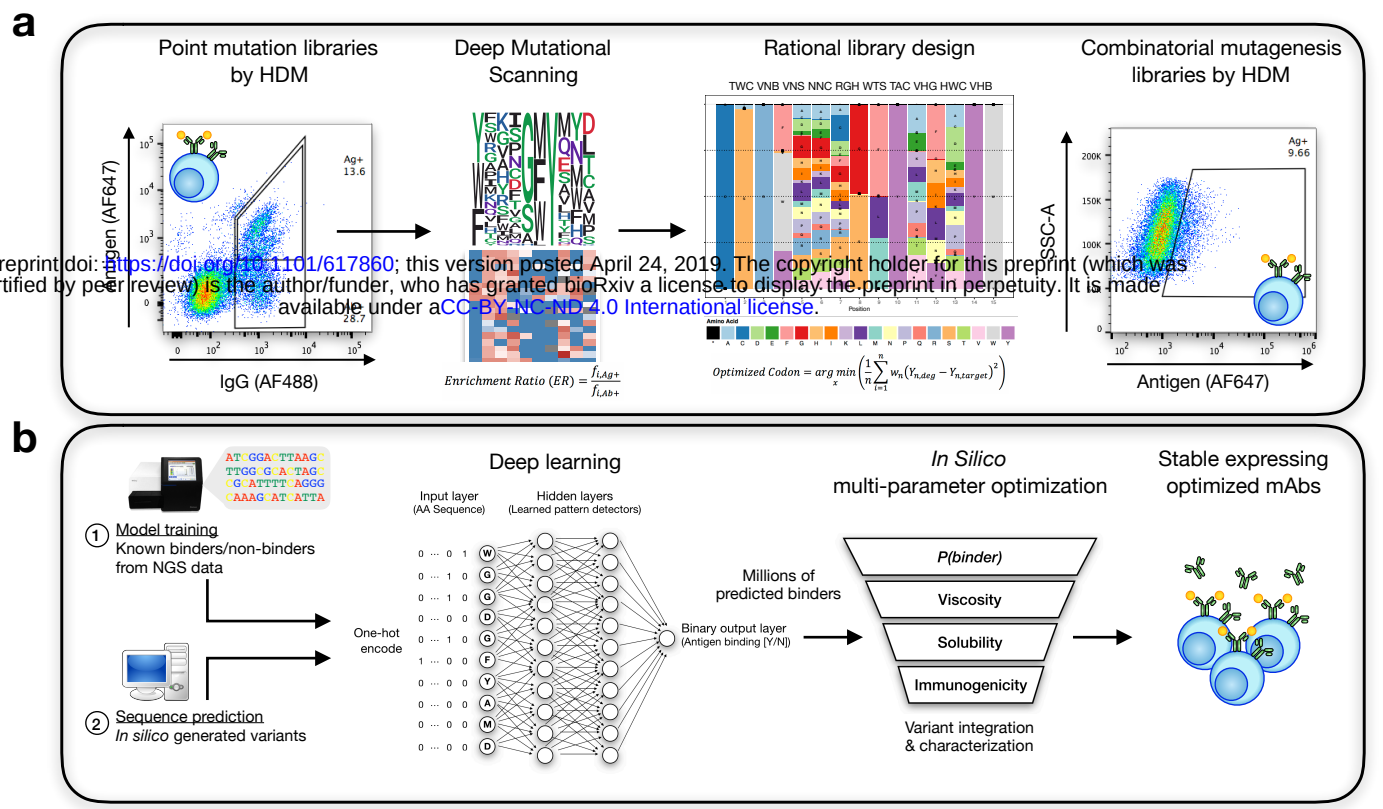


Figure 1: Implementing deep learning to predict antibody target specificity

(a) Generating quality data capable of training accurate models. First, deep mutational scanning assesses the impact mutations have on protein function across many different positions. These insights can then be applied to combinatorial mutagenesis strategies to guide protein library design capable of producing thousands of binding variants. **(b)** Sequence information for binders and non-binders can then be used to train deep neural networks to accurately predict antigen specificity of unknown antibody variants, producing millions of predicted binders. These binders can then be subjected to any available in silico methods for predicted various developability attributes.

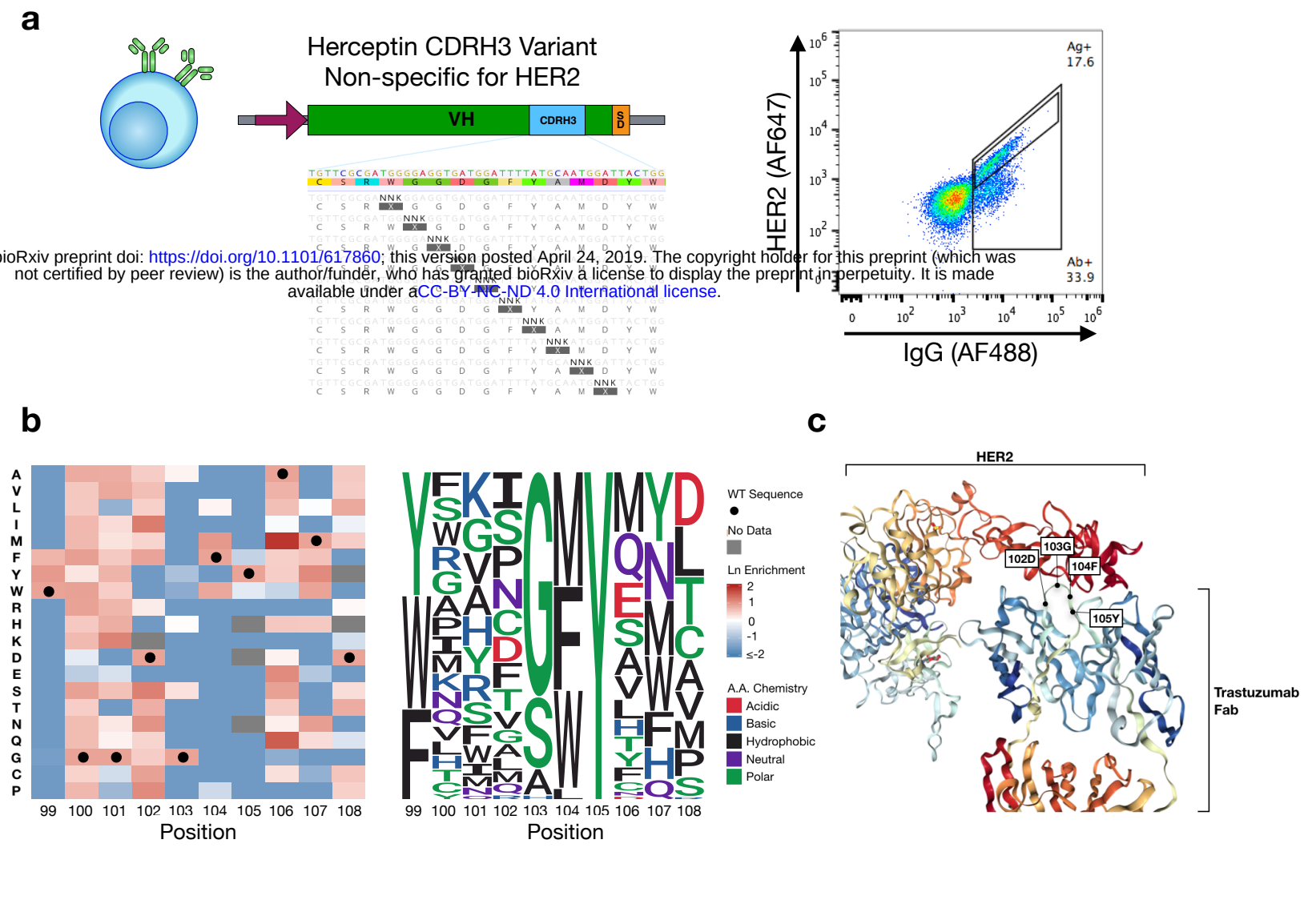


Figure 2: Deep mutational scanning reveals specificity determining residues

(a) Flow cytometry profile following integration of tiled mutations by homology-directed mutagenesis. Antigen specific variants underwent 3 rounds of enrichment (Supplementary Fig. 2) **(b)** Corresponding heatmap (left) following sequencing analysis of the pre-sorted (Ab+) and post-sorted (Ag+) populations (Supplementary Table 1). Wild type amino acids are marked by black circles. The resulting sequence logo plot (right) generated by positively enriched mutations per position. **(c)** 3D protein structure of trastuzumab in complex with its target antigen, HER2[23,24]. Locations of surface exposed residues: 102D, 103G, 104F, and 105Y are given[40].

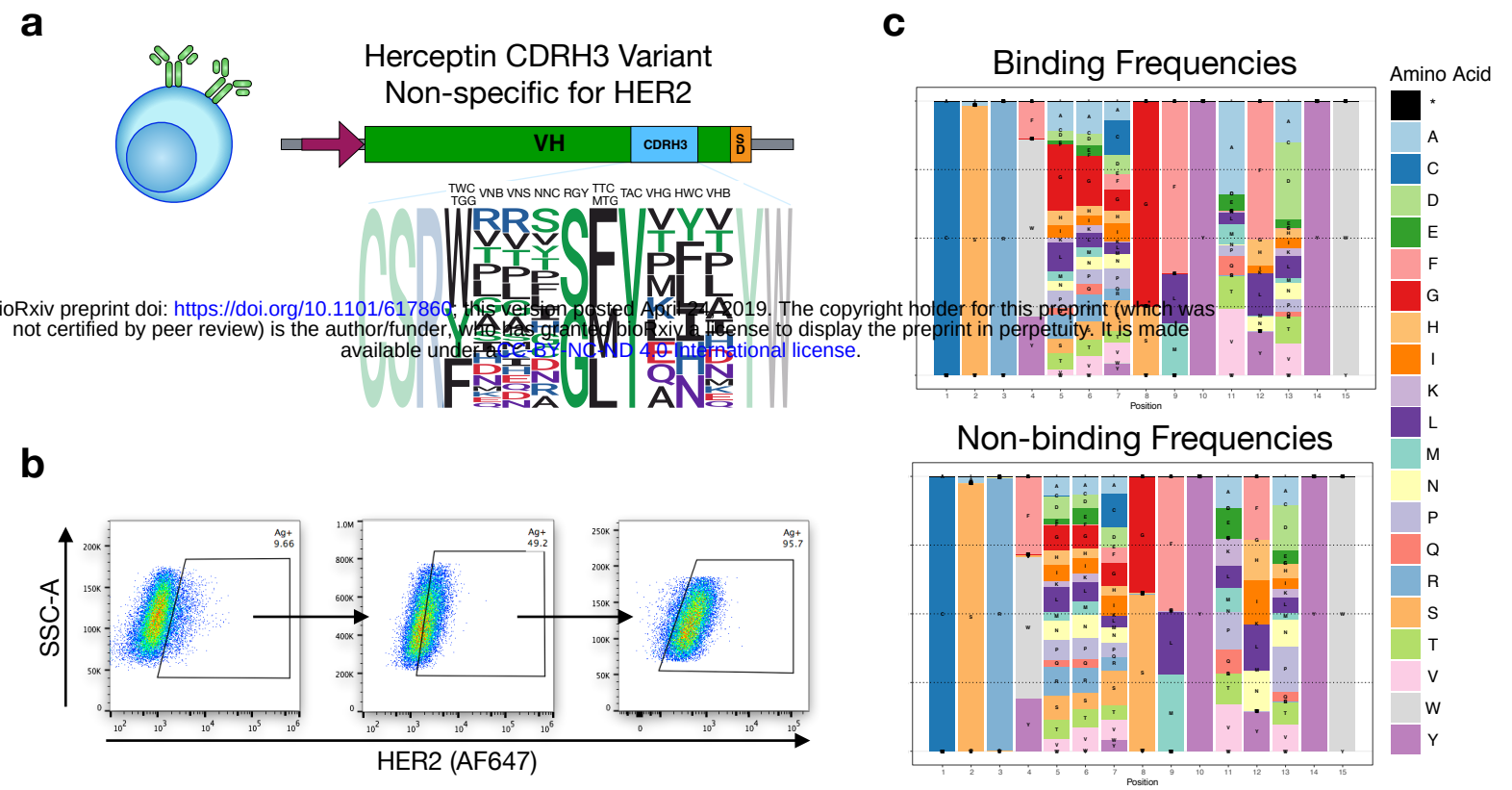
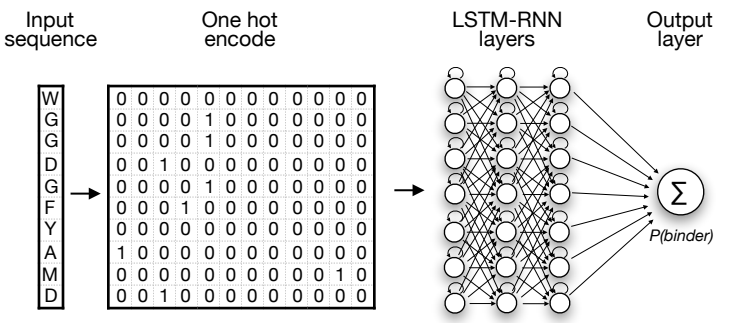


Figure 3: Combinatorial mutagenesis libraries generate data enriched with binding variants

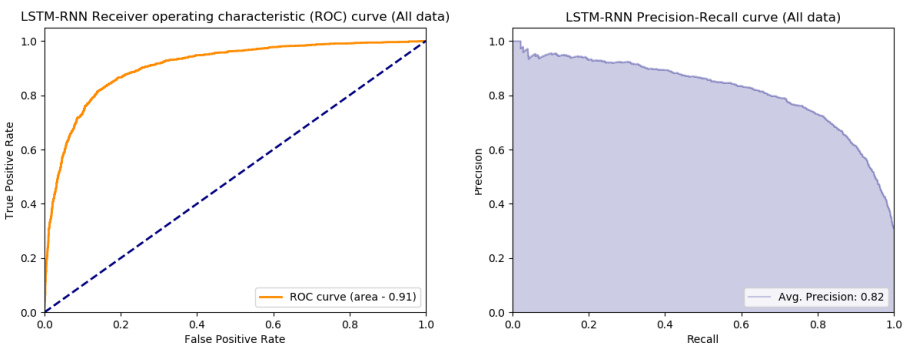
(a) Combinatorial mutagenesis libraries are designed from enrichment ratios observed in DMS data and integrated into the trastuzumab variant by homology-directed mutagenesis. **(b)** Flow cytometry plots resulting from transfection of a rationally designed library. Two rounds of enrichment were performed to produce a library of antigen specific variants. Deep sequencing was performed on the library (Ab+), non-binding variants (Ag-), and binding variants after 1 and 2 rounds of enrichment (Ag+1, Ag+2) (Supplementary Fig. 3, Supplementary Table 2). **(c)** Amino acid frequency plots of antigen binding variants and non-binding variants reveals nearly indistinguishable amino acid usages across all positions.

a

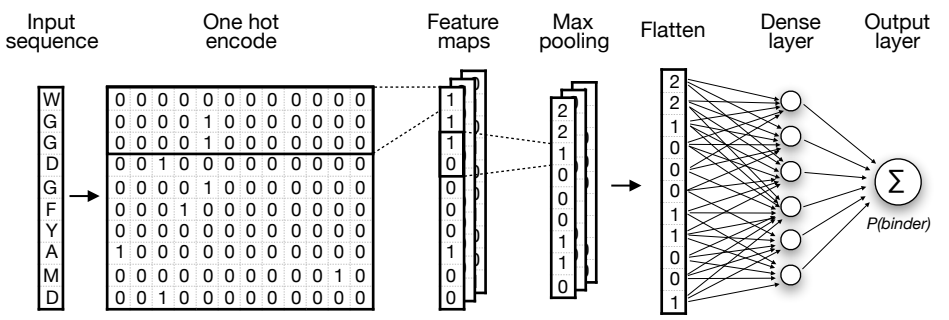


Parameters:	No. sequences: 36,497 Training/test split: 0.7 Class split: 0.31	Matrix size: (10, 20) Kernel size: 3 Stride: 1	No. Hidden layers: 3 Nodes per layer: 40 Dropout rate: 0.1	Activation: Sigmoid
Fitting:	Epochs: 20	Batch size: 32	Optimizer: RMSprop	Loss function: Binary cross-entropy

b



c



Parameters:	No. sequences: 36,497 Training/test split: 0.7 Class split: 0.31	Matrix size: (10, 20) Kernel size: 3 Stride: 1	Matrix size: (10,1,400) No. Filters: 400 Dropout rate: 0.5	Matrix size: (9,1,400) Pool size: 2 Stride: 1	No. Nodes: 50 Activation: ReLU	Activation: Sigmoid
Fitting:	Epochs: 20	Batch size: 16	Optimizer: Adam	Loss function: Binary cross-entropy		

d

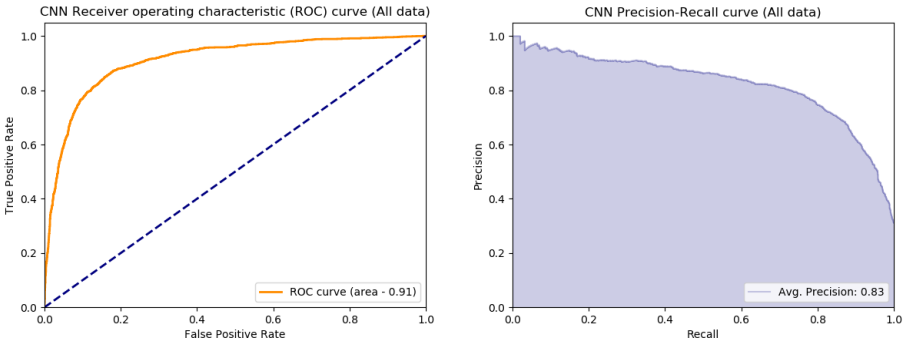
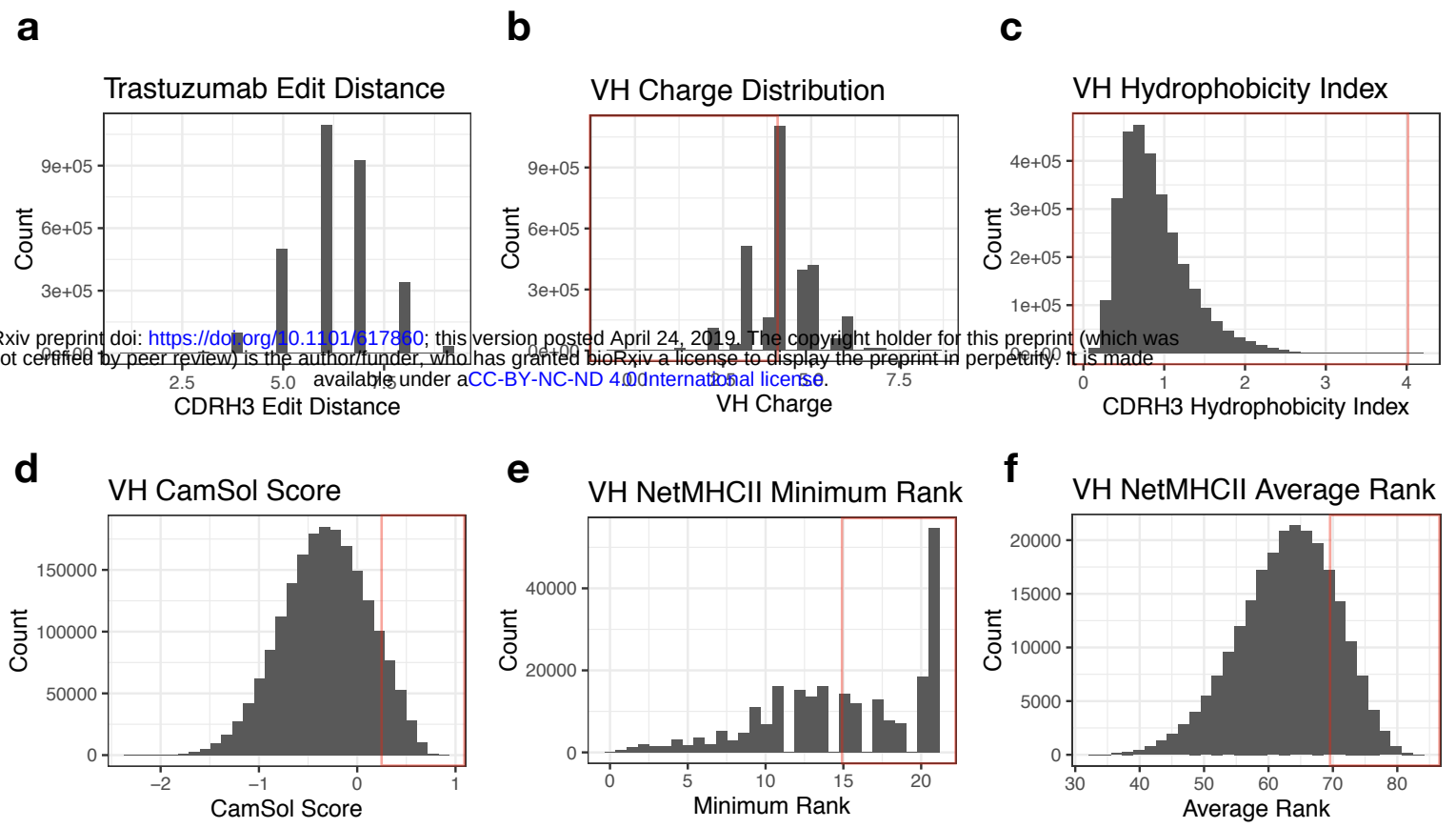


Figure 4: Deep learning models accurately predict antigen specificity

The selected network architectures and their model performance curves for classification of binding and non-binding sequences. Model training was performed on 70% of the data and testing was performed by withholding the remaining 30% and then comparing the model's classification of test sequences with the known classification. In lieu of adjusting the data set to a defined class split of binding/non-binding sequences, all known information was utilized to train and test the networks (approx. class split of 31%). (**a**) LSTM-RNN architecture and parameters used for model fitting. (**b**) ROC (receiver operating character) curve and PR (precision-recall) curve observed on the classification of sequences in the test set by the LSTM-RNN. (**c**) CNN architecture and parameters used for model fitting. (**d**) ROC curve and PR curve observed on the classification of sequences in the test set by the CNN. The high values observed for the ROC area under curve (AUC) and average precision of both networks represent robust measures of model accuracy and precision.



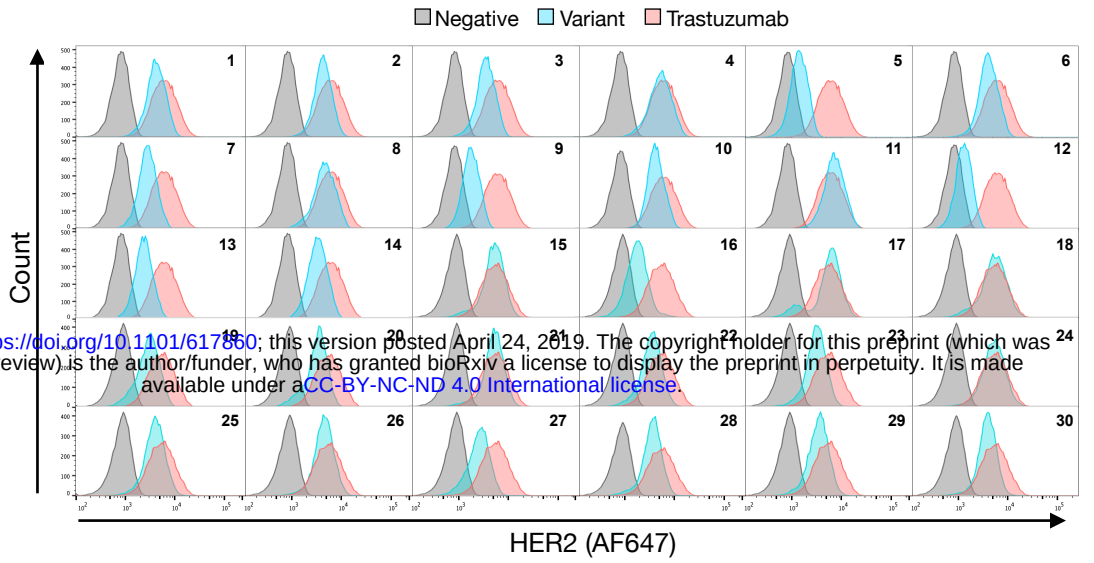
g

Filtering Parameters	No. Sequences
RNN *P*(binder) > 0.5 CNN *P*(binder) > 0.5	8,552,268 9,519,642
RNN *P*(binder) > 0.75 CNN *P*(binder) > 0.75	4,315,323 5,218,706
RNN, CNN consensus	3,002,476
Net charge < 4.2 Hydrophobicity Index < 4.0	1,927,942
Solubility score > 0.2	236,066
Minimum % Rank > 15 Average % Rank > 70 (Across all 15-mers)	40,588

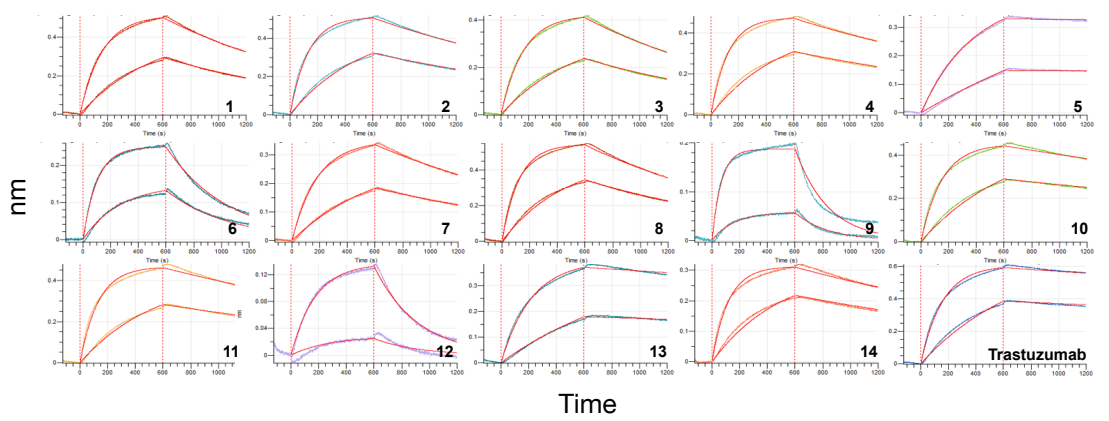
Figure 5: In silico screening of predicted binders produces multi-parameter optimized variants

Antigen specific predictions yield variants with a wide range In silico calculated parameters for developability. The following histograms show the parameters distributions of all predicted variants at the different stages of filtering. Red boxes indicate filtering cut-offs. **(a)** Levenshtein distance from wild-type trastuzumab. **(b)** Net charge of the VH domain. **(c)** CDRH3 hydrophobicity index. **(d)** CamSol intrinsic solubility score. **(e)** Minimum NetMHCIIpan % Rank across all possible 15-mers. **(f)** Average NetMHCIIpan % Rank across all possible 15-mers. **(g)** Filtering parameters and the number of sequences at the corresponding stage of filtering.

a



b



c

Variant	CDRH3 Sequence	RNN *P*(Binder)	CNN *P*(Binder)	LD	Net Charge	FvCSP	HI Sum	CamSol Score	NetMHCII Minimum	NetMHCII Average	KD (nM)	ka (1/Ms)	kd (1/s)
1	WQEAAFYAHD	0.833	0.930	5	6.02	9.06	1.07	0.45	21	63.4	3.31	2.22E+05	7.34E-04
2	YGGRGMYQLE	0.906	0.973	6	7.22	12.71	0.81	0.49	21	68.1	1.74	2.90E+05	5.02E-04
3	WNQLGMYVND	0.874	0.888	6	6.19	9.58	1.02	0.37	11	54.9	3.36	2.23E+05	7.47E-04
4	YGMRGLYAYE	0.942	0.948	6	7.22	12.71	0.92	0.48	13	50.5	1.47	3.10E+05	4.55E-04
5	WMETGFYTHD	0.895	0.966	5	6.02	9.06	0.97	0.38	20	66.9	0.14	1.06E+05	1.45E-05
6	WNAPSFYAND	0.885	0.977	5	6.19	9.58	1.02	0.57	17	64.4	7.43	3.02E+05	2.25E-03
7	WQGIGLYELD	0.942	0.889	5	5.25	6.70	1.28	0.46	12	60.9	3.20	1.95E+05	6.24E-04
8	WAMLGMYAHD	0.847	0.779	5	6.97	11.94	1.33	0.24	10	44.1	2.32	3.01E+05	6.96E-04
9	YQANGLYAYE	0.959	0.886	7	6.22	9.67	0.98	0.42	14	55.8	9.77	4.34E+05	4.24E-03
10	YRAVGFYTND	0.783	0.871	6	7.19	12.62	0.95	0.43	9	51.2	0.94	2.56E+05	2.41E-04
11	WAPYGLYAHD	0.948	0.959	5	6.97	11.94	1.19	0.41	19	59.0	1.28	2.92E+05	3.72E-04
12	WDGPAFYELD	0.860	0.936	5	4.27	3.73	1.13	0.53	13	70.2	31.1	1.03E+05	3.21E-03
13	WGIHSFYEHD	0.844	0.869	5	6.80	11.42	1.09	0.42	20	64.2	0.54	1.79E+05	9.63E-05
14	YGEYGMYVNK	0.888	0.870	7	7.22	12.71	0.85	0.47	21	65.0	1.24	3.23E+05	4.01E-04
15	WRDRGFYEYD	0.858	0.974	5	6.27	9.81	0.78	0.66	14	67.4	n.d.	n.d.	n.d.
16	WEEYGLYVHD	0.933	0.992	6	5.07	6.18	1.04	0.37	12	68.3	n.d.	n.d.	n.d.
17	YASAGMYTHD	0.927	0.883	7	6.97	11.94	0.89	0.54	19	61.7	n.d.	n.d.	n.d.
18	YGDAGMYALK	0.973	0.995	6	7.19	12.62	1.04	0.55	19	63.4	n.d.	n.d.	n.d.
19	WQLGGMYTHD	0.919	0.941	6	6.97	11.94	0.98	0.41	21	67.4	n.d.	n.d.	n.d.
20	WNSDGLYAYE	0.864	0.961	5	5.24	6.70	0.98	0.51	3	67.2	n.d.	n.d.	n.d.
21	WQRGGFYVND	0.956	0.993	5	7.19	12.62	0.94	0.47	9	64.1	n.d.	n.d.	n.d.
22	YGARGFYQND	0.892	0.789	5	7.19	12.61	0.82	0.69	13	65.2	n.d.	n.d.	n.d.
23	YAGPGMYTNQ	0.870	0.830	7	7.17	12.55	0.80	0.46	20	74.1	n.d.	n.d.	n.d.
24	WNPHGLYVND	0.939	0.974	6	6.97	11.93	1.01	0.49	13	65.9	n.d.	n.d.	n.d.
25	YGSNGLYANQ	0.914	0.908	6	7.17	12.55	0.90	0.51	21	70.4	n.d.	n.d.	n.d.
26	WPKVGLYTND	0.853	0.865	6	7.19	12.62	0.99	0.62	15	64.5	n.d.	n.d.	n.d.
27	WGIVSFYEND	0.871	0.873	5	5.24	6.70	1.24	0.20	17	59.6	n.d.	n.d.	n.d.
28	YSMPGMYTNA	0.848	0.938	8	7.17	12.55	0.85	0.37	21	62.8	n.d.	n.d.	n.d.
29	WAEAGMYEFD	0.880	0.915	6	4.30	3.83	1.12	0.48	17	65.6	n.d.	n.d.	n.d.
30	WPMCGMYTHD	0.835	0.855	6	6.97	11.93	0.88	0.30	21	69.5	n.d.	n.d.	n.d.
Trastuzumab	WGGDGFYAMD	0.962	0.938	0	5.21	6.61	0.92	0.49	11	68.3	0.40	2.51E+05	1.01E-04

Figure 6: Neural network predicted sequences are experimentally validated to be antigen-specific

30 variants were randomly selected and integrated into individual hybridoma cells lines by separately transfecting ssODN donor sequences with gRNA. **(a)** Out of the 30 sequences selected and integrated, all 30 bound the target antigen indicated by flow cytometry. **(b)** Affinities for 14 of the 30 variant sequences were determined by biolayer interferometry (BLI). The majority of sequences measured exude affinities in the single nano molar or sub-nanomolar range **(c)** A final table of the 30 variants randomly selected with their developability parameters. Values are shaded green to red according to their measure of developability. (n.d., not determined).