Deep learning-guided selection of antibody therapies with enhanced resistance to current and prospective SARS-CoV-2 Omicron variants

Working Paper

Author(s):

Frei, Lester; Gao, Beichen; Han, Jiami; Taft, Joseph M.; Irvine, Edward B.; Weber, Cédric (); Kumar, Rachita (); Eisinger, Benedikt N.; Reddy, Sai T.

Publication date: 2023-10-10

Permanent link: https://doi.org/10.3929/ethz-b-000652262

Rights / license: Creative Commons Attribution-NoDerivatives 4.0 International

Originally published in: bioRxiv, https://doi.org/10.1101/2023.10.09.561492

# 1 Deep learning-guided selection of antibody therapies with enhanced resistance

# 2 to current and prospective SARS-CoV-2 Omicron variants

- 3
- 4 Lester Frei<sup>1, 2,\*</sup>, Beichen Gao<sup>1, 2,\*</sup>, Jiami Han<sup>1, 2</sup>, Joseph M. Taft<sup>1, 2</sup>, Edward B. Irvine<sup>1</sup>, Cédric R. Weber<sup>3</sup>,
- 5 Rachita K. Kumar<sup>1</sup>, Benedikt N. Eisinger<sup>1</sup> and Sai T. Reddy<sup>1, 2, #</sup>
- <sup>6</sup> <sup>1</sup>Department of Biosystems Science and Engineering, ETH Zurich; Basel 4058, Switzerland.
- 7 <sup>2</sup>Botnar Research Centre for Child Health; Basel 4058, Switzerland.
- 8 <sup>3</sup>Alloy Therapeutics (Switzerland) AG, Allschwil 4123, Switzerland
- 9 \*equal contribution
- 10 #corresponding author: <u>sai.reddy@ethz.ch</u>
- 11
- 12

# 13 ABSTRACT

#### 14

15 Most COVID-19 antibody therapies rely on binding the SARS-CoV-2 receptor binding domain (RBD). However, heavily mutated variants such as Omicron and its sublineages, which are characterized by an ever increasing number 16 of mutations in the RBD, have rendered prior antibody therapies ineffective, leaving no clinically approved antibody 17 18 treatments for SARS-CoV-2. Therefore, the capacity of therapeutic antibody candidates to bind and neutralize current and prospective SARS-CoV-2 variants is a critical factor for drug development. Here, we present a deep learning-19 20 guided approach to identify antibodies with enhanced resistance to SARS-CoV-2 evolution. We apply deep mutational 21 learning (DML), a machine learning-guided protein engineering method to interrogate a massive sequence space of 22 combinatorial RBD mutations and predict their impact on angiotensin-converting enzyme 2 (ACE2) binding and 23 antibody escape. A high mutational distance library was constructed based on the full-length RBD of Omicron BA.1, 24 which was experimentally screened for binding to the ACE2 receptor or neutralizing antibodies, followed by deep sequencing. The resulting data was used to train ensemble deep learning models that could accurately predict binding 25 or escape for a panel of therapeutic antibody candidates targeting diverse RBD epitopes. Furthermore, antibody breadth 26 was assessed by predicting binding or escape to synthetic lineages that represent millions of sequences generated using 27 in silico evolution, revealing combinations with complementary and enhanced resistance to viral evolution. This deep 28 29 learning approach may enable the design of next-generation antibody therapies that remain effective against future SARS-CoV-2 variants. 30

Frei, Gao et al., Deep learning on Omicron variants, 2023.

### 31 INTRODUCTION

- 32 The onset of the COVID-19 pandemic spurred the rapid discovery, development and clinical approval of several antibody therapies. The monoclonal antibody LY-CoV555 (bamlanavimab) (Eli Lilly)<sup>1</sup> and the combination therapy 33 consisting of REGN10933 (casirivimab) and REGN10987 (imdevimab) (Regeneron)<sup>2</sup> were among the first to receive 34 35 Emergency Use Authorization (EUA) from the United States FDA in late 2020. The primary mechanism of action for these therapies consist of virus neutralization by binding to specific epitopes of the RBD of SARS-CoV-2 spike 36 37 (S) protein, thus inhibiting viral entry into host cells via the ACE2 receptor. However, the emergence of SARS-CoV-38 2 variants such as Beta, Gamma and Delta, each characterized by numerous mutations in the RBD, exhibited reduced 39 sensitivity to neutralizing antibodies, including LY-CoV555<sup>3,4</sup>, whose EUA was subsequently revoked. Of note, antibody combination therapies such as those from Regeneron and Eli Lilly (LY-CoV555+LY-CoV16 (etesevimab)) 40 41 were more resilient to viral variants and maintained their EUA throughout most of 2021<sup>3</sup>. However, the emergence 42 and rapid spread of Omicron BA.1 in late 2021, a variant which has a staggering 35 mutations in the S protein, 15 of 43 which are in the RBD resulted in substantial escape from nearly all clinically approved antibody therapies<sup>5</sup>. This includes the combination therapies from Regeneron and Eli Lilly, which also had their EUAs subsequently revoked<sup>6</sup>. 44 45 Even antibody therapies with exceptional breadth, which were initially discovered against the ancestral SARS-CoV-2 (Wu-Hu-1) and retained neutralizing activity against BA.1 – S309 (sotrovimab) (GSK/Vir)<sup>7</sup> and LY-CoV1404 46 47 (bebtelovimab) (Eli Lilly)<sup>8</sup> - lost efficacy against subsequent Omicron sublineages (e.g., BA.2, BA.4/5, and BO.1.1)<sup>9,10</sup> and had their clinical use authorization revoked. Despite there being a critical need for antibody therapies 48
- with weakened immune systems)<sup>11-15</sup>, since March 2023, there are no antibody therapies with an active clinical

for the protection of at-risk populations (young children, the elderly, individuals with chronic illnesses, and those

- whit weakened minimum systems) , since match 2023, there are no untroody theraptes with an active emited
- 51 authorization for COVID- $19^{16}$ .

52 The ephemeral clinical life span of COVID-19 antibody therapies has emphasized that, in addition to established 53 metrics for antibody therapeutics (e.g. neutralization potency, affinity, and developability)<sup>17</sup>, it is imperative to evaluate antibody breadth (ability of an antibody to bind to divergent SARS-CoV-2 variants) at early stages of 54 55 clinical development. This may enable selection of lead candidates that have the most potential to maintain activity against a rapidly mutating SARS-CoV-2. To address this, high-throughput protein engineering techniques such as 56 57 deep mutational scanning (DMS)<sup>18</sup> have been extensively employed to profile the impact of single position mutations in the RBD on ACE2-binding and antibody escape<sup>5,19-24</sup>. While DMS has proven effective for profiling single 58 mutations, many SARS-CoV-2 variants that have emerged possess multiple mutations in the RBD. For example the 59 60 aforementioned Omicron BA.1 lineage, or the recently identified BA.2.86, which possesses an astonishing 13 RBD mutations relative to its closest Omicron variant (BA.2) and 26 RBD mutations relative to ancestral Wu-Hu-125-27. 61 62 Experimental screening of combinatorial RBD mutagenesis libraries (e.g., using yeast surface display) vastly undersamples the theoretical protein sequence space, therefore computational approaches are increasingly being 63 64 employed in concert. For instance, experimental measurements such as DMS data have been used to calculate 65 statistical estimators<sup>28</sup> or to train machine learning models that make predictions on ACE2 binding and antibody escape<sup>29–31</sup>. While such computational tools enable interrogation of a larger mutational landscape of SARS-CoV-2, 66 67 their primary reliance on datasets that largely consist of single mutations from DMS experiments limits their ability 68 to capture the effects of combinatorial mutations, especially in the context of high mutational variants such as Omicron sublineages (e.g., BA.1, BA.4/5, BA.2.86). 69

70

Frei, Gao et al., Deep learning on Omicron variants, 2023.

- 71 Here, we apply deep mutational learning (DML), which combines yeast display screening, deep sequencing and
- 72 machine learning to address the emergence of Omicron BA.1 and its many sublineages. We expand the scope of
- 73 DML from screening short, focused mutagenesis libraries<sup>32</sup> to screening combinatorial libraries spanning the entire
- 74 RBD for binding to ACE2 or binding/escape from antibodies. Ensemble deep learning models utilizing dilated
- residual network blocks were trained with deep sequencing data and shown to make accurate predictions for ACE2
- <sup>76</sup> binding and antibody escape. Next, deep learning was used to determine the breadth of second-generation antibodies
- 77 (with known binding to BA.1) across a massive sequence landscape of BA.1-derived synthetic lineages, allowing the
- rational selection of specific antibody combinations that optimally cover the RBD mutational sequence space. This
- approach provides a powerful tool to guide the selection of antibody therapies that have enhanced resistance to both
- 80 current and future high mutational variants of SARS-CoV-2.
- 81

#### 82 **RESULTS**

#### 83 Design and construction of a high distance Omicron BA.1 RBD library

- A mutagenesis library was constructed based on BA.1, covering the entire 201 amino acid (aa) RBD region
- 85 (positions 331 531 of SARS-CoV-2 S protein). To maximize the interrogated RBD sequence space, the library
- design was entirely synthetic and unbiased, as it did not consider evolutionary data or previous experimental
- findings. For the construction of the library, the RBD sequence was split into 11-12 fragments, each with an
- approximate length of 48 nucleotides (nt) (Supplementary Table 1). For each fragment, 136 different single-stranded
- 89 oligonucleotides (ssODN) were designed, where each ssODN had either one codon or all combinations of two
- 90 codons replaced by fully degenerate NNK codons (N = A, G, C, or T; K = G or T) (Fig. 1a) (Methods). For each
- 91 fragment, ssODNs were amplified using PCR to generate double-stranded DNA. Each fragment was flanked by
- 92 recognition sites for the type II-S restriction enzyme BsmBI, thus enabling assembly into full-length RBD regions by
- 93 Golden Gate assembly (GGA)<sup>33</sup>. GGA utilizes type II-S restriction enzymes capable of cleaving DNA outside their
- 94 recognition sequence, thereby allowing the resulting DNA overhangs to have any sequence. Based on the overhangs,
- 95 individual fragments were assembled by DNA ligase to full-length RBD sequences with high fidelity<sup>34,35</sup>. The
- 96 restriction sites were eliminated during the process, thus enabling scarless assembly of full length RBD sequences
- 97 (Fig. 1b, Methods)<sup>34</sup>. This approach yielded approximately 98% correctly assembled RBD sequences
- 98 (Supplementary Fig. 1). Since GGA required four nt homology between individual fragments for ligation, this led to
- 99 portions of the sequence which needed to remain constant, thereby restricting library diversity<sup>36</sup>. To overcome this
- 100 limitation, four sub-libraries were designed and individually assembled. Using sub-library 1 as a reference, sub-
- 101 library 2 is shifted by 12 nt, sub-library 3 by 24 nt and sub-library 4 by 36 nt. These sub-libraries provided an
- 102 increase in the mutational space covered by the RBD combinatorial mutagenesis library, since at the GGA homology
- 103 for a given library, the remaining three libraries can have mutations (Fig. 1c).

Frei, Gao et al., Deep learning on Omicron variants, 2023.



104 105

106 Figure 1. Construction of a high edit distance (ED) synthetic variant library based on Omicron BA.1 RBD. a,

107 The RBD sequence was split into 11-12 fragments, each being approximately 48 nt in length. For each fragment, a 108 ssODN library with either zero, one or two mutations was designed. **b**, To introduce mutations, NNK codons were tiled 109 across the fragments (1). Each fragment was flanked by BsmBl sites (2). The ssODNs were flanked by primer binding 110 sites for double stranded synthesis through PCR (primers are represented by black arrows and primer binding sites 111 are peach colored) (3). The type II-S restriction enzyme BsmBI gives rise to orthogonal four nt overhangs, which are 112 used by a ligase to assemble individual fragments into full-length RBD sequences (4). c, The use of GGA for library 113 construction required the presence of constant regions for ligation between fragments (in black), thereby restricting 114 the library diversity. To overcome this drawback, four staggered sub-libraries were constructed. Due to limitations in 115 sequencing length, it was further necessary to split the RBD into two separate libraries. The extent of seg-library A is 116 indicated in orange and seq-library B in cyan. The primer binding sites for deep sequencing are indicated using 117 orange and cyan arrows. d, Targeted sequencing of seq-libraries A and B showed comprehensive mutational 118 coverage for both libraries. The same color scheme as in (c) was used to indicate the extent of both libraries. e, To 119 adjust the mutational rate of the library, three different conditions were tested. Different amounts of fragments with 120 zero, one or two mutations were pooled in different ratios which yielded libraries with different mutational distributions.

121

122 The current read length of Illumina does not allow coverage of the entire RBD with a single sequencing read (paired-

- 123 end). Therefore, two separate sequencing libraries (seq-library A and B) were individually constructed. The seq-
- 124 library A and B possessed mutations in positions 331 475 and 386 531, respectively (Fig. 1c). The seq-libraries
- 125 were constructed separately but all subsequent steps were performed in a pooled fashion. Following deep
- 126 sequencing, complete mutational coverage for each residue was observed in both seq-libraries (Fig. 1d).
- 127 Interestingly, the mutational frequency is somewhat variable across the seq-libraries, showing a marked decrease in

- 128 mutations every 16 residues. The low mutational frequencies line up with GGA homologies of sub-library 1. We
- 129 hypothesize that when pooling the sub-libraries, sub-library 1 was more prominent than the other sub-libraries and
- 130 therefore less mutations at these sites are observed.
- 131
- 132 Next, to optimize the number of mutations per RBD sequence, titration of the fragment assembly step was
- 133 performed. Wild-type (WT) fragments (BA.1 sequence) and fragments with one and two mutations respectively were
- 134 pooled in different ratios for assembly. Separately, assembly was performed with 60%, 70% and 80% of WT
- 135 fragments, with the remaining percentage split evenly between fragments with one and two mutations. Deep
- 136 sequencing of these libraries revealed a clear trend in mutational distribution based on the different ratios,
- 137 highlighting the tunable nature of our approach. Based on these results, all subsequent work was carried out using the
- 138 60% WT library as it has the highest mean number of mutations, therefore providing an appropriate approximation
- 139 for extensively mutated Omicron sublineages.



Frei, Gao et al., Deep learning on Omicron variants, 2023.

# 141 Figure 2. Screening RBD libraries for ACE2 binding and antibody escape by yeast display and deep

- 142 **sequencing. a**,**b**, Workflow for sorting of yeast display RBD libraries and FACS dot plots for **a**, ACE2 and **b**,
- 143 antibodies Brii-198 and ZCB11. Gating schemes correspond to binding and non-binding (escape) RBD variant
- populations. **c**,**d**,**e** Heatmaps depict the binding score of each aa per position of full-length RBD following sorting and
- deep sequencing of libraries for c, ACE2 d, Brii-198 e, and ZCB11; higher binding score indicates greater frequency in
- 146 the binding population vs non-binding population. WT BA.1 residues are in gray. **f**, Heatmaps for seq-libraries A and B
- 147 depict binding scores for ACE2 and antibodies of key mutations seen in major Omicron sublineage variants.
- 148

# 149 Screening RBD libraries for ACE2 binding and antibody escape

- 150 Co-transformation of yeast cells (S. cerevisiae, strain EBY100) using the PCR amplified RBD library and linearized
- 151 plasmid yielded more than 2 x 10<sup>8</sup> transformants (Methods). Yeast surface display of RBD variants was achieved
- 152 through C-terminal fusion to Aga2<sup>37</sup>. Next, fluorescence-activated cell sorting (FACS) was used to isolate yeast cells
- 153 expressing RBD variants that either retained binding or completely lost binding to dimeric soluble human ACE2
- 154 (Fig. 2a). Notably, RBD variants with only partial binding to ACE2 were not isolated, as such intermediate
- 155 populations could not be confidently classified as either binding or non-binding. Removing these variants is essential
- 156 to obtain cleanly labeled datasets for training supervised machine learning models.
- 157

158 Since binding to ACE2 is a prerequisite for cell entry and subsequent viral replication, only this population is

159 biologically relevant. Thus, only the ACE2-binding population was used in following FACS sorts to isolate RBD

- 160 variants that either retained binding or completely lost binding (escape) activity to a panel of eight neutralizing
- 161 antibodies (Fig. 2b, Supplementary Fig. 2 and Supplementary Table 2). The antibodies selected target different
- 162 epitopes, and are well characterized for their neutralizing activity to BA.1 and its sublineages, which provide a good
- 163 internal control to assess the accuracy of our method<sup>38-40</sup>. The panel consists of the following antibodies: A23-58.1<sup>41</sup>,
- 164 COV2-2196<sup>42</sup>, Brii-198<sup>43</sup>, ZCB11<sup>44</sup>, 2-7<sup>45</sup>, S2X259<sup>46</sup>, ADG20<sup>47</sup>, and S2H97<sup>20</sup>.
- 165

166 Following ACE2 and monoclonal antibody sorting, pure populations of RBD variants (binding and non-binding)

- 167 were subjected to deep sequencing (Supplementary Table 3). Reads covering the RBD sequence were then extracted
- 168 from the NGS data and heatmaps were constructed depicting binding scores (relative aa frequencies per position in
- 169 the RBD of binding vs non-binding variants) (Fig. 2c-e and Supplementary Fig. 3). The heatmaps demonstrate nearly
- 170 complete coverage of mutations across the RBD within all sorted populations. A heterogeneous distribution of
- 171 mutations is observed for ACE2 binding, with no specific positions or mutations showing dominance (Fig. 2c). This
- agrees with previous studies that suggest the Q498R and N501Y mutations present in BA.1 exhibit strong epistatic
- 173 effects that compensate for many mutations that cause loss of binding<sup>48</sup>. In contrast, for certain antibodies, clear
- 174 mutational patterns could be observed, including escape mutations that correspond with previous DMS studies (Fig.
- 175 2d-f and Supplementary Fig. 3). For example, RBD escape variants for Brii-198 are enriched for mutations in
- positions 346 and 452 (Fig. 2d), which are present in BA.1 and BA.4/BA.5, respectively and correspond to previous
- 177 work that shows they drive a drastic loss of binding to Brii-198<sup>49</sup>. In contrast, enrichment of these escape mutations
- are not observed for antibody 2-7 (Supplementary Fig. 3), even though Brii-198 and 2-7 share a similar epitope,
- 179 suggesting that the binding modality between these two antibodies are different, which is also reflected by their
- 180 difference in resistance to Omicron variants (e.g., 2-7 shows strong binding to BA.2 and BA.4/BA.5, while Brii-198
- 181 does not bind BA.2.12 and BA.4/BA.5)<sup>39,50</sup>. Similarly, the F486V mutation, which has been demonstrated to
- drastically reduce the neutralization potency of ZCB11 by over 2000-fold<sup>10</sup>, is highly enriched in the RBD escape
- 183 population (Fig. 2e, f). These mutations are also seen in A23-58.1 and COV2-2196, which bind to a similar epitope

- 184 (Supplementary Fig. 3). Lastly, for ADG20, we observe a high enrichment of escape mutations in 408(Fig. 2f,
- 185 Supplementary Fig. 3); this position is also mutated in BA.2 and BA.4/BA.5 variants, which have been shown to
- 186 have drastically reduced neutralization by ADG20<sup>10</sup>.
- 187
- 188 While heatmap analysis allows specific mutational patterns to be linked with antibody escape profiles, the high-
- 189 dimensional nature and potentially higher order impact of combinatorial mutations is not reflected in this format.
- 190 It is apparent that protein epistasis and combinatorial mutations can modify the effect of known escape mutations,
- 191 either amplifying or reducing antibody binding. For example, individual RBD mutations (G339D, S371F, S373P,
- 192 S375, K417N, N440K, G446S, S477N, T478K, E484A, Q493R, G496S, Q498R, N501Y, Y505H) in BA.1 and
- BA.1.1 do not enhance escape to COV2-2196, with each mutation causing an average fold reduction of 2.2, but
- 194 together cause over 200-fold reduction in neutralization<sup>51</sup>. Conversely, the introduction of the single R493Q mutation
- in BA.2 substantially rescued the neutralizing activities of Brii-198, REGN10933, COV2-2196 and ZCB11<sup>10</sup>. Thus,
- 196 while the heatmaps indicate specific mutational contributions to antibody escape, other techniques such as deep
- 197 learning are required to capture the high-dimensional nature of combinatorial mutations, and generalize to future
- 198 mutations.
- 199



Frei, Gao et al., Deep learning on Omicron variants, 2023.

201 Figure 3. Training and testing of deep learning ensemble models for prediction of ACE2 binding and antibody 202 escape based on full-length RBD sequences. a, Deep sequencing data of sorted yeast display libraries are 203 encoded by one-hot encoding and used to train CNN models with several dilated convolutional residual blocks. The 204 models perform a final classification by predicting binding or non-binding to ACE2 or antibodies based on the encoded 205 RBD sequence. b, Performance of CNN models trained on all datasets shown by accuracy, Matthews Correlation 206 Coefficient (MCC) and precision. Scores are a result of five rounds of cross-validation with mean performance 207 displayed, and standard deviation indicated by error bars. c, Majority voting by an ensemble of models is used to 208 determine the final label for each variant. d. Predicted labels of antibodies to well-characterized Omicron variants: 209 colors indicate final labels, and mis-classifications are marked with an "X". e, Comparison of predicted labels to experimental  $K_d$  reported in He et al. (2023)<sup>39</sup> for antibodies 2-7, COV2-2196, S2H97, and S2E12 (as a proxy for A23-210 211 58.1), region highlighted in gray indicates model "sensitivity" threshold.

212

### 213 Deep learning ensemble models accurately predict ACE2 binding and antibody escape

214 To address the high dimensionality of our dataset and to understand epistatic effects between mutations in the full RBD mutational sequence space, which is far too vast to be comprehensively screened experimentally, we trained 215 216 deep learning ensemble models. Deep sequencing data from FACS-isolated yeast populations underwent pre-217 processing and quality filtering prior to being used as training data for machine learning. In the datasets for all 218 antibodies, using the BA.1 RBD sequence as a reference, the mean rate of mutations ranged between ED two (ED<sub>2</sub>) 219 and three ED<sub>3</sub>, with a max ED<sub>8</sub> (Methods and Supplementary Fig. 5 and 6). Following nucleotide to protein 220 translation, one-hot encoding was performed to convert as sequences into an input matrix for machine and deep 221 learning models (Fig. 3a). Supervised machine learning models were trained to predict the probability (P) that a 222 specific RBD sequence will bind to ACE2 or a given antibody. A higher P signifies a stronger correlation with 223 binding, whereas a lower P corresponds to non-binding (escape). The machine learning models tested included K-224 nearest neighbor (KNN), logistic regression (Log Reg), naive Bayes (NB), support vector machines (SVM) and 225 Random Forests (RF). Additionally, as a baseline for deep learning models, a multilayer perceptron (MLP) model was also tested. Finally, we implemented a convolutional neural network (CNN) inspired by ProtCNN<sup>52</sup>, which 226 leverages residual neural network blocks and dilated convolutions to learn global information across the full RBD 227 228 sequence (Fig. 3a).

229

Each model was trained using an 80/10/10 train-validate-test split of data. Inputs were one-hot encoded RBD

sequences, with the CNN using a 2D matrix and others using a 1D flattened vector. For initial benchmarking, a

232 collection of different baseline machine learning models were trained on each dataset with hyperparameter

233 optimization through random search, and were evaluated with 5-fold cross validation based on several common

234 metrics (accuracy, F1, MCC, precision and recall). In the baseline machine learning models, class balancing was

achieved by random subsampling from the majority class. Unsampled majority class sequences were set aside and

236 merged with the held-out test set for use in model evaluation. Following training, most of the baseline models

resulted in relatively high accuracy scores (0.7-0.9) across all datasets, however for smaller datasets (under 20,000

sequences) substantially lower values of F1 (0.2-0.3) and MCC (0.2-0.4) were observed (Supplementary Fig. 7). In

239 contrast, the baseline MLP and CNN deep learning models performed substantially better, including large

240 improvements in F1 and MCC scores (Fig. 3b and Supplementary Fig. 7). While in most cases, the MLP models

resulted in relatively high MCC scores (up to > 0.9), CNN models performed substantially better, with MCC scores

242 up to 0.15 higher than MLPs (Supplementary Fig. 7).

243

Having determined that the CNN models performed superior to the machine learning models and MLP, we next

245 applied an exhaustive hyperparameter search on CNN models to optimize their performance (Supplementary Table

- 4). Training data was balanced through rejection sampling, while the held-out test set remained imbalanced to
- 247 accurately evaluate F1 and MCC scores. To prevent data leakage during training, the held-out test set was fixed and
- 248 multiple models were trained on different training-validation splits of the remaining dataset to make sure each model
- learned slightly different parameters of the data. When tested on the held-out test set, the final models yielded robust
- 250 predictive performance up to an ED of eight from the WT BA.1 sequence (Supplementary Fig. 8).
- 251
- For our final ensemble, we selected three CNN models from each library with the highest MCC scores to
- 253 generate the predicted labels for each variant through majority voting (Fig. 3c). In short, each model outputs *P* of
- binding for each input sequence, and labels are assigned based on a threshold. Here, P > 0.75 was classified as
- binding, P < 0.25 was classified as non-binding (escape), and those in between were labeled as "uncertain". The final
- 256 classification label was taken as the majority label across the three models. An RBD variant was assigned a predicted
- 257 "escape" label if either the ensemble models of seq-library A or seq-library B predicted escape, and assigned a
- 258 predicted "binding" label only if both models predicted binding. This leads to a more conservative prediction of
- antibody binding to variants, and minimizes false-positives. We tested the performance of the ensemble models on
- 260 published experimental data of antibody binding (or neutralization) to Omicron sublineages<sup>10,38,49,53–56</sup>. In general, the
- 261 ensemble model predictions performed well, assigning accurate labels to over 80% of the antibody-variant pairs,
- with only four mis-classifications (Fig. 3d). Three of these mis-classifications were false-negatives, which is likely
- due to the more conservative approach used for binding classification (Fig. 3d, Supplementary File). Comparing
- model predictions with published data on antibody affinity values (equilibrium dissociation constant, *K*<sub>d</sub>), revealed
- that uncertain and mis-classifications were confined to antibodies with intermediate affinities ( $K_d = 75 250$  nM),
- suggesting that there may be a sensitivity limit correlated with lower antibody affinity (Fig. 3e).



Frei, Gao et al., Deep learning on Omicron variants, 2023.

Figure 4. Evaluating antibody breadth on synthetic Omicron lineages. a, Example of a synthetic lineage tree of 268 269 sequences generated containing mutations unseen in major Omicron variants, with heatmap indicating the deep 270 learning predictions of binding or escape for individual antibodies. b, Total mean predicted breadth of individual 271 antibodies and combinations on synthetic lineages generated from 2022 mutational probabilities. c, The fraction (%) of 272 sequences bound by individual antibodies at different ED from BA.1. d, UMAP displays a subsample of ZCB11 273 escape variants in protein sequence space with antibody-specific binding clusters highlighted. e, Sequence logos 274 show the top 25 positions with greatest Kullback-Leibler (KL)-divergence in ZCB11 escape variants at ED<sub>6</sub>, and 275 sequences re-captured by Brii-198, ADG20 and A23-58.1, f. The top 50 predicted mutations ranked by their escape 276 scores (see Methods) from the generated synthetic lineages, with new mutations seen in the BA.2.86 variant 277 highlighted.

### 278 Designing antibody combinations by predicting resistance to synthetic Omicron lineages

- 279 After validating the performance of CNN models on test and validation data, we next deployed them to evaluate the
- resistance of antibodies to viral evolution. While antibody breadth is normally evaluated retroactively based on
- 281 neutralization or binding to previously observed variants, here we aimed to leverage this machine learning-guided
- 282 protein engineering approach to prospectively characterize and assess the breadth of antibodies against Omicron
- variants that may emerge in the future. This was achieved by generating synthetic lineages stemming from BA.1.
- 284 Since the potential sequence space of combinatorial RBD mutations is exceedingly massive, it was necessary to
- reduce this to a relevant subspace, therefore mutational probabilities were calculated across the RBD using SARS-
- 286 CoV-2 genome sequencing data (available on Global Initiative on Sharing Avian Influenza Data, GISAID
- [www.gisaid.org]) and used to generate synthetic lineages that mimic natural mutational frequencies. Starting with
- the BA.1 sequence, mutational frequencies from 2021 and 2022 were utilized to generate ten sets of 250,000
- synthetic RBD sequences through six rounds of *in silico* evolution, where the 100 variants with the highest predicted
- 290 score for ACE2 binding (averaged across the ensemble CNN models) in each round were used as seed sequences for
- the next round of mutations. Next, the ensemble deep learning models were used to predict antibody binding or
- 292 escape (or uncertain classification) for the synthetic variants. This provides an estimation of each individual
- antibody's binding breadth in the generated sequence space and thus correlates with resistance to prospective
- 294 Omicron lineages (Fig. 4a,b, Supplementary Fig. 9).
- 295

296 Since several of the clinically used antibody therapies for COVID-19 consisted of a cocktail of two antibodies (e.g.,

- 297 LY-CoV555+LY-CoV16, REGN10933+REGN10987, COV2-2130+COV2-2196), we also determined antibody
- breadth across all two-way combinations. For the 2022-based synthetic lineages, ZCB11 showed the greatest
- predicted breadth, followed by A23-58.1, Brii-198 and ADG20 (Fig. 4b). The ensemble models predict very low
- 300 breadth for 2-7 and COV2-2196, despite both maintaining binding to BA.2 and beyond<sup>39</sup>. This is likely due to the
- 301 high uncertainty of these models. The predicted coverage of ZCB11 corresponds well with experimental
- measurements that show it maintains high affinities and neutralization to several Omicron variants (BA.2, BA.4/5)<sup>10</sup>.
- 303 Similarly, Brii-198 and A23-58.1 have been shown to bind BA.2, BA.2.12 and BA.2.75 variants<sup>40</sup>, aligning with the
- 304 predictions of their relatively high breadth. Examining breadth profiles of each antibody as a function of ED revealed
- differing profiles, such as ZCB11 and Brii-198 maintaining high breadth at larger ED (>ED4), while A23-58.1 and
- 306 ADG20 have substantially lower breadth at large ED (Fig. 4c). The predicted breadth of several antibodies were
- 307 substantially different for synthetic lineages generated using 2021 mutational probabilities. For example, the breadth
- 308 of ADG20 is substantially higher as it is predicted to bind over 50% of variants, while the breadth of Brii-198 and
- A23-58.1 is reduced by 9% and 15%, respectively (Supplementary Fig. 9 and 10). This suggests that correctly
- 310 anticipating antigenic drift and changes in mutational frequencies play an important role in determining breadth
- 311 predictions.

Frei, Gao et al., Deep learning on Omicron variants, 2023.

- 312 313 It is worth noting that calculating the breadth of antibody combinations is not simply additive. For example, while 314 Brii-198 ranks lower than A23-58.1 in total breadth, Brii-198 provides more complementary coverage to ZCB11 315 (Brii-198 binds to more variants that escape ZCB11), resulting in an overall increase in variant coverage in a 316 simulated cocktail. Examining the distribution of escape variants for ZCB11 at ED<sub>6</sub>, where it sees its most significant 317 breadth reduction— the three other highly ranked antibodies (A23-58.1, Brii-198 and ADG20) re-establish coverage
  - 517 breaden reddenon- the three other ngmy ranked antibodies (A25-56.1, Din-176 and AD626) re-establish coverage
  - 318 over unique clusters in the sequence space (Fig. 4d). However, only ADG20 and Brii-198 cover and mitigate variants
  - that include the key F468V mutation (e.g., BA.4/5). Furthermore, Brii-198 covers the most diverse clusters that
  - 320 contain additional critical mutations at the F468 position, in addition to the surrounding residues in this epitope (Fig.
  - 4e and Supplementary Fig. 11). Thus, while any of the three antibodies would be complementary to ZCB11 by
  - nature of targeting a different epitope<sup>10</sup>, our breadth analysis aids in identifying the most complementary antibody by variant coverage.
  - 324
  - To quantify the impact of how individual mutations can drive antibody escape, an escape score  $(S_m^{\wedge})$  was
  - 326 computed for each mutation (*m*) within the synthetic lineages. This metric is a normalized product of the number of
  - 327 antibodies escaped by a given mutation and the mutation's frequency within the lineage (see Methods). When
  - examining individual RBD mutations across the synthetic lineages (Fig. 4f), it was revealed that T523P has the
  - highest escape score. Comparatively, DMS results showed that mutations at position 523 have a slightly negative
  - influence on RBD protein expression level <sup>19</sup>, which may explain its low occurrence in natural variants, having only
  - been observed in 70 sequences in the GISAID database. Furthermore, the combination of D339R, F486A and T523P
  - mutations in the simulated BA.1 lineages caused the most antibody escape among mutations not previously observed in major variants (Fig. 4f). Out of these, the positions 339 and 486 are mutated in BA.2.75 and XBB and their
  - sublineages. The top 50 mutations with the highest escape scores include K356T and R403K, which are present in
  - 335 the recently reported and highly mutated BA.2.86 variant and had not been previously reported in any other major

  - variant (Fig. 4f). Additionally, positions V445 and N481 were also mutated in BA.2.86. Taken together, this suggests
  - that DML-derived escape scores may reveal mutations or positions that emerge in future variants.
  - 338

### 339 **DISCUSSION**

- 340 The emergence of SARS-CoV-2 lineages with a high number of mutations has resulted in substantial viral immune
- evasion, including ineffective neutralization by previously developed therapeutic antibodies<sup>5</sup>. This rapid pace of viral
- 342 evolution has underscored the need for novel approaches to adequately profile antibody candidates and predict their
- robustness to emerging variants early on during drug development. To this end, we leverage DML, a machine
- 344 learning-guided protein engineering method to prospectively evaluate clinically relevant antibodies for their breadth
- 345 against potential future Omicron variants across a large mutational sequence space.
- 346
- 347 We first demonstrate the feasibility of assembling full-length RBD mutagenesis libraries with high fidelity using a
- large number of relatively short ssODNs in a one-pot reaction and obtaining library sizes in excess of 10<sup>8</sup>. This is
- 349 despite the fact that previous studies have reported a decrease in GGA when increasing the number of DNA
- 350 fragments<sup>39</sup>. Screening of these libraries for ACE2 binding and antibody escape yielded high-dimensional data sets
- 351 with combinatorial mutations spanning the entire RBD sequence, which is not obtainable through frequently
- 352 employed approaches such as DMS. In addition, the RBD library design can be updated to accommodate mutations

Frei, Gao et al., Deep learning on Omicron variants, 2023.

- present in emerging variants, and the average number of mutations can be titrated to generate data suitable for the
- training of machine learning models. This library design and screening approach could also be exploited to profile
- viral surface proteins from other rapidly evolving viruses such as influenza or HIV, two viruses which undergo
- 356 substantial antigenic drift that drives their immune escape<sup>57–59</sup>.
- 357 So far, the breadth of SARS-CoV-2 therapeutics has been assessed through the use of past variants and observed
- 358 mutations<sup>20,60–62</sup>. Measuring breadth in this way does not adequately predict long-term resistance against future
- 359 variants. The deployment of ensemble deep learning models to make predictions on synthetic mutational trajectories
- 360 of the RBD enabled an effective quantitative method to evaluate the breadth of each antibody based on its coverage
- of RBD mutational sequence space. DML predictions confirm that ZCB11 has exceptionally broad breadth to major
- 362 Omicron lineages that emerged in 2022, while many other antibodies fail against Omicron variants<sup>39</sup>. Furthermore,
- 363 our results suggest that the standard structure-based approach of selecting antibodies targeting different epitopes in a
- 364 cocktail does not sufficiently determine which combinations offer the most cumulative breadth. High breadth
- 365 cocktails would ensure that even if a variant escapes one antibody in the cocktail, it has a high chance to be re-
- captured by the other antibody thus potentially maintaining the clinical effectiveness of the therapy. For example,
- this occured with the combination antibody therapy from Eli Lilly (LY-CoV555+LY-CoV16), which continued to be
- used clinically when only a single antibody in the combination was effective after the emergence of Beta, Gamma
- and Delta variants<sup>22,63</sup>. Interestingly, a comprehensive search through a SARS-CoV-2 antibody database (Cov-
- AbDab, accessed April 2023)<sup>64</sup> reveals that a number of neutralizing antibodies discovered early in the pandemic
- from patients infected with the ancestral Wu-Hu-1 are still able to neutralize Omicron variants such as BA.5, BQ.1
- and XBB.1. DML could therefore be a powerful tool to identify such variant-resistant antibodies for therapeutic
- development.
- 374

Analysis of DML breadth predictions also highlights specific and positional mutations that are associated with 375 376 greater immune escape, with four such mutations being observed in the recently discovered and highly mutated 377 BA.2.86 variant. In contrast, other recently published deep learning methods, which rely on models trained using a combination of DMS and protein structure data, were able to only correctly forecast one new mutation each that 378 379 appeared in the XBB.1.5 and BQ.1 variants, respectively<sup>30,31,65</sup>. While this demonstrates the value of using protein 380 structural information to better infer higher-order effects between mutations, these models are still limited by the use 381 of low-distance (most often single-mutation) DMS data. Thus, it would be worthwhile to explore whether the use of 382 combinatorial DML data can further improve the accuracy and forecasting performance of models trained using a 383 multi-task objective, similar to those mentioned above.

- 384
- 385 The accuracy of antibody breadth predictions is dependent on having an accurate forecast of future mutations in the
- RBD. The use of deep learning models that predict ACE2 binding allowed us to capture evolutionary pressures
- 387 correlated with host receptor binding, which is a mandatory feature of any emerging SARS-CoV-2 variant<sup>66</sup>.
- 388 However, a myriad of other factors impact antigenic drift and variant emergence, such as transmissibility, host cell
- infectivity, crossover, reproductive rate, etc.<sup>67</sup>, thus generating training data related to these factors, for example
- through the use of an advanced pseudovirus mutational library screening system<sup>68</sup>, may further support the
- 391 generation of deep learning models that can predict future mutations and variants with higher accuracy.
- 392

Frei, Gao et al., Deep learning on Omicron variants, 2023.

### 393 METHODS

### 394 Construction of a high distance Omicron RBD library for yeast surface display

- 395 Synthetic ssODNs (oPools from IDT) were designed with either one or all possible combinations of two degenerate
- 396 NNK codons for each fragment (Supplementary Table 1). For each fragment, 136 ssODNs were designed (16 single
- 397 NNK codons and  $120 = \binom{16}{2}$  double NNK codon combinations). Each fragment was flanked by BsmBI recognition
- 398 sites and ~20 nt for second strand synthesis through PCR. For high fidelity library assembly, the overhangs were
- 399 optimized using the NEB ligase fidelity viewer (https://ligasefidelity.neb.com/viewset/run.cgi). Using the NEBridge<sup>®</sup>
- 400 Golden Gate Assembly Kit (NEB, E1602), individual fragments were assembled to full-length RBD gene segments.
- 401 A custom entry vector based on pYTK001 (addgene, Kit #1000000061) was designed. Double stranded fragments
- 402 were mixed with 75 ng entry vector in a 2:1 molar ratio. As suggested by the manufacturer's instructions, 2 µL NEB
- 403 Golden Gate Enzyme Mix was used. For the assembly, the following protocol was used: (42°C, 5 min  $\rightarrow$  16°C, 5
- 404 min) x  $30 \rightarrow 60^{\circ}$ C, 5 min. The assembled libraries were transformed into *E. coli* DH5 $\alpha$  ElecroMAX (Thermo Fisher
- 405 Scientific, 11319019), resulting in  $\sim 4 \times 10^8$  transformants. According to the manufacturer's instructions (Zymo,
- 406 D4201), the RBD library plasmid was extracted from *E. coli*.
- 407

408 The RBD library was PCR amplified and the yeast display vector (pYD1) was linearized using the restriction

- 409 enzyme BamHI (Thermo Fisher Scientific, FD0054). Both insert and backbone were column purified according to
- 410 the manufacturer's instructions (D4033) and drop dialyzed for 2 h using nuclease-free water (Millipore
- 411 VSWP02500). The RBD library insert and linearized pYD1 backbone were co-transformed into yeast (S. cerevisiae,
- 412 strain EBY100) using a previously described protocol<sup>69</sup>. Briefly, EBY100 (ATCC, MYA-4941) was grown overnight
- 413 in YPD [20 g/L glucose (Sigma-Aldrich, G8270), 20 g/L vegetable peptone (Sigma-Aldrich, 19942), and 10 g/L
- 414 yeast extract (Sigma-Aldrich, Y1625) in deionized water]. On the day of the library preparation, yeast cells from the
- 415 overnight culture were inoculated in 300 mL YPD at an OD<sub>600</sub> of 0.3. The cells were grown to an OD<sub>600</sub> of 1.6 before
- 416 washing the cells twice with 300 mL ice cold 1 M Sorbitol solution (Sigma-Aldrich, S1876). In a subsequent step,
- 417 the cells were conditioned using a solution containing 100 mM lithium acetate (Sigma-Aldrich, L6883) and 10 mM
- 418 DTT (Roche, 10197777001) for 30 min at 30 °C. This was followed by a third wash using 300 mL ice cold 1 M
- 419 Sorbitol solution. Using 50 µg insert and 10 µg pYD1 backbone, electrocompetent EBY100 were transformed using
- 420 2 mm electroporation cuvettes (Sigma-Aldrich, Z706086). The cells were recovered for 1 h in in recovery medium
- 421 (YPD:1 M Sorbitol solution mixed in a 1:1 ratio) before passageing the cells into selective SD-CAA medium [20 g/L
- 422 glucose (Sigma-Aldrich, G8270), 8.56 g/L NaH<sub>2</sub>PO<sub>4</sub>·H<sub>2</sub>O (Roth, K300.1), 6.77 g/L Na<sub>2</sub>HPO<sub>4</sub>·2H<sub>2</sub>O (Sigma-Aldrich,
- 423 1.06580), 6.7 g/L yeast nitrogen base without amino acids (Sigma-Aldrich, Y0626) and 5 g/L casamino acids (Gibco,
- 424 223120) in deionized water]. The cells were grown for 2 days at 30 °C. To estimate the transformation efficiency,
- 425 dilution plating was performed. Approximately  $2 \ge 10^8$  transformants were obtained.
- 426

## 427 Screening RBD libraries for ACE2-binding or non-binding

- 428 Yeast cells containing the RBD library plasmid were grown in SD-CAA for 18 24 h at 30°C. Surface display of
- 429 Omicron RBD was induced by passageing the cells into SG-CAA medium [20 g/L galactose (Sigma-Aldrich,
- 430 G0625), 8.56 g/L NaH<sub>2</sub>PO<sub>4</sub>·H<sub>2</sub>O, 6.77 g/L Na<sub>2</sub>HPO<sub>4</sub>·2H<sub>2</sub>O, 6.7 g/L yeast nitrogen base without amino acids and 5
- 431 g/L casamino acids in deionized water]. The cells were incubated at 23°C for 48 hours, as previously described<sup>37</sup>.
- 432 Approximately 10<sup>9</sup> cells were spun down by centrifugation at 3500 x g for 3 min and washed once with 5 mL cold
- 433 wash buffer [DPBS (PAN Biotech, P04-53500)+0.5% BSA (Sigma-Aldrich, A2153)+2 mM EDTA (Biosolve,

Frei, Gao et al., Deep learning on Omicron variants, 2023.

- 434 051423)+0.1% Tween20 (Sigma Aldrich, P1379)]. Next, cells were labeled with 50 nM of biotinylated human ACE2
- 435 protein (Acro Biosystems, AC2-H82E6) for 30 minutes at 4°C at 700 RPM on a shaker (Eppendorf, ThermoMixer
- 436 C). The cells were subsequently washed. In a secondary staining step, cells were labeled with Streptavidin-
- 437 Phycoerythrin (PE) (Biolegend 405203) (1:80 diluted) and anti-FLAG Tag Allophycocyanin (APC) (Biolegend
- 438 637308) (1:200 dilution) at 4°C for 30 min at 700 RPM. Afterwards, cells were centrifuged at 3500 x g for 3 min.
- 439 The supernatant was discarded and the tube was protected from light and stored on ice until sorting. Binding
- 440 (PE+/APC+) and non-binding (PE-/APC+) populations of yeast cells were collected by FACS (BD FACSAria
- 441 Fusion or BD Influx) (Fig. 2a,b and Supplementary Fig. 2). Collected cells were pelleted at 3500 x g for 3 min to
- 442 remove the FACS buffer. The cells were resuspended using SD-CAA and grown for two days at 30°C. The sorting
- 443 process was repeated until the desired populations were pure.
- 444

# 445 Screening RBD libraries for antibody binding or escape

- 446 The ACE2-binding population of yeast cells expressing the RBD library was grown and induced as described above.
- 447 Approximately 10<sup>8</sup> cells were pelleted by centrifugation at 3500 x g for 3 min at 4°C and washed once with 1 mL
- 448 wash buffer. The washed cells were incubated with antibodies (concentrations listed in Supplementary Table 2).
- 449 Suitable concentrations approximately corresponding to the EC<sub>90</sub> were experimentally determined beforehand
- 450 (Supplementary Fig. 12). Cells were incubated for 30 min at 4 °C and 700 RPM. After an additional washing step, a
- 451 secondary stain was performed using 5 ng/ml anti-human IgG-AlexaFluor647 (AF647) (Jackson Immunoresearch,
- 452 109-605-098) (1:200 dilution). The cells were incubated for 30 minutes at 4 °C and 700 RPM. Subsequently, cells
- 453 were washed and stained in a tertiary staining step using 1 ng/ml anti-FLAG-PE (1:200 dilution) for 30 min at 4 °C
- 454 and 700 RPM. Cells were pelleted by centrifugation at 3500 x g for 3 min at 4 °C. The supernatant was discarded and
- the tube was protected from light and stored on ice until sorting. Cells expressing RBD that maintained antibody-
- 456 binding (AF647+/PE+) or showed a complete loss of antibody binding (AF647-/PE+) were isolated using FACS (BD
- 457 Aria Fusion or Influx BD). Collected cells were pelleted by centrifugation at 3500 x g for 3 min at room temperature.
- 458 The FACS buffer was discarded and the cells were resuspended using SD-CAA. The cells were cultured for 48 h at
- 459 30 °C. The sorting process was repeated once for the binding population and twice for the non-binding population.
- 460 This procedure yielded pure binding and non-binding (escape) populations.
- 461

### 462 Deep sequencing of RBD libraries

- The pYD1 plasmid encoding the RBD library was extracted from yeast cells per manufacturer's instructions (Zymo, D2004). The mutagenized part of the RBD was PCR amplified using custom designed primers for seq-library A and seq-library B (Supplementary Table 5). In a second PCR amplification step, sample specific barcodes (Illumina
- Nextera) were introduced, which allowed pooling of individual populations for sequencing. The populations were
- 467 sequenced using the Illumina MiSeq v 3 kit which allows for 2 X 300 paired-end sequencing.
- 468

## 469 Preprocessing of deep sequencing data

- 470 Sequencing reads were paired, quality trimmed and merged using the BBTools suite<sup>70</sup> with a quality threshold of
- 471 qphred R>25. RBD nt sequences were then extracted using custom R scripts, followed by translation to aa sequences.
- 472 Read counts per sequence were calculated and singletons (read count = 1) were discarded. Sequencing datasets used
- 473 for training machine and deep learning models were created by combining the binding and non-binding datasets.
- 474 Sequences present in both populations were removed.

Frei, Gao et al., Deep learning on Omicron variants, 2023.

#### 475 476 Binding scores for heatmaps shown in Fig. 2c-e were created by calculating aa counts per position in the RBD from both binding and non-binding sequences. WT (BA.1) aa residues were then removed, relative frequencies were 477 478 calculated with a pseudocount of 1 added, and final binding scores were calculated as binding frequencies divided by 479 non-binding frequencies. The results were then log-transformed before plotting in the heatmap for visualization. 480 481 Training and testing machine and deep learning models All machine learning code and models were built in Python (3.10.4)<sup>71</sup>. For data processing and visualization, numpy 482 483 (1.23.3), pandas (1.4.4), matplotlib (3.5.3) and seaborn (0.12.0) packages were used. Baseline benchmarking models 484 were built using Scikit-Learn (1.0.2), while Keras (2.9.0) and Tensorflow (2.9.1) were used to build the MLP and 485 CNN models. 486 487 Each model was trained using 80/10/10 train-val-test data random splits. RBD library protein sequences (from seq-488 library A or B deep sequencing data) were one-hot encoded prior to being used as inputs into the models. For the 489 CNN, the 2D one-hot encoded matrix was used as the input, while for others, the matrix was flattened into a one-490 dimensional vector. All reported model performances were evaluated using 5-fold cross-validation, and evaluated 491 based on the metrics for accuracy, f1, MCC, precision, and recall. 492 493 When training baseline machine learning models, class balancing was performed through random downsampling 494 from the majority class so that it was equal to the counts from the minority class; this was performed at each ED. 495 RBD sequences that were not sampled from the majority class were then reserved separately as additional "unseen 496 sequences". These were then combined with the held-out test set during model evaluation to ensure that the models 497 could perform well with an imbalanced test set. Hyperparameter optimization was performed during model training 498 using up to 30 rounds of RandomSearchCV (from Scikit-Learn), and the best model performances were kept for 499 comparison to deep learning models. 500 501 To train the deep learning models, exhaustive hyperparameter search was performed on the CNN models to optimize 502 performance through the hyperparameters listed in (Supplementary Table 4). The training dataset was balanced at 503 different ratios (see Minority Ratio row, Supplementary Table 3) while validation and test sets remained unbalanced 504 to appropriately evaluate MCC, precision and recall scores on imbalanced data. Dataset balancing was performed 505 through rejection sampling using a custom dataset sampler created in Tensorflow. To prevent data leakage during 506 training of the models for ensembles, the held-out test set was fixed, while multiple models were trained on random 507 splits of the training and validation sets to make sure each model learned slightly different parameters of the dataset, 508 while being evaluated on the same held-out test sequences. 509 510 Predictions made with ensemble deep learning models 511 Natural and in silico generated synthetic RBD variant sequences were assigned "binding", "escape" and "uncertain" 512 labels for ACE2 and antibodies using an ensemble of trained models. For a given RBD sequence, each model assigns 513 a binding label if output P > 0.75, escape if output P < 0.25, or uncertain otherwise. For each of the two libraries 514 (seq-library A and B), the three models with the highest MCC scores were used to independently assign labels to

515 each sequence, followed by majority voting, where the most common label was taken as the label for each variant.

Frei, Gao et al., Deep learning on Omicron variants, 2023.

- 516 The labels from models trained with seq-library A or seq-library B were used to determine the final label for each
- 517 variant: "binding" if both libraries agree on a "binding" label, "escape" if either library predicts "escape", and
- 518 "unsure" otherwise. For experimentally measured variants, antibodies-variant pairs were labeled as "escape" if their
- 519 measured  $K_d$  was > 100nM or IC50 was > 1ug/mL.
- 520
- 521
- 522

# 523 Calculating mutational probabilities of the RBD based on SARS-CoV-2 genome data

- 524 To generate the mutational probability matrices used for synthetic lineages, SARS-CoV-2 spike protein sequences
- 525 were obtained from the GISAID database (most recent access of June 2023) The regions corresponding to the RBD
- 526 were extracted, along with the date when each sequence was deposited into the database. Sequences were separated
- 527 by the year they were added (e.g., 2021 or 2022). From these sequences, mutations were counted at each position per
- 528 position, and per aa. Mutational frequencies at each position were calculated using these counts. Finally, a log
- 529 softmax function was applied to obtain mutational probabilities for each position. For each position, only residues
- that were observed in GISAID sequences were counted, while all unseen residues were not included in the softmax
- transform, preventing them from being generated in synthetic lineages.
- 532

### 533 In silico generation of synthetic Omicron lineages

- 534 Using BA.1 as the initial seed variant, *in silico* sequences were generated in a stepwise fashion over six rounds of
- 535 mutations. In the first round, single mutations were randomly generated across the RBD. Positions and aa for each
- 536 mutational round were selected using probabilities from the 2021 or 2022 substitution matrices; as a control,
- 537 sequences were also generated using no substitution matrix (where all mutations were sampled from a uniform
- 538 probabilities distribution). Then binding probability scores were assigned to variants in each generation by taking the
- average of all P predicted by each of the ACE2 models in the ensemble. The top 100 variants ranked by ACE2-
- 540 binding *P* were used as seed sequences for the next round of mutations. For each round, new variants were only
- accepted if they contained mutations not previously seen in other generated variants, or else the process was repeated
- square again and new mutations selected until the maximum number of variants were reached (250,000).
- 543

### 544 Calculating escape scores

- n An escape score  $(S_m)$  was calculated that aims to quantify the impact of a given mutation on driving escape from
- 546 the antibodies tested herein and was calculated by:

547 
$$S_m = \frac{\sum_{E=0}^a \frac{\left(E * f * d\right)}{n}}{N}$$

- 548
- 549  $S_m$  is the escape score of a mutation m, E is the number of antibodies that are predicted to escape from m, and within 550 the group of sequences with the same number of E, f is the frequency that m appeared in the sequence group,  $\underline{d}$  is the 551 mean of sequence ED from BA.1, n is the number of sequences, N is how many times one mutation appeared in
- different groups of *E*, *a* is the total number of groups, according to how many antibodies were tested (here, a = 6).
- 553 For better visualization, the adjusted escape score was used (Fig 4a, f) and is calculated by the following equation:
- 554  $S_{adj} = log_{(10)}S_m + 7.$
- 555

Frei, Gao et al., Deep learning on Omicron variants, 2023.

### 556 Additional statistical analysis and plots

- 557 Statistical analysis was performed using Python (3.10.4) with the Scipy package (1.9.3). Dimensionality reduction
- 558 was performed using UMAP-learn (0.5.3). Graphics were generated using matplotlib (3.5.3), seaborn (0.12.0), and
- 559 ggtree (3.8.0). Sequence logo plots were created using Seq2Logo (5.29.8)<sup>72</sup> or the dsmlogo package from the Bloom
- 560 Lab (<u>https://github.com/jbloomlab/dmslogo</u>).
- 561
- 562 The KL-divergence was calculated by adapting a recently described method<sup>30</sup>. In short, a probability-weighted KL
- logo plot was used to visualize differences between a subset of sequences to the full background dataset. Let M1 =
- 564 (f<sub>1</sub>, f<sub>2</sub>, f<sub>3</sub>..., f<sub>n</sub>) represent the position frequency matrix (PFM) of the background sequence set, where the length of
- the initial sequence is n = 201 and each frequency  $f_i = (a_1, a_2, a_3, \dots, a_{20})^T$ , represents the frequency of each aa per
- position i. At the same time, M2 =  $(f_1, f_2, f_3, ..., f_n)$  represents the PFM of the subset of sequences, each  $f_i = (a_1, f_2)$
- 567  $a_2$ ,  $a_3$ ,  $\dots a_{20}$ )<sup>T</sup>. The KL divergence at each position is computed as:

568 
$$D_{KL}(f_i'||f_i) = \sum_{i=1}^{20} a_i' \cdot \ln\left(\frac{a_i'}{a_i}\right)$$

- 569 The KL divergence is used to set the total height at each position in the logo plots (Fig. 4e). The height and direction
- of each aa letter are calculated through probability-weighted normalization as part of the Seq2Logo package using:

. .

571 
$$h(a_i') = \frac{a_i' \cdot \ln\left(\frac{a_i}{a_i}\right)}{\sum_{i=1}^{20} a_i' \cdot \left|\ln\left(\frac{a_i'}{a_i}\right)\right|} D_{KL}(f_i'||f_i)$$

572

#### 573 Data availability

- 574 The main data supporting the results in this study are available within the paper and its Supplementary Information.
- 575 The raw and analysed datasets generated during the study will be made available at: https://github.com/LSSI-
- 576 ETH/Omicron\_DML.
- 577

### 578 Code availability

- 579 The code and models used to perform the work in this study will be available at the following:
- 580 <u>https://github.com/LSSI-ETH/Omicron DML</u>.
- 581

### 582 Acknowledgments

- 583 We thank the ETH Zurich D-BSSE Single Cell Unit and the ETH Zurich D-BSSE Genomics Facility for support.
- 584 This work was supported by the Botnar Research Centre for Child Health (FTC COVID-19, to S.T.R.)
- 585

## 586 Author contributions

- 587 L.F., B.G., J.H., J.M.T and S.T.R. developed the methodology. L.F., J.M.T. designed and generated mutagenesis
- 588 libraries, L.F., performed screening experiments, B.G. and J.H., analyzed the sequencing data and performed deep-
- 589 learning analyses. L.F., B.G., J.H. and S.T.R. wrote the manuscript, with input from all other authors.
- 590

# 591 **Competing interests**

- 592 C.R.W. is an employee of Alloy Therapeutics (Switzerland). C.R.W. and S.T.R. may hold shares of Alloy
- 593 Therapeutics. S.T.R. is on the scientific advisory board of Alloy Therapeutics.

Frei, Gao et al., Deep learning on Omicron variants, 2023.

#### 594 SUPPLEMENTARY FIGURES





# 596

Supplementary Figure 1. After assembling the RBD sequence from short fragments and transformation into *E. coli*,
single colonies were picked and colony PCRs (cPCR) were performed. For the amplification, primers binding
directly upstream and downstream of the RBD were used. As a control, WT BA.1 plasmid was used. When running

600 the cPCR products on a 2% agarose gel, 46 out of 47 reactions showed the right band size of 663 base pairs (bp)

601 (wrongly assembled variant highlighted in red), roughly corresponding to 98% correctly assembled full length RBD

602 sequences.

Frei, Gao et al., Deep learning on Omicron variants, 2023.



604 Supplementary Figure 2. Representative FACS dot plots of yeast RBD libraries during antibody screening; sorting

605 gates for binding and non-binding (escape) populations are shown.

Frei, Gao et al., Deep learning on Omicron variants, 2023.



Frei, Gao et al., Deep learning on Omicron variants, 2023.



Supplementary Figure 3. Heatmaps showing binding scores per position across the RBD for libraries sorted against
 each target (ACE2 or antibodies, respectively). Blue regions indicate mutations seen in greater frequency in the
 binding variant pool, while red regions indicate mutations with greater frequency in escape variants. WT (BA.1)

- 611 residues are depicted by grey boxes.
- 612



Frei, Gao et al., Deep learning on Omicron variants, 2023.

- 614 Supplementary Figure 4. Line plots show the frequencies of selected mutations in the binding and escape fractions
- 615 of the deep sequencing data. The selected mutations have been observed in previously identified Omicron
- 616 sublineages.
- 617

#### Total Unique Counts after NGS processing





619 Supplementary Figure 5. Total unique sequences (aa) in each deep sequencing dataset (following pre-processing).

Frei, Gao et al., Deep learning on Omicron variants, 2023.



620

621 Supplementary Figure 6. Number of unique sequences (aa) in each dataset per ED from WT BA.1 RBD sequence.

622 To allow visual comparison between datasets, the maximum of the y-axis in all antibody datasets has been set to the

623 highest count in all datasets (20,000).

Frei, Gao et al., Deep learning on Omicron variants, 2023.



624

625 Supplementary Figure 7. Barplots show MCC scores of all baseline machine learning models: Logistic Regression

626 (Log Reg), Naive Bayes (NB), Radial Basis Function kernel SVM (RBF), Random Forest (RF), Stochastic Gradient

627 Descent (SGD), and deep learning models: MLP and CNN, for **a**, seq-library A and **b**, seq-library B. All scores were

628 evaluated through 5-fold cross-validation with a 80/10/10 train-val-test split.

Frei, Gao et al., Deep learning on Omicron variants, 2023.



629

630 Supplementary Figure 8. CNN model performances on test sequences based on ED from BA.1; shown area,

631 accuracy, **b**, MCC, and **c**, precision. All scores shown are combined results from 5-fold cross-validation with a

632 80/10/10 train-val-test split.

Frei, Gao et al., Deep learning on Omicron variants, 2023.



Proprotion of labels by distance in BA.1 lineages generated with 2021 mutational frequencies



с

Proprotion of labels by distance in BA.1 lineages generated with random probabilities



634

635 Supplementary Figure 9. Percent of predicted binding and escape variants per ED (from BA.1) for each antibody.

636 Predictions were run on 10 sets of synthetic lineages: BA.1-derived lineages based on GISAID mutational

Frei, Gao et al., Deep learning on Omicron variants, 2023.

- 637 frequencies from **a**, 2021, **b**, 2022 or **c**, randomized probabilities (see Methods). Each synthetic lineage contains up
- 638 to 250,000 sequences.
- 639



640

641 Supplementary Figure 10. a, predicted total antibody breadth and b, antibody breadth per ED (from BA.1) on

synthetic lineages (BA.1-derived lineages based on 2021 GISAID mutational frequencies, see Methods). **c**, predicted

total antibody breadth and **d**, antibody breadth per ED (from BA.1) on randomized synthetic lineages (BA.1-derived

644 lineages based on uniform random sampling frequencies, see Methods).

Frei, Gao et al., Deep learning on Omicron variants, 2023.





646 Supplementary Figure 11. UMAPs show synthetic lineage variants in protein sequence space. a, sequences from 647 Figure 4e, coloured to indicate their ED relative to WT (BA.1), and b, coloured to highlight sequences that bind (in 648 gray) or escape (in red) from ZCB11. c, dimensionality reduced subsample of sequences taken from synthetic 649 lineages from 2021, 2022 or random (none) probabilities coloured by their ED (from BA.1) and d, by the 650 probabilities used to generate each lineage ("None" indicates sequences that were generated by random sampling 651 from a uniform probability distribution across the RBD, with all as substitutions allowed)

651 652



Concentration [nM]

Frei, Gao et al., Deep learning on Omicron variants, 2023.

- 654 Supplementary Figure 12. Titration curves of individual antibodies tested against yeast-displayed Omicron BA.1
- 655 RBD.
- 656

### 657 SUPPLEMENTARY TABLES

	Sub-library 1	Sub-library 2	Sub-library 3	Sub-library 4
Fragment 1	tgatgatagctatcggcacacgtctcge tccAATATCACGAACCTT TGTCCTTTCGATGAGGT CTTCAATGCTACTAGAT tcgcagagacggaactgagtcggcgc cgatg	tgatgatagctatcggcacacgtctcg <b>ct</b> <b>cc</b> AATATCACGAACCTTT GTCCTTTCGATGAGGTCT TCAATGCTACTAGATTC GCATCCGTG <b>tatg</b> cgagacgga actgagtcggcgccgatg	tgatgatagctatcggcacacgtctcgc tccAATATCACGAACCTT TGTCCTTTCGATGAGGT CTTCAATGCTACTAGAT TGGCATCCGTGTATGCA TGGAATAgaaaggagacggaa ctgagtcggcgcccgatg	
Fragment 2	gatagcggactttcggtcaacgtctctt cgcATCCGTGTATGCATG GAATAGAAAGAGAAATT AGTAATTGTGTAGCGGA CTacagtgagacgtctgaatgtacaa gcaaccc	gatagcggacttteggtcaacgtetegt atgCATGGAATAGAAAGA GAATTAGTAATTGTGTA GCGGACTACTCTGTACTT TAtaactgagacgtetgaatgtacaag caacce	gatagcggactttcggtcaacgtctcag     gatagcggactttcggtcaacgtctcag       aaaGAGAATTAGTAATTG     cttCAATGCTACTAGAT       TGTAGCGGACTACAGTG     CGCATCCGTGTATGCA       TACTTTATAACTTGGCCC     GGAATAGAAAAGAGAA       q     ccttcgagacgtctgaatgtacaagcaa       ccc     gcaaccc	
Fragment 3	aagtgggccgagcctggactcgtctct	aagtgggccgagcctggactcgtctcat	aagtgggccgagcctggactcgtctcc	aagtgggccgagcctggactcgtctca
	acagTGTACTTTATAACT	aacTTGGCCCCCTTCTTTA	ccttCTTTACATTCAAGTG	gtaaTTGTGTAGCGGACT
	TGGCCCCCCTTCTTACA	CATTCAAGTGTTACGGT	TTACGGTGTATCTCCCA	ACAGTGTACTTTATAAC
	TTCAAGTGTTACGGTGT	GTATCTCCCACCAAGTtg	CCAAGTTGAATGATCTA	TTGGCCCCCTTCTTTAC
	ATCtcccagagacgcagctggttcc	aatgagacgcagctggttcctgcgtga	Tgctttgagacgcagctggttcctgcgt	ATtcaaggagacgcagctggttcctg
	tgcgtgagc	gc	gagc	cgtgagc
Fragment 4	tcgagactcgggatgacagccgtetee	tcgagactcgggatgacagccgtctctt	tcgagactcgggatgacagccgtctctg	tcgagactcgggatgacagccgtctctt
	tcccACCAAGTTGAATGA	gaaTGATCTATGCTTTAC	cttTACAAACGTTTACGCC	caaGTGTTACGGTGTATC
	TCTATGCTTTACAAACG	AAACGTTTACGCCGATA	GATAGTTTCGTAATTAG	TCCCACCAAGTTGAATG
	TTTACGCCGATAGTTTC	GTTTCGTAATTAGAGGC	AGGCGATGAAGTGCGTC	ATCTATGCTTTACAAAC
	Gtaattgagacgtcatagtacetegg	Gatgaagagacgtcatagtacctcggg	agatcgagacgtcatagtacctcgggt	Gtttacgagacgtcatagtacctcggg
	gtacca	tacca	acca	tacca
Fragment 5	acttactcaggttattgcttcgtctcgta	acttactcaggttattgcttcgtctcgatg	acttactcaggttattgcttcgtctcc <b>aga</b>	acttactcaggttattgcttcgtctcgttt
	atTAGAGGCGATGAAGT	aAGTGCGTCAGATCGCA	tCGCACCAGGCCAGACG	aCGCCGATAGTTTCGTA
	GCGTCAGATCGCACCA	CCAGGCCAGACGGGCAA	GGCAATATAGCAGATTA	ATTAGAGGCGATGAAGT
	GGCCAGACGGGCAATA	TATAGCAGATTATAATTa	TAATTATAAGCTGCCTG	GCGTCAGATCGCACCAG
	TAGCagattgagacggaacgccc	taaggagacggaacgcccatctagcg	A <b>tgac</b> tgagacggaacgcccatctag	gccaggagacggaacgccatctagc
	atctagcggctg	gctg	cggctg	ggctg
Fragment 6	gcgtcttgaatgctcggtcccgtctcca gatTATAATTATAAGCTG CCTGATGACTTCACCGG CTGTGTGGATAGCTTGGA Acagcagagacggcttgcgaagtcta cattgg	gcgtcttgaatgctcggtcccgtctctat aaGCTGCCTGATGACTTC ACCGGCTGTGTGTGATAGC TTGGAACAGCAATAAAC tagatgagacggcttgcgaagtctacat tgg	gcgtcttgaatgctcggtcccgtctca <b>tg</b> acTTCACCGGCTGTGTGA TAGCTTGGAACAGCAAT AAACTAGATTCCAAG <b>gtg</b> tcgagacggcttgcgaagtctacattgg	gcgtcttgaatgctcggtcccgtctcgg ccaGACGGGCAATATAG CAGATTATAATTATAAG CTGCCTGATGACTTCAC CGGctgtggagacggcttgcgaagt ctacattgg
Fragment 7	tatatgaatgcgacctagaacgtctca <b>c</b>	tatatgaatgcgacctagaacgtctccta	tatatgaatgcgacctagaacgtctcgg	tatatgaatgcgacctagaacgtctcgc
	agcAATAAACTAGATTCC	gaTTCCAAGGTGTCTGGC	tgtCTGGCAATTACAATT	tgtGTGATAGCTTGGAAC
	AAGGTGTCTGGCAATTA	AATTACAATTATTTGTAC	ATTTGTACCGTCTGTTCC	AGCAATAAACTAGATTC
	CAATTATTTGTACCGTCt	CGTCTGTTCCGTAAAAgc	GTAAAAGCAATTTGAAA	CAAGGTGTCTGGCAATta
	gttcgagacgacggccgggaaaggt	aatgagacgacggccgggaaaggtac	Ccatttgagacgacggccgggaaag	caagagacgacggccgggaaaggta
	acgcg	gcg	gtacgcg	cgcg
Fragment 8	cgcggtatgggaggatcaagcgtctc	cgcggtatgggaggatcaagcgtctca	cgcggtatgggaggatcaagcgtctcc	cgcggtatgggaggatcaagcgtctct
	ctgttCCGTAAAAGCAATT	gcaaTTTGAAACCATTTGA	cattTGAAAGAGACATAA	tacaATTATTTGTACCGTC
	TGAAACCATTTGAAAG	AAGAGACATAAGCACTG	GCACTGAAATTTACCAA	TGTTCCGTAAAAGCAAT
	AGACATAAGCACTGAA	AAATTTACCAAGCAGGG	GCAGGGAACAAACCGTG	TTGAAACCATTTGAAAG
	ATTTaccaagagacggggccaata	aacaagagacggggccaatagagag	CAacggcgagacggggccaataga	AGAcataagagacggggccaatag
	gagaggctcct	gctcct	gaggctcct	agaggctcct
Fragment 9	ctctcactcgctaggaggcacgtctcta	ctctcactcgctaggaggcacgtctcga	ctctcactcgctaggaggcacgtctcaa	ctctcactcgctaggaggcacgtctcac
	ccaAGCAGGGAACAAAC	acaAACCGTGCAACGGCG	cggCGTAGCTGGCTTTAA	ataAGCACTGAAATTTAC
	CGTGCAACGGCGTAGCT	TAGCTGGCTTTAACTGTT	CTGTTATTTCCCATTAAG	CAAGCAGGGAACAAAC
	GGCTTTAACTGTTATTT	ATTTCCCATTAAGATCTT	ATCTTATAGTTTCAGAC	CGTGCAACGGCGTAGCT
	CCcattagagacgaatgtaaaacaat	Atagttgagacgaatgtaaaacaatggt	Ctacgtgagacgaatgtaaaacaatgg	ggcttgagacgaatgtaaaacaatggtt
	ggttact	tact	ttact	act

Frei, Gao et al., Deep learning on Omicron variants, 2023.

Fragment 10	gcatcgatacataaaacatgcgttctcce attAAGATCTTATAGTTT CAGACCTACGTATGGA GTCGGGCATCAGCCGTA CCgtgttgagacgctgtccatcggtt gcccaaa	gcatcgatacataaaacatgcgtctca <b>ta</b> gtTTCAGACCTACGTATG GAGTCGGGCATCAGCCG TACCGTGTTGTGGGTTC <b>tttc</b> agagacgctgtccatcggttgcccaaa	gcatcgatacataaaacatgcgtctcct acgTATGGAGTCGGGCAT CAGCCGTACCGTGTTGT GGTTCTTTCATTTGAACt gctggagacgctgtccatcggttgccc aaa	gcatcgatacataaaacatgcgttct <b>g</b> gctTTAACTGTTATTTCCC ATTAAGATCTTATAGTT TCAGACCTACGTATGGA gtcgggagacgctgtccatcggttgcc caaa
Fragment 11	gtgttaagtgtctatcaccccgtctccgt gtTGTGGTTCTTTCATTT GAACTGCTGCACGCGCC CGCAACCGTATGCGGG CCGAAGAAATCAACGga ttagagacggtcgctgtactaatagttg t	gtgttaagtgtctatcaccccgtctccttt cATTTGAACTGCTGCACG CGCCCGCAACCGTATGC GGGCCGAAGAAATCAAC Ggattagagacggtcgctgtactaata gttgt	gtgttaagtgtctatcaccccgtctcctg ctGCACGCGCCCCGCAACC GTATGCGGGCCGAAGAA ATCAACGgattagagacggtcgc tgtactaatagttgt	gtgttaagtgtctatcaccccgtctca <b>gt</b> c <b>g</b> GGCATCAGCCGTACC GTGTTGTGGGTTCTTTCAT TTGAACTGCTGCACGCG C <b>ccgc</b> agagacggggccgtttcccgc atataa
Fragment 12	-	-	-	acgccaggttgtatccgcatcgtctccc cgcAACCGTATGCGGGCC GAAGAAATCAACGgattag agacggtcgctgtactaatagttgt

659

9 Supplementary Table 1. Sequences for fragments by sub-library. Sequences marked with uppercase letters are

derived from the RBD open reading frame. The NNK codons are exclusively in this region. Bold lowercase

sequences are the four nt homologies for GGA. The remaining lowercase sequences contain BsmBI recognition sites

- and primer binding sites for double strand synthesis.
- 663

Therapeutic antibodies	Concentration [nM]
S2X259	12.5
S2H97	2.5
COV2-2196	12.5
ZCB11	6.25
2-7	60
ADG20	7.5
A23-58.1	5
Brii-198	10

- 664 **Supplementary Table 2:** Antibody concentrations used for FACS of yeast displayed RBD libraries. The
- 665 concentrations were determined based on the titration curves shown in Supplementary Fig. 12.

Frei, Gao et al., Deep learning on Omicron variants, 2023.

		Paired and fil	tered Reads
Population	Antigen	Binding	Non-Binding
seq-Library A	ACE2	1.92E+06	1.75E+06
seq-Library B	ACE2	2.32E+06	1.60E+06
seq-Library A, ACE2 binding	2-7	3.27E+05	5.52E+05
seq-Library A, ACE2 binding	A23-58.1	1.20E+06	1.08E+06
seq-Library A, ACE2 binding	ADG20	8.37E+05	8.83E+05
seq-Library A, ACE2 binding	Brii-198	9.40E+05	3.59E+05
seq-Library A, ACE2 binding	COV2-2196	5.59E+05	9.40E+05
seq-Library A, ACE2 binding	S2H97	5.37E+05	7.23E+05
seq-Library A, ACE2 binding	S2X259	6.22E+05	5.59E+05
seq-Library A, ACE2 binding	ZCB11	5.52E+05	8.30E+05
seq-Library B, ACE2 binding	2-7	1.02E+06	9.32E+05
seq-Library B, ACE2 binding	A23-58.1	9.12E+05	7.38E+05
seq-Library B, ACE2 binding	ADG20	9.38E+05	9.03E+05
seq-Library B, ACE2 binding	Brii-198	9.70E+05	5.48E+05
seq-Library B, ACE2 binding	COV2-2196	7.66E+05	9.70E+05
seq-Library B, ACE2 binding	S2H97	8.42E+05	4.43E+05
seq-Library B, ACE2 binding	S2X259	7.50E+05	7.66E+05
seq-Library B, ACE2 binding	ZCB11	9.32E+05	6.09E+05

## **Supplementary Table 3:** Deep sequencing statistics for sorted RBD libraries.

## 

Parameter	MLP	ProtCNN
Learning Rate	0.01-0.0001	
Optimizer Adam, SGD		, SGD
Minority Ratio (dataset balance)	0.1 - 0.5	
Epochs	15 - 75	
Test/Val ratio	0.1-0.25	

MLP Parameters		
Dense Dimensions	32-512	
# Dense Layers	1-3	
Dense Dropout	0-0.5	

CNN Parameters	]	
Kernel Size		3-21
Stride		1-3
Filter Number		32-512
Padding		"Same", 1
Pool Size		1-3
Pool Stride		1-3
Dilation Rate		2-5
Residual Blocks		1-3

# **Supplementary Table 4:** Hyperparameter Search Conditions for CNN and MLP models

Primer name	Sequence (5' to 3')
seq-library A fwd	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGATGTGCCCGATTATGCG
seq-library A rev	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGCCGTTGCACGGTTTGTT
seq-library B fwd	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGTGTTACGGTGTATCTCCC

	sec	l-library B rev	GTCTCGTGGGGCTCGGAGATGTGTATAAGAGACAGCTCACTTGTCATCATCGTC C			
672	Sup	Supplementary Table 5: Primers used to amplify seq-libraries A and B in a targeted fashion for subsequent deep				
673	sequencing.					
674						
675	Sup	plementary File:				
676	Exp	erimentally measure	d binding affinity and neutralization of antibodies against SARS-CoV-2 variants from			
677	publ	lications.				
678	1.	Jones, B. E. et al.	The neutralizing antibody, LY-CoV555, protects against SARS-CoV-2 infection in			
679		nonhuman prima	tes. Sci. Transl. Med. 13, (2021).			
680	2.	Hansen, J. et al. S	Studies in humanized mice and convalescent humans yield a SARS-CoV-2 antibody			
681		cocktail. Science	<b>369</b> , 1010–1014 (2020).			
682	3.	Hoffmann, M. et	al. SARS-CoV-2 variants B.1.351 and P.1 escape from neutralizing antibodies. Cell			
683		<b>184</b> , 2384–2393.	e12 (2021).			
684	4.	McCallum, M. et	al. Molecular basis of immune evasion by the Delta and Kappa SARS-CoV-2			
685		variants. Science	<b>374</b> , 1621–1626 (2021).			
686	5.	Cao, Y. et al. On	nicron escapes the majority of existing SARS-CoV-2 neutralizing antibodies. Nature			
687		<b>602</b> , 657–663 (20	021).			
688	6.	Shrestha, L. B., F	Foster, C., Rawlinson, W., Tedla, N. & Bull, R. A. Evolution of the SARS-CoV-2			
689		omicron variants	BA.1 to BA.5: Implications for immune escape and transmission. Rev. Med. Virol.			
690		<b>32</b> , (2022).				
691	7.	Pinto, D. et al. C	ross-neutralization of SARS-CoV-2 by a human monoclonal SARS-CoV antibody.			
692		<i>Nature</i> <b>583</b> , 290–	-295 (2020).			
693	8.	Westendorf, K. e	t al. LY-CoV1404 (bebtelovimab) potently neutralizes SARS-CoV-2 variants. Cell			
694		<i>Rep.</i> <b>39</b> , 110812	(2022).			
695	9.	Arora, P. et al. O	micron sublineage BQ.1.1 resistance to monoclonal antibodies. Lancet Infect. Dis.			
696		<b>23</b> , 22–23 (2023)				
697	10.	Wang, Q. et al. A	antibody evasion by SARS-CoV-2 Omicron subvariants BA.2.12.1, BA.4 and BA.5.			
698		<i>Nature</i> <b>608</b> , 603-	-608 (2022).			
699	11.	CDC. COVID-19	Vaccines for People Who Are Moderately or Severely Immunocompromised.			
700		Centers for Dised	ase Control and Prevention https://www.cdc.gov/coronavirus/2019-			
701		ncov/vaccines/ree	$commendations/immuno.html?s\_cid=10483:immunocompromised\%20 and\%20 covid$			
702		%20vaccine:sem	.ga:p:RG:GM:gen:PTN:FY21 (2023).			
703	12.	Lee, A. R. Y. B.	et al. Efficacy of covid-19 vaccines in immunocompromised patients: systematic			
704		review and meta-	analysis. BMJ <b>376</b> , e068632 (2022).			
705	13.	Martinelli, S., Pa	scucci, D. & Laurenti, P. Humoral response after a fourth dose of SARS-CoV-2			
706		vaccine in immu	nocompromised patients. Results of a systematic review. Front Public Health 11,			
707		1108546 (2023).				

Frei, Gao et al., Deep learning on Omicron variants, 2023.

708 14. Casadevall, A. & Focosi, D. SARS-CoV-2 variants resistant to monoclonal antibodies in immunocompromised patients constitute a public health concern. J. Clin. Invest. 133, (2023). 709 710 15. Considerations for implementing and adjusting public health and social measures in the context of 711 COVID-19. https://www.who.int/publications/i/item/who-2019-ncov-adjusting-ph-measures-2023.1 712 (2023). 16. Anti-SARS-CoV-2 Monoclonal Antibodies. COVID-19 Treatment Guidelines 713 714 https://www.covid19treatmentguidelines.nih.gov/tables/variants-and-susceptibility-to-mabs/. 715 17. Jain, T. et al. Biophysical properties of the clinical-stage antibody landscape. Proc. Natl. Acad. Sci. 716 U. S. A. 114, 944–949 (2017). 18. Hanning, K. R., Minot, M., Warrender, A. K., Kelton, W. & Reddy, S. T. Deep mutational scanning 717 for therapeutic antibody engineering. Trends Pharmacol. Sci. 43, 123-135 (2022). 718 719 19. Starr, T. N. et al. Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. Cell 182, 1295-1310.e20 (2020). 720 721 20. Starr, T. N. et al. SARS-CoV-2 RBD antibodies that maximize breadth and resistance to escape. 722 Nature 597, 97-102 (2021). 723 21. Starr, T. N. et al. Prospective mapping of viral mutations that escape antibodies used to treat COVID-724 19. Science 371, 850-854 (2021). 22. Starr, T. N., Greaney, A. J., Dingens, A. S. & Bloom, J. D. Complete map of SARS-CoV-2 RBD 725 726 mutations that escape the monoclonal antibody LY-CoV555 and its cocktail with LY-CoV016. Cell 727 *Rep Med* **2**, 100255 (2021). 728 23. Greaney, A. J. et al. Mapping mutations to the SARS-CoV-2 RBD that escape binding by different 729 classes of antibodies. Nat. Commun. 12, 4196 (2021). 24. Francino-Urdaniz, I. M. et al. One-shot identification of SARS-CoV-2 S RBD escape mutants using 730 yeast screening. Cell Rep. 36, 109627 (2021). 731 732 25. Callaway, E. Why a highly mutated coronavirus variant has scientists on alert. *Nature* (2023) doi:10.1038/d41586-023-02656-9. 733 26. Yang, S. et al. Antigenicity and infectivity characterization of SARS-CoV-2 BA.2.86. bioRxiv 734 735 2023.09.01.555815 (2023) doi:10.1101/2023.09.01.555815. 736 27. Uriu, K. et al. Transmissibility, infectivity, and immune evasion of the SARS-CoV-2 BA.2.86 variant. Lancet Infect. Dis. 0, (2023). 737 28. Greaney, A. J., Starr, T. N. & Bloom, J. D. An antibody-escape estimator for mutations to the SARS-738 CoV-2 receptor-binding domain. Virus Evol 8, veac021 (2022). 739 29. Makowski, E. K., Schardt, J. S., Smith, M. D. & Tessier, P. M. Mutational analysis of SARS-CoV-2 740 741 variants of concern reveals key tradeoffs between receptor affinity and antibody escape. PLoS 742 Comput. Biol. 18, e1010160 (2022). 30. Han, W. et al. Predicting the antigenic evolution of SARS-COV-2 with deep learning. Nat. Commun. 743 744 14, 3478 (2023). 31. Wang, G. et al. Deep-learning-enabled protein-protein interaction analysis for prediction of SARS-745

- 746 CoV-2 infectivity and variant evolution. *Nat. Med.* **29**, 2007–2018 (2023).
- 747 32. Taft, J. M. et al. Deep mutational learning predicts ACE2 binding and antibody escape to
- combinatorial mutations in the SARS-CoV-2 receptor-binding domain. *Cell* 185, 4008–4022.e14
  (2022).
- 33. Engler, C., Kandzia, R. & Marillonnet, S. A one pot, one step, precision cloning method with high
  throughput capability. *PLoS One* 3, e3647 (2008).
- 752 34. Pryor, J. M., Potapov, V., Bilotti, K., Pokhrel, N. & Lohman, G. J. S. Rapid 40 kb Genome
  753 Construction from 52 Parts through Data-optimized Assembly Design. *ACS Synth. Biol.* 11, 2036–
  754 2042 (2022).
- Taylor, G. M., Mordaka, P. M. & Heap, J. T. Start-Stop Assembly: a functionally scarless DNA
   assembly system optimized for metabolic engineering. *Nucleic Acids Res.* 47, e17 (2019).
- 36. Engler, C., Gruetzner, R., Kandzia, R. & Marillonnet, S. Golden gate shuffling: a one-pot DNA
  shuffling method based on type IIs restriction enzymes. *PLoS One* 4, e5553 (2009).
- 37. Boder, E. T. & Wittrup, K. D. Yeast surface display for screening combinatorial polypeptide
  libraries. *Nat. Biotechnol.* 15, 553–557 (1997).
- 38. Tzou, P. L., Tao, K., Kosakovsky Pond, S. L. & Shafer, R. W. Coronavirus Resistance Database
  (CoV-RDB): SARS-CoV-2 susceptibility to monoclonal antibodies, convalescent plasma, and plasma
  from vaccinated persons. *PLoS One* 17, e0261045 (2022).
- An updated atlas of antibody evasion by SARS-CoV-2 Omicron sub-variants including BQ.1.1 and
   XBB. *Cell Reports Medicine* 4, 100991 (2023).
- Chen, Y. *et al.* Broadly neutralizing antibodies to SARS-CoV-2 and other human coronaviruses. *Nat. Rev. Immunol.* 23, 189–199 (2022).
- 41. Wang, L. *et al.* Ultrapotent antibodies against diverse and highly transmissible SARS-CoV-2
  variants. *Science* 373, (2021).
- 42. Zost, S. J. *et al.* Potently neutralizing and protective human antibodies against SARS-CoV-2. *Nature*584, 443–449 (2020).
- 43. Ju, B. *et al.* Human neutralizing antibodies elicited by SARS-CoV-2 infection. *Nature* 584, 115–119
  (2020).
- 44. Zhou, B. *et al.* A broadly neutralizing antibody protects Syrian hamsters against SARS-CoV-2
  Omicron challenge. *Nat. Commun.* 13, 1–14 (2022).
- 45. Liu, L. *et al.* Potent neutralizing antibodies against multiple epitopes on SARS-CoV-2 spike. *Nature*584, 450–456 (2020).
- 46. Tortorici, M. A. *et al.* Broad sarbecovirus neutralization by a human monoclonal antibody. *Nature*597, 103–108 (2021).
- 47. Rappazzo, C. G. *et al.* Broad and potent activity against SARS-like viruses by an engineered human
  monoclonal antibody. *Science* 371, 823–829 (2021).
- 48. Moulana, A. *et al.* Compensatory epistasis maintains ACE2 affinity in SARS-CoV-2 Omicron BA.1. *Nat. Commun.* 13, 7011 (2022).

- 49. Cao, Y. *et al.* BA.2.12.1, BA.4 and BA.5 escape antibodies elicited by Omicron infection. *Nature*608, 593–602 (2022).
- 50. Wang, Q. *et al.* Antibody evasion by SARS-CoV-2 Omicron subvariants BA.2.12.1, BA.4 and BA.5. *Nature* 608, 603–608 (2022).
- 51. Cox, M. *et al.* SARS-CoV-2 variant evasion of monoclonal antibodies based on in vitro studies. *Nat. Rev. Microbiol.* 21, 112–124 (2022).
- 52. Bileschi, M. L. *et al.* Using deep learning to annotate the protein universe. *Nat. Biotechnol.* 40, 932–
  937 (2022).
- An updated atlas of antibody evasion by SARS-CoV-2 Omicron sub-variants including BQ.1.1 and
   XBB. *Cell Reports Medicine* 4, 100991 (2023).
- 54. Sheward, D. J. *et al.* Evasion of neutralising antibodies by omicron sublineage BA.2.75. *Lancet Infect. Dis.* 22, 1421–1422 (2022).
- 55. Wang, Q. *et al.* Alarming antibody evasion properties of rising SARS-CoV-2 BQ and XBB
  subvariants. *Cell* 186, 279–286.e8 (2023).
- 56. Wang, Q. *et al.* Antigenic characterization of the SARS-CoV-2 Omicron subvariant BA.2.75. *Cell Host Microbe* 30, 1512–1517.e4 (2022).
- 57. Dufloo, J. *et al.* Broadly neutralizing anti-HIV-1 antibodies tether viral particles at the surface of
  infected cells. *Nat. Commun.* 13, 1–11 (2022).
- 58. Meijers, M., Vanshylla, K., Gruell, H., Klein, F. & Lässig, M. Predicting in vivo escape dynamics of
  HIV-1 from a broadly neutralizing antibody. *Proc. Natl. Acad. Sci. U. S. A.* 118, (2021).
- 59. Doud, M. B., Lee, J. M. & Bloom, J. D. How single mutations affect viral escape from broad and
  narrow antibodies to H1 influenza hemagglutinin. *Nat. Commun.* 9, 1–12 (2018).
- 60. Underwood, A. P. *et al.* Durability and breadth of neutralisation following multiple antigen exposures
  to SARS-CoV-2 infection and/or COVID-19 vaccination. *EBioMedicine* **89**, 104475 (2023).
- 61. Chen, Y. *et al.* Immune recall improves antibody durability and breadth to SARS-CoV-2 variants. *Sci Immunol* 7, eabp8328 (2022).
- 62. Hastie, K. M. *et al.* Defining variant-resistant epitopes targeted by SARS-CoV-2 antibodies: A global
  consortium study. *Science* 374, 472–478 (2021).
- 812 63. Planas, D. *et al.* Reduced sensitivity of SARS-CoV-2 variant Delta to antibody neutralization. *Nature*813 **596**, 276–280 (2021).
- Raybould, M. I. J., Kovaltsuk, A., Marks, C. & Deane, C. M. CoV-AbDab: the coronavirus antibody
  database. *Bioinformatics* 37, 734–735 (2021).
- 816 65. Yue, C. *et al.* ACE2 binding and antibody evasion in enhanced transmissibility of XBB.1.5. *Lancet*817 *Infect. Dis.* 23, 278–280 (2023).
- 66. Ma, W., Fu, H., Jian, F., Cao, Y. & Li, M. Immune evasion and ACE2 binding affinity contribute to
  SARS-CoV-2 evolution. *Nat Ecol Evol* 7, 1457–1466 (2023).
- 67. Carabelli, A. M. *et al.* SARS-CoV-2 variant biology: immune escape, transmission and fitness. *Nat. Rev. Microbiol.* 21, 162–177 (2023).

- 822 68. A pseudovirus system enables deep mutational scanning of the full SARS-CoV-2 spike. Cell 186,
- 823 1263–1278.e20 (2023).
- 69. Benatuil, L., Perez, J. M., Belk, J. & Hsieh, C.-M. An improved yeast transformation method for the generation of very large human antibody libraries. *Protein Eng. Des. Sel.* **23**, 155–159 (2010).
- 826 70. BBMap. *SourceForge* https://sourceforge.net/projects/bbmap/ (2022).
- 827 71. Van Rossum, G. & Drake, F. L., Jr. *The Python Language Reference Manual*. (Network Theory.,
  828 2011).
- 829 72. Thomsen, M. C. F. & Nielsen, M. Seq2Logo: a method for construction and visualization of amino
- acid binding motifs and sequence profiles including sequence weighting, pseudo counts and two-
- sided representation of amino acid enrichment and depletion. *Nucleic Acids Res.* 40, W281–7 (2012).