Modeling mutational effects on biochemical phenotypes using convolutional neural networks: application to SARS-CoV-2

Bo Wang[1] and Eric R. Gamazon[1, 2, 3, 4]

[1]Division of Genetic Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA

[2]Vanderbit Genetics Institute, Vanderbilt University Medical Center, Nashville, TN, USA

[3]Data Science Institute, Vanderbilt University Medical Center, Nashville, TN

[4]Clare Hall, University of Cambridge, Cambridge, United Kingdom

Send correspondence to:

Eric R. Gamazon <ericgamazon@gmail.com>

ABSTRACT

Biochemical phenotypes are major indexes for protein structure and function characterization. They are determined, at least in part, by the intrinsic physicochemical properties of amino acids and may be reflected in the protein three-dimensional structure. Modeling mutational effects on biochemical phenotypes is a critical step for understanding protein function and disease mechanism as well as enabling drug discovery. Deep Mutational Scanning (DMS) experiments have been performed on SARS-CoV-2's spike receptor binding domain and the human ACE2 zinc-binding peptidase domain – both central players in viral infection and evolution and antibody evasion - quantifying how mutations impact binding affinity and protein expression. Here, we modeled biochemical phenotypes from massively parallel assays, using convolutional neural networks trained on protein sequence mutations in the virus and human host. We found that neural networks are significantly predictive of binding affinity, protein expression, and antibody escape, learning complex interactions and higher-order features that are difficult to capture with conventional methods from structural biology. Integrating the intrinsic physicochemical properties of amino acids, including hydrophobicity, solvent-accessible surface area, and long-range non-bonded energy per atom, significantly improved prediction (empirical $p<0.01$) though there was such a strong dependence on the sequence data alone to yield reasonably good prediction. We observed concordance of the DMS data and our neural network predictions with an independent study on intermolecular interactions from molecular dynamics (multiple 500 ns or 1 µs all-atom) simulations of the spike protein-ACE2 interface, with critical implications for the use of deep learning to dissect molecular mechanisms. The mutation- or genetically- determined component of a biochemical phenotype estimated from the neural networks has improved causal inference properties relative to the original phenotype and can facilitate crucial insights into disease pathophysiology and therapeutic design.

Introduction

Since the initial outbreak, the SARS-CoV-2 virus has rapidly spread worldwide causing a global public health crisis, the coronavirus disease 2019 (COVID-19). *In vitro* and cryo-electron microscopy studies have established that the betacoronavirus uses the human cell-surface protein angiotensin converting enzyme 2 (ACE2) to gain entry into target cells[1–3]. Therefore, precise characterization of the interaction between the Receptor Binding Domain (RBD) of the viral spike glycoprotein and the ACE2 complex is of critical importance in understanding COVID-19 pathophysiology[3]. Not surprisingly, several drug candidates that target either the virus or the receptor have been developed on the basis of the ACE2 binding. With improved understanding of this key molecular interaction, two major therapeutic strategies have been pursued, including 1) engineering high-affinity ACE2 decoy or developing antibody cocktail treatments and 2) screening new or repurposing existing inhibitors targeting the binding interface[4,5]. Establishing the sequence-structure-phenotype relationship for the spike RBD and the ACE2 receptor is essential for both strategies, in which the sequence mutational effect on receptor affinity and other biochemical phenotypes is the major component[6–10].

Comprehensive understanding of how variants, including single mutations, affect disease-relevant biochemical phenotypes would go a long way towards clarifying molecular mechanisms of disease as well as downstream adverse complications and guiding pharmacological interventions. In addition, elucidating the mutational effect may shed light on selective pressures determining the evolutionary trajectory of the coronavirus as well as identify risk factors for viral infection and maladaptive host response to COVID-19 in human populations[11]. Deep Mutational Scanning (DMS) systematically evaluates the effect of mutant versions of the protein on measured biochemical phenotypes [12,13,6,14]. High-throughput mutagenesis in DMS makes it possible to assess the phenotypic consequences of each possible amino acid mutation in a protein, generating large datasets that can

reveal the sequence-function landscape. The development of computational approaches to learn the complex and non-linear features of this map can enable high-throughput inference of basic protein properties. Statistical and machine learning methods, including deep learning, have attracted significant attention due to their predictive power[15]. A recently developed supervised learning framework tailored to DMS datasets, convolutional neural networks demonstrated spectacular performance, consistent with other recent studies of mutational effect[16,17].

DMS experiments on both the SARS-CoV-2 spike glycoprotein and the ACE2 receptor have been performed, providing an important basis for further investigations of mutational effects [4,7,18]. In this work, we conducted systematic modeling of the mutational effects of the RBD in the viral spike protein and of the ACE2 receptor on biochemical phenotypes, extending a supervised learning framework[16]. Three classes of critical phenotypes -- binding affinity, protein expression, and antibody escape -- were systematically analyzed within the sequence-structure-function paradigm that informs much of proteomic and structural biology studies. Neural networks were also leveraged to learn (a) the complex functional landscape of the viral spike protein's RBD and the host cell-surface receptor ACE2 and (b) the antibody-escape map of RBD mutations, including mutations that undergo selection during viral proliferation in the presence of antibodies and mutations that have been circulating in human populations[19–21]. Finally, the mutation- or genetically- determined phenotypes (i.e., the component of a phenotype defined by the sequence versus a technical confounder or the environmental component) estimated from the neural networks were exploited to extract key insights into the molecular mechanisms of COVID-19 infection and severity.

Results

Overview of framework

Briefly, we describe the framework (Figure 1A). Here we consider a set of N training samples consisting of (protein) sequences, $\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_\mathrm{N} \in \mathrm{X}$, with corresponding biochemical phenotype values, $\mathrm{y}_1, \mathrm{y}_2, \ldots, \mathrm{y}_\mathrm{N} \in \mathrm{Y}$ (where phenotype is fixed and chosen from binding affinity, protein expression, or an antibody escape measure), where the pairs are sampled independently and identically distributed (i.i.d.) from a joint distribution

$$(\mathrm{x_i}\ , \mathrm{y_i}) \sim \Lambda$$

A supervised learning method seeks to find, from a pre-defined class of functions, a function $\mathrm{g} : \mathrm{X} \rightarrow \mathrm{Y}$, which maps the input sequence x to the phenotype value y and minimizes the expected deviation

$$\mathrm{E}(\mathrm{L}(\mathrm{y}, \mathrm{g}(\mathrm{x}))) = \int \mathrm{L}(\mathrm{y}, \mathrm{g}(\mathrm{x}))\Lambda(\mathrm{x}, \mathrm{y})\mathrm{dxdy}$$

from y under a loss function $\mathrm{L} : \mathrm{Y}\ \mathrm{x}\ \mathrm{Y} \rightarrow \mathrm{R}$. Regularization (e.g., dropout) is also performed to prevent overfitting and enhance generalization error. The expected loss is generally approximated by the average of the $\mathrm{L}(\mathrm{y}_i, \mathrm{g}(\mathrm{x}_i))$ across the training samples, i.e., the empirical loss, given that the distribution Λ is typically not known. Neural networks constitute the class of functions defined by the composition of affine maps (Ψ_i) and activation functions (φ) (which are typically non-linear).

$$\mathrm{g}(\mathrm{x}) = \Psi_\mathrm{M}(\varphi(\Psi_{\mathrm{M-1}}\varphi(\ldots\ \varphi(\Psi_1(\mathrm{x})\ \ldots) \qquad (1)$$

(Note linear regression can be trivially included as a member of this class of functions.) The minimization problem, i.e., the search for an optimal function g, is framed as a search for a set of weights $\mathbf{W}_{(k)}$ and biases $b_{(k)}$ (Methods).

Application of deep learning framework to the spike RBD and the ACE2 PD complex

The crystal structure of the spike RBD — ACE2 PD complex was retrieved from the Protein Data Bank (PDB; PDB ID: 6M0J)[2]. Our study investigated two biochemical phenotypes, binding affinity and protein expression – which are critical to viral evolution, receptor recognition, and host-virus interaction

– and a class of antibody-escape phenotypes, which is critical to understanding the antigenic consequences of mutations and facilitating effective vaccine design (Figure 1A). Specifically, we trained on single-mutation functional score profiles, consisting of all possible amino acid mutations and the corresponding phenotypic measurements (quantified relative binding score for the ACE2 catalytic zinc-binding peptidase domain (PD), binding affinity and expression for the spike RBD, and the "escape fraction" from each of 10 human monoclonal antibodies) measured by DMS experiments. The protein sequences x_i were encoded by one-hot encoding and an external curated featurization table, i.e., AAindex, which summarizes the intrinsic physicochemical properties (e.g., hydrophobicity, solvent-accessible surface area, and long-range non-bonded energy per atom) of amino acids (Figure 1B; Methods).

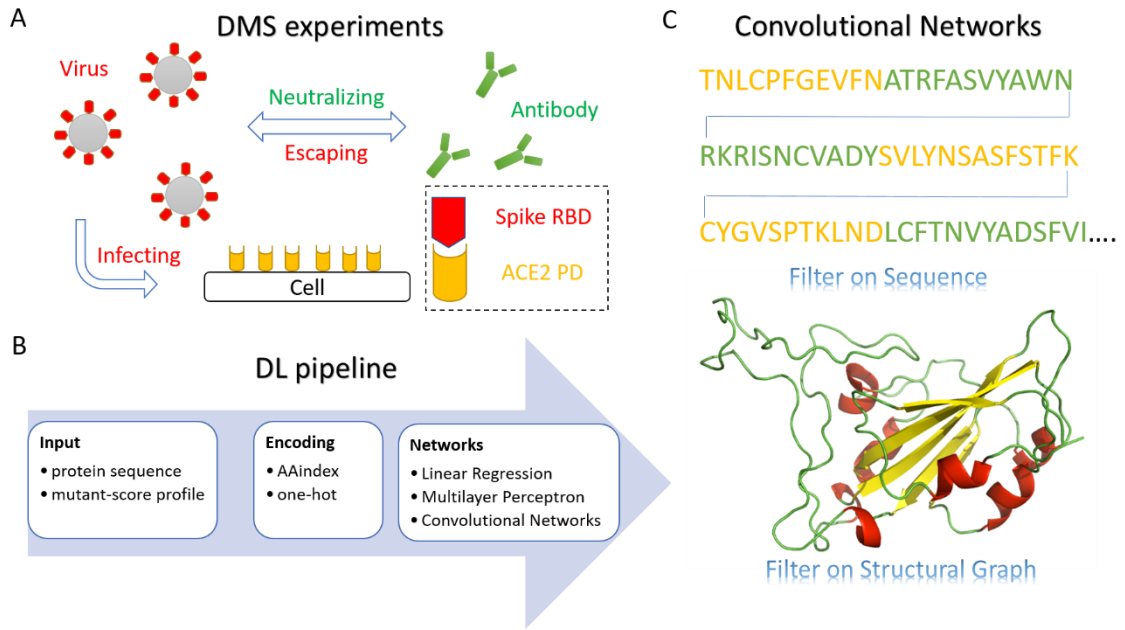


Figure 1. Overview of the framework. A We evaluated three biochemical phenotypes generated from Deep Mutational Scanning (DMS) experiments: binding affinity, protein expression, and antibody escape. These biochemical phenotypes may determine the evolutionary trajectory of the coronavirus, viral infection, and maladaptive host response. **B** We quantified the extent to which state-of-the-art neural network methodologies can predict these phenotypes, using only sequence mutation data. The predicted biomedical phenotypes are mutation- or genetically- determined traits, which are therefore of intrinsic interest because they are determined solely by the (amino-acid) sequence data. We encoded the protein sequence input using AAindex, consisting of 566 intrinsic (data-independent) physicochemical properties of the amino acids, such as hydrophobicity and long-range non-bonded

energy per atom, and one-hot encoding. We then trained neural networks on the mutation-phenotype data from the DMS experiments. **C** We trained sequence convolutional neural networks and graph convolutional neural networks (along with [fully connected] multilayer perceptrons and linear regression) on the DMS data. Shown are the convoluted sequences and structural motifs for the two types of architecture, respectively, differentiated by color.

We generated independent models, which allow prediction of mutational effects on binding affinity, expression, or antibody escape (for each of the 10 antibodies) of previously uncharacterized variants. We implemented linear regression and several neural network architectures: multilayer perceptron (fully connected), sequence convolutional neural network, and graph convolutional neural network. While linear regression considers only the additive contributions of sequence residues, multilayer perceptron can capture nonlinear relationships between sequence (input) and phenotype (output). Due to their distinctive architecture, convolutional neural networks are even more powerful to learn higher-level features. Filters for sequence convolutional neural networks slide along the one-dimensional input of protein sequence and learn from sequence fragments while graph convolutional neural networks act on the structural coordinates of proteins (Figure 1C). In the graph structure representation $\mathrm{G} = (\mathrm{V}, \mathrm{E})$ of the spatial structure of the protein, the sequence residues are represented as nodes V while the interactions for pairs of residues (determined by a distance threshold) are represented as edges E. Graph convolutional neural networks learn segments of protein structures, i.e., motifs, with the convolutional kernels capturing translational invariance.

Evaluating models learned from Deep Mutational Scanning data

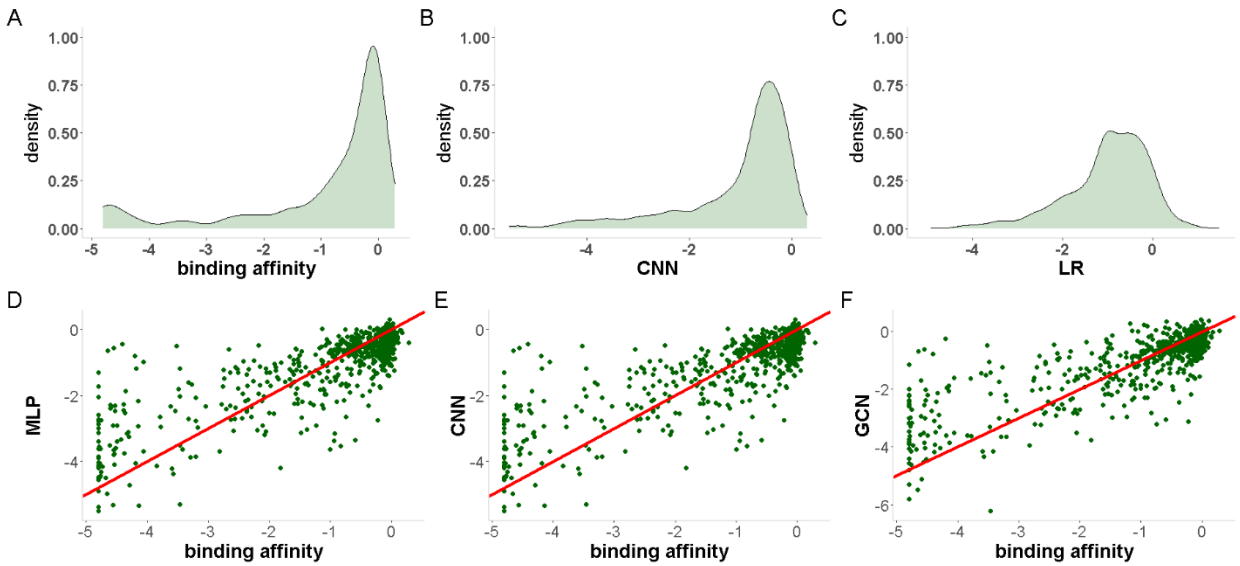


Figure 2. Predictive performance of neural networks on binding affinity. A Distribution of ACE2 binding affinity in the test set from single mutations in the spike glycoprotein RBD as measured by DMS. The predicted biochemical phenotype from **B** neural networks, e.g., CNN, was much closer in distribution to the original phenotype than that from **C** linear regression. Notably, **D-F** the 3 neural network architecture models trained here showed significant (Spearman) correlation between the observed (x-axis) and predicted (y-axis) phenotype in an independent test dataset. Red line is the diagonal ($y = x$) line. CNN: convolutional neural network (Table S1, Network 2); LR: linear regression; MLP: multilayer perceptron; GCN: graph convolutional network (Table S1, Network 8).

We performed the supervised learning methods on the various spike RBD and ACE2 PD datasets, comprising mutant-score profiles of binding affinity, expression, and antibody escape. For each model (defined by a phenotype from a DMS dataset), the dataset was randomly split into 60% training, 20% tuning, and 20% testing subsets. Trained models were evaluated by predictive performance on the test subset (e.g., Figure 2A). The supervised learning methods demonstrated varied predictive performance in the case of the spike RBD, where the Spearman correlation coefficient between the true and predicted phenotype ranged from 0.49 (for linear regression) to 0.78 (for convolutional neural network) for ACE2 binding affinity (Table S1). Neural networks provided much better prediction performance than linear regression, indicating that binding affinity is characterized by interactions (non-additive effects) among residues and non-linear mappings of sequence to function. Relative to the distribution of binding

affinity (Figure 2A) from the RBD mutations, the predicted phenotypes from the neural networks had greater fidelity to the original than linear regression (Figure 2B and 2C). Notably, the amino acid sequence mutational data alone could significantly predict binding affinity ($p < 2.2x10^{-16}$), using multilayer perceptron (Spearman correlation $\rho = 0.52$), sequence convolutional neural network ($\rho = 0.69$), and graph convolutional neural network ($\rho = 0.70$) (Figure 2D, 2E, and 2F, Table S1).

For the spike glycoprotein expression, the predictive performance (Spearman correlation coefficient) ranged from 0.44 (for linear regression) to 0.84 (for graph convolutional network). Binding affinity to ACE2 and spike glycoprotein expression were significantly correlated ($\rho = 0.65$) in the training set, but the relationship was found to be highly nonlinear (Figure 3A). Though both phenotypes reflect protein stability under a wide range of biochemical contexts, binding affinity is more determined by the surface residues on or near the binding sites as well as by allosteric effects induced by inner or remote residues. As in the case of binding affinity, the neural network models significantly predicted RBD expression ($p < 2.2x10^{-16}$), using multilayer perceptron (Spearman correlation $\rho = 0.50$), sequence convolutional neural network ($\rho = 0.682$), and graph convolutional neural network ($\rho = 0.679$) (Figure 3B, 3C, and 3D, Table S1). Despite the range in prediction performance of the models, the estimates of the mutation-determined component of RBD expression derived from them were highly correlated (Figure 3E).

Previous studies have shown that mechanistic insight of protein mutational effects is not easily transferable to studying the effects of variants in other proteins[14]. We hypothesized that models trained on one species-phenotype are not suitable for predicting another species-phenotype. To evaluate this hypothesis, we trained a model using the avGFP DMS dataset[22], in which the score of the green fluorescent protein from *Aequorea victoria*, avGFP, is fluorescence. The avGFP sequence was slightly modified to match the input dimension of the spike RBD (Methods). Therefore, the expression-based score profile of the spike RBD could be used as an external test dataset, with its combination of amino-

acid sequence unseen during the training and an unrelated phenotype. As hypothesized, the sequence convolutional model from avGFP failed to predict the expression of mutated sequences (Spearman correlation $\rho = -0.0346$). This finding implies that the biochemical phenotypes differed in their dependence on protein amino-acid sequence length, type, or composition, with interaction from non-sequence physicochemical environments (for example, pH condition and ionic density) potentially accounting for the difference.

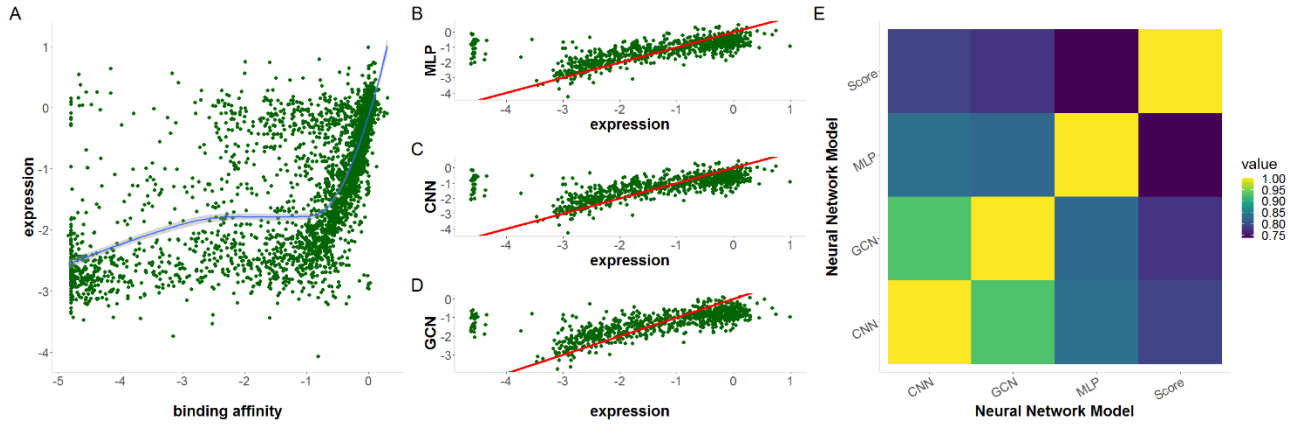


Figure 3. Predictive performance of neural networks on protein expression. A Binding affinity to ACE2 and spike glycoprotein expression, both induced by a protein sequence mutation on the RBD, were significantly correlated (Spearman correlation $\rho = 0.65$), but the relationship between the two phenotypes was highly nonlinear. We fit a curve through the points, using Locally Estimated Scatterplot Smoothing (LOESS), a method for local polynomial regression. **B-D** We trained three neural network architecture models, which could significantly predict (Spearman correlation $\mathrm{p} < 2.2\mathrm{x}10^{-16}$) protein expression in an independent test set. Red line is the diagonal ($y = x$) line. **E** Pairwise correlation between the neural network models. MLP: multilayer perceptron; CNN: convolutional neural network (Table S1, Network 2); GCN: graph convolutional network (Table S1, Network 8).

Influence of network architectures on biochemical phenotype prediction

All mutational sequences were encoded by the externally defined AAindex, a set of 566 quantitative indices of physicochemical properties of amino acids, which were used to train the

empirical mutational scores on the DMS data[23]. Thus, we examined the impact of the quality of the AAindex on prediction performance. We shuffled the entire AAindex so that the intrinsic information for each amino acid would be completely lost and replaced with random noise (Methods). We generated 100 pseudo AAindex tables, on which sequence neural networks were trained, using a given choice of configurations, i.e., number of layers and filter dimensions (Figure 4A). For each choice of network configurations, the distribution followed a bell-shaped curve, with the prediction performance from the actual AAindex located at the far-right and outperforming any of the shuffled tables. In other words, integrating the AAindex information significantly improved (empirical $p<0.01$) the overall performance of the neural networks.

The prediction substantially improved with a larger filter dimension on a single hidden layer sequence convolutional network. This observation implies that the neural networks were able to learn more information from longer sequence fragments, which is consistent with the expression of the spike RBD reflecting complex non-linear relationships of the sequence (Network 1, 2 Figure 4A). Even with a shuffled AAindex table, i.e., in which the intrinsic properties of the amino acids were essentially lost, a larger filter dimension led to enhanced learning from the expression phenotype, with increased mean and reduced standard deviation for the empirical performance distribution, indicating highly strong dependence of expression on the sequence data. Deeper neural networks with a fixed filter dimension also showed improved prediction on expression (Network 2-4, Figure 4A). Integration of accurate AAindex into the input for a deeper neural network resulted in further performance gain relative to shuffled AAindex, confirming that the neural network was leveraging the externally-defined intrinsic properties. To further reveal the information learned from neural networks by increasing the number of hidden layers, we applied Uniform Manifold Approximation and Projection (UMAP) to visualize the latent space in the last layer[16,24]. Notably, a deeper network could better differentiate high- and low-scoring sequence mutants as distinct clusters (Network 1-4, Figure 4B).

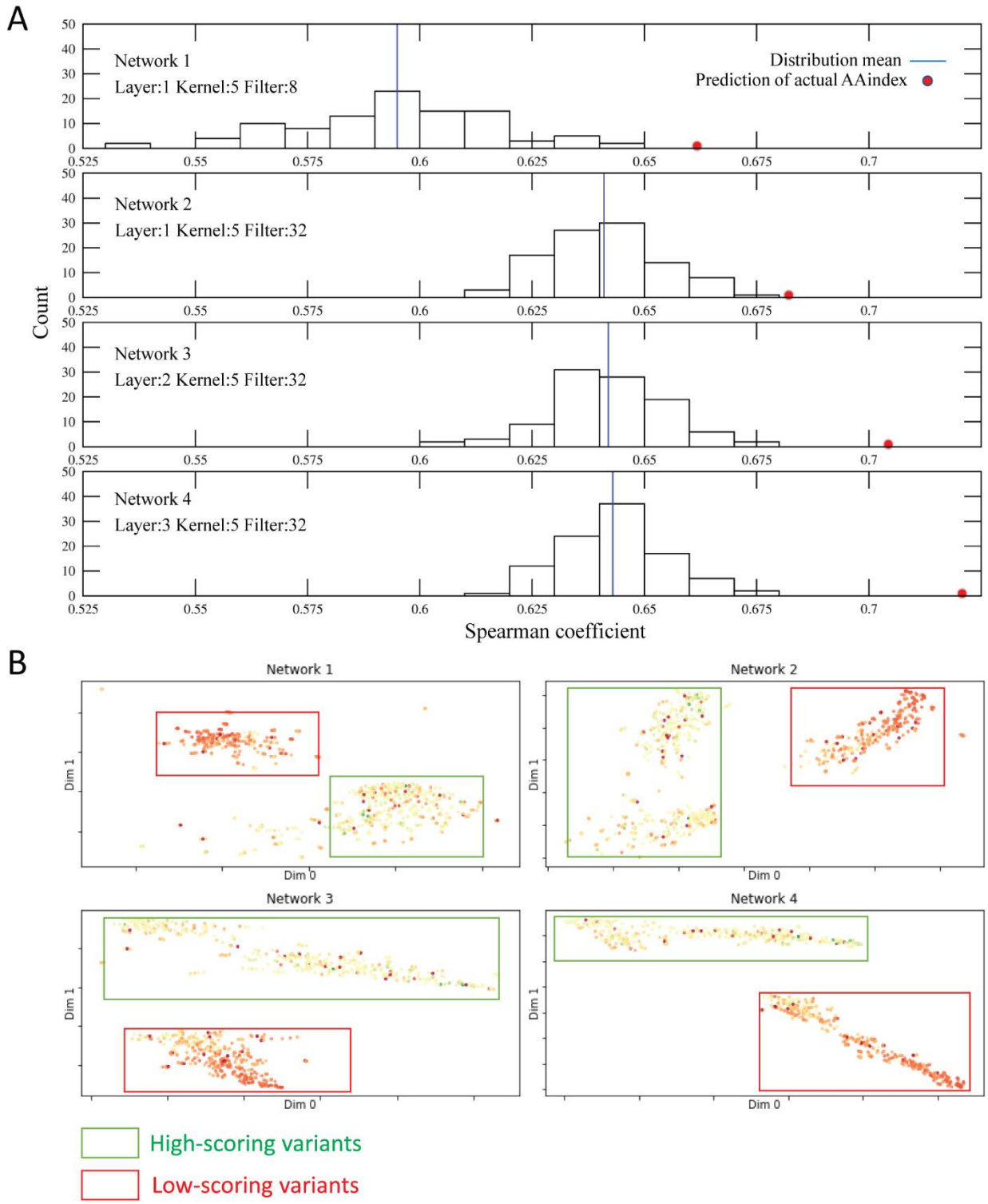


Figure 4. Effect of neural network architectures on predicting biochemical phenotypes. A Distribution of predictive performance (Spearman correlation coefficient) of sequence convolutional neural networks with shuffled AAindex (shown as a histogram) and the actual AAindex (shown as a red dot), training on the spike RBD expression dataset. Each plot corresponds to a specific choice of architecture,

including the number of layers and the filter dimension. Increasing the filter dimension or the number of layers improved the prediction performance. Furthermore, encoding the input data with the actual AAIndex, which encodes the externally-defined intrinsic physicochemical properties of the amino acids, consistently led to performance gain. Nevertheless, we observed a strong dependence, on the input sequence data alone, of the original biochemical phenotype, with reasonable performance across all choices of network configurations. **B** Uniform Manifold Approximation and Projection (UMAP) projection of the latent space of the sequence convolutional networks on the last internal layer of the network. Mutants are colored according to the actual phenotype score, with green and red representing high-scoring and low-scoring variants, respectively.

Mutational effects on biochemical phenotypes

SARS-CoV-2 infection leads to complex pathophysiological consequences which are far from being fully understood. Elucidating mutational effects on biochemical phenotypes for the spike RBD and the ACE2 PD is critical in improving our understanding of the virus-host interactions as well as host immune response. Furthermore, with the spike glycoprotein as the primary target due to its role in mediating coronavirus entry into host cells, neutralizing Antibodies (Abs) may provide protection from SARS-CoV-2 in humans[25]. Here, we highlight the results from sequence convolutional neural networks that showed the optimal performance among all tested models and further interpret them in connection with existing experimental characterizations, molecular modeling, and host genetic inferences. In addition, we leveraged the neural network derived phenotypes to evaluate their utility in downstream analyses because, as inferred phenotypes, they are solely determined by the sequence mutation data and are thus less biased by possible confounding artifacts[26,27], including batch effects, environmental (non-sequence physicochemical) determinants, and other technical sources of heterogeneity.

Key mutations on spike RBD. SARS-CoV-2 evolution promotes virulence and enhances escape from neutralizing Abs. Interface mutations (Figure 5A and 5D), specifically those at contact sites, are more likely to have direct impact on infectivity. The mutational effects of these key binding residues of the spike RBD (to ACE2 PD) were systematically evaluated by our neural networks (Figure 5B and 5E).

Predicted phenotypic effects were consistent with DMS measurements[7]; for example, mutations at G502 showed the most extreme impact on receptor binding affinity (see figure 5B). In addition, the key spike RBD binding residues were consistent with results of interfacial interactions of the spike RBD bound to ACE2 from molecular dynamics simulations[28] (performed for 500 nanoseconds in triplicates or 1 μs all-atom, using Desmond). Intermolecular interactions, including hydrophobic interactions, hydrogen bonds, π- π, and cation-π, were dynamically observed in formation and breakage in the simulations. Notably, the key binding residues of the spike RBD formed consistent polar interactions with the corresponding residues of ACE2 during the molecular dynamics simulation trajectories (Figure 5C).

Neutralizing Abs secreted from immune cells may also be found to be in complex with the spike RBD. Consistent with the experimental characterizations[19], the convolutional neural network derived (i.e., estimated, mutation-determined) antibody-escape phenotype for the 10 human monoclonal Abs have differentiated target "hotspots" on the surface of the spike RBD (Figure 6A). The neural network prediction performance on antibody escape of the spike RBD varied between the Abs (Spearman correlation coefficient between 0.09 and 0.32) (Figure 6B). The variability in performance might be due, at least in part, to the unique physicochemical properties of the 10 monoclonal antibodies as well as their specific relative dependence on sequence mutation and other (non-sequence) determinants.

As receptor recognition and antibody escape are closely related viral activities, we implemented Random Forest to calculate the *feature importance* of the 10 convolutional neural network derived mutation-determined antibody-escape phenotypes as predictors of binding affinity towards the ACE2 receptor. We hypothesized that those Abs that shield the receptor by pre-occupying the spike RBD – ACE2 PD binding interface would weigh more in their contributions to the prediction. Notably, we found that the top 3 Abs (COV2-2499, COV2-2832, and COV2-2050), all entirely targeting the interface based on experimental characterizations[19], among the neural network derived antibody-escape predictors

accounted for nearly 60% of the feature importance score (Figure 6C). In particular, key residues on the spike RBD that are critical binding sites with ACE2 PD include G446, Y449, G496, Q498, and T500 for escape mutations of the Ab COV-2499. We found these residues to be the sites of the peaks for the neural network derived antibody-escape phenotype for COV-2499, i.e., the predicted escape mutations (Figure 6A). The importance of these residues for viral binding to the host receptor was also supported by results on intermolecular interactions from the molecular dynamics simulations (Figure 5C).

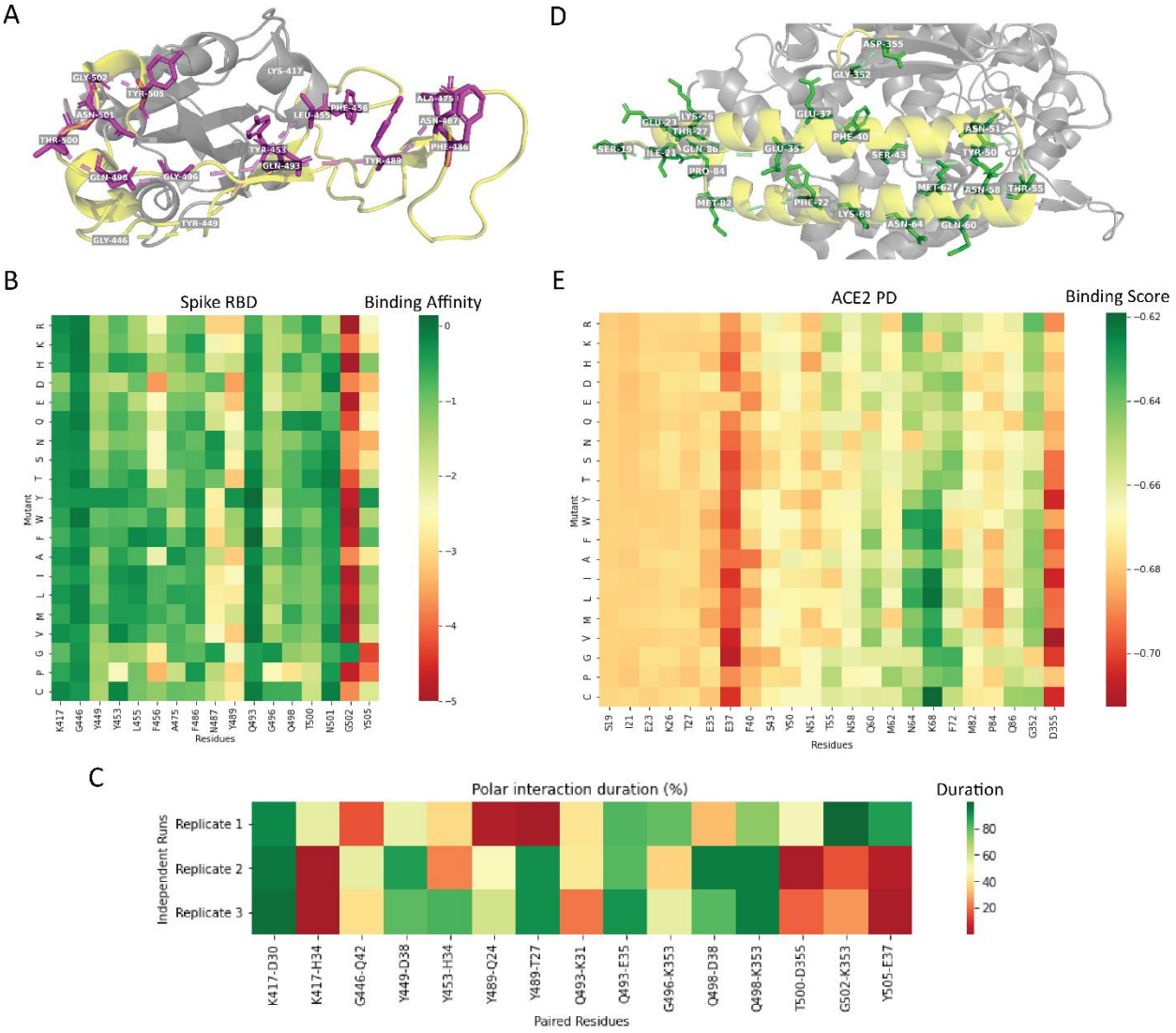


Figure 5. Mutational effects at key interface residues on the spike RBD and ACE2 PD. A Human ACE2-interacting residues (magenta) of the spike RBD interface (yellow) with **B** their predicted mutational effects on receptor affinity from a convolutional neural network with optimal performance (Network 3,

Table S1). **C** Paired residues between the spike RBD and the ACE2 PD that retained polar interactions and the interaction duration percentage from molecular dynamics simulation trajectories. Briefly, 500 nanosecond molecular dynamics simulations had been performed in triplicates using Desmond (ref 28), with the structure of the spike RBD bound to ACE2 solvated with single point charge water molecules and with use of the OPLS forcefield. Intermolecular interactions (e.g., hydrophobic interactions, hydrogen bonds, π- π, cation-π) were dynamically observed in formation and breakage. We leveraged the residues of the spike RBD forming consistent polar interactions with the corresponding residues of ACE2 to evaluate concordance with the DMS measurements and our neural network predictions. Each paired residues included here were in polar interaction for at least 40% of the simulation duration in at least one replicate. **D** Human polymorphisms (shown in green stick representation), based on the gnomAD resource (v2) from large-scale sequencing projects, of the ACE2 PD interface with **E** the corresponding residues' predicted mutational effects on binding affinity from a convolutional neural network with optimal performance (Network 5, Table S2).

Key mutations on ACE2 PD. We comprehensively evaluated mutational effects on the relative binding of ACE2 PD. In addition to key binding residues paired with those of the spike RBD (Figure 5C), we investigated mutations with high allele frequencies in human populations (Figure 5D and 5E). While linear regression (Spearman correlation $\rho = -0.014$) and the multilayer perceptron (Spearman correlation $\rho = -0.044$) failed to make accurate predictions, the best performance achieved from sequence convolutional neural networks was 0.37 (Table S2). The lower phenotypic variance explained by the neural network models on ACE2 PD binding implies a complex sequence-function relationship, with possible contributions from sequence mutations and other (environmental) determinants. Notably, the empirical distribution of the score of the spike RBD and the ACE2 PD (Figure S1) is consistent with the spike RBD displaying more non-conserved mutations than the ACE2 PD, which fits our intuitive understanding of the properties of the virus (which must undergo adaptation to transmission and replication in the host) and the membrane receptor (with its conserved function).

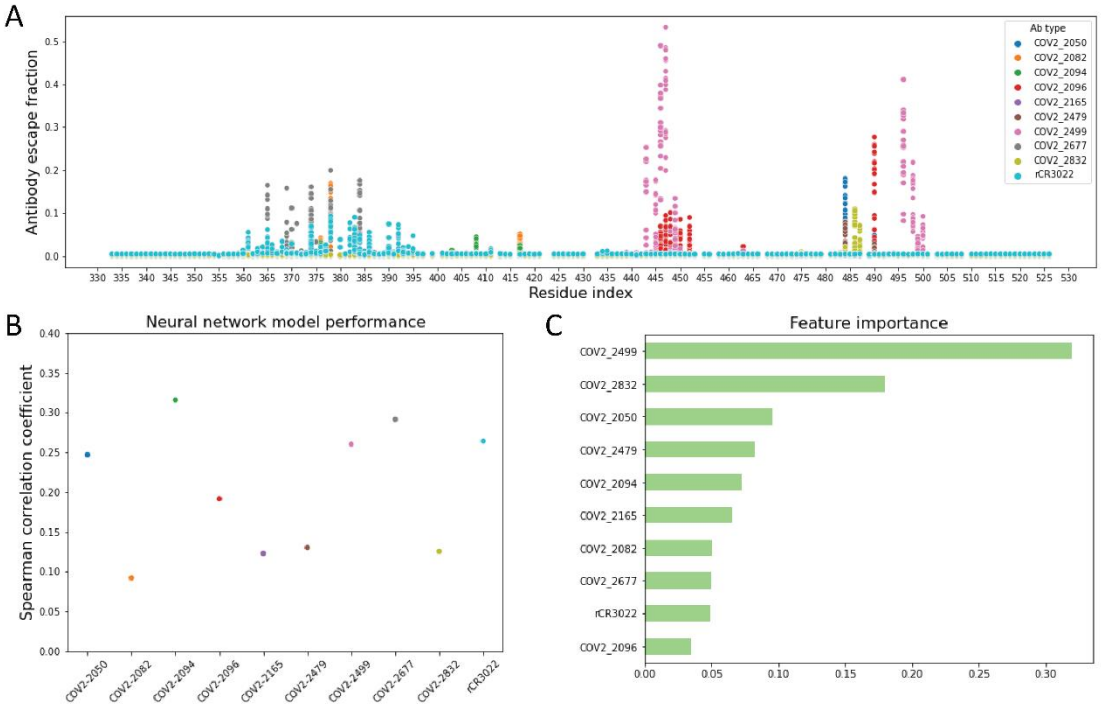


Figure 6. Neural network derived antibody-escape phenotypes, performance prediction, and feature importance. A Maps of the convolutional neural network derived (i.e., estimated, mutation-determined) escape fraction phenotype on the spike RBD for the different Abs (shown as different colors). Shifting peaks in different colors indicate differentiated binding sites for the Abs on the spike RBD surface. Key residues on the spike RBD that are critical binding sites with ACE2 PD include G446, Y449, G496, Q498, and T500 for escape mutations of COV-2499. Note these residues are the sites of the peaks of the neural network derived antibody-escape phenotype for COV-2499, i.e., the predicted escape mutations. The importance of these residues for viral binding to the host receptor is also supported by results on intermolecular interactions from the molecular dynamics simulations (Figure 5C). **B** Predictive performance (calculated as Spearman correlation coefficient between actual and predicted values) of our neural network model on each of the 10 Abs (Table S3). **C** Feature importance of the convolutional neural network derived (i.e., estimated, mutation-determined) antibody-escape phenotype for each of the 10 Abs as derived from a joint Random Forest regression model in predicting binding affinity towards the ACE2 receptor. The Abs that shield the receptor by entirely targeting the binding interface ranked in the top 3 neural network derived antibody-escape predictors.

Discussion

In this work, we leveraged neural networks to model the biochemical phenotypes of the spike RBD and the ACE2 PD. Mutations can directly influence protein conformational dynamics, stability, and associations/dissociations, i.e., the sequence-structure-phenotype relationship[28]. We examined the prediction performance of several neural network architectures designed to learn the features of the sequence-function landscape. Convolutional neural networks demonstrated superior performance over

linear regression or multilayer perceptron in “decoding” the mutational information from sequence to phenotype. Furthermore, the sequence convolutional neural network, in general, outperformed the graph convolutional neural network (Tables S1 and S2). The performance of the graph convolutional neural network might be related to the graph model architecture as well as the physicochemical nature of the predicted phenotype, i.e., the node average model does not sufficiently capture relevant features from a structural perspective for binding affinity. Capturing non-local motifs or long-range dependencies such as allosteric effects in the structural information warrants further study.

Our deep learning experiments confirmed the quality of the AAindex, which captures the intrinsic properties of each amino aside that are independent of the particular sequence, structure, or DMS effect scores. Our shuffling experiments showed that the AAindex significantly improved the prediction. Alternatively put, the use of a randomly generated AAindex was found to be sub-optimal, with the physicochemical properties encoded in the actual AAindex having a significantly greater impact on phenotype prediction. Nevertheless, we observed strong dependence of the biochemical phenotypes on the input sequence data alone, from which the convolutional neural networks could extract features and perform phenotype prediction reasonably well.

Investigating mutational effects is essential for deriving mechanistic insights and elucidating disease pathways. Molecular modeling tools such as molecular dynamics simulations and docking are frequently used towards these ends, with the clear benefit of obtaining atomic-detail understandings of biomolecular systems[9,28]. However, these tools are typically impeded by system complexity and computational demand, so that only a partial structure could be simulated for a limited number of mutational systems. As structural interactions comprise the central elements of molecular modeling, residues on the binding interface with high binding frequency or affinity are likely to be the focus, while dissociation effects like antibody escape are challenging to model. In fact, it has been reported that the rapid spread of COVID-19 has more to do with asymptomatic and pre-symptomatic transmission than

enhanced receptor binding[6]. In addition, allosteric effects are difficult to capture in molecular modeling yet critical to viral infection and evolution. It has been reported that the D614G variant that is located far from the RBD of the spike glycoprotein can facilitate ACE2 infection.[29] Similarly, mutations near the ACE2's chloride-binding sites, which are located far from the spike interface, may also alter the spike RBD binding[18]. Some mutations could become prevalent as selection facilitates receptor binding. For instance, the N501Y mutation, which sits between two key binding residues, i.e., T500 and G502 (Figure 5A, 5B and 5C), has been reported to be responsible for the rapid spread of the virus in southeastern England, threatening the efficacies of the existing vaccines[30,31]. Deep learning on DMS, as presented here, is a time-efficient and economical way to model high-dimensional mutational data to gain insights into therapeutic design (Figure 6).

It is worth noting that the tested neural networks achieved much better prediction performance on the spike RBD than on the ACE2 PD. The difference in prediction performance on the two interacting components of the protein complex could result from difference in the data quality, i.e., the DMS experiments were performed in two separate studies. However, the difference may also be due to a difference in the intrinsic properties of the virus and the receptor. Specifically, the human ACE2 phenotype may be less dependent on (or determined by) sequence mutation data. This finding was also supported by the fact that linear regression or fully connected networks failed to make good prediction for the ACE2 PD.

Methods

Datasets

We performed our deep learning experiments on publicly available spike RBD and ACE2 PD DMS datasets[7,18,19]. Binding affinity and protein expression of the spike RBD as well as the relative binding of ACE2 PD dataset are considered. We also leveraged a deep mutational scan of the RBD to determine

how variants from spike RBD responded to antibody binding from 10 human monoclonal antibodies, of which nine target SARS-CoV-2 and one targets SARS-CoV. The original mutational effect scores - "escape fraction" - have been normalized via Box Cox transformation.

Models on the avGFP protein (an independent DMS dataset) were also retrained and tested in the RBD expression dataset[16,22]. For this purpose, the wild-type sequence for avGFP was reduced so that the overall sequence length matched that of the spike RBD. Accordingly, the original DMS dataset for avGFP was modified to only include the variant-score profile of the sliced sequence.

Protein sequence encoding and structural description

AAindex is a resource of 566 physicochemical properties (e.g., polarizability parameter, residue volume, solvation free energy, and other attributes) for each of the 20 amino acids[32]. The AAindex applied in this work underwent dimensionality reduction (from 566 properties to 19 Principal Components)[16]. The wild-type and mutated protein sequence were encoded using one-hot encoding and the AAindex.

Network architectures and model training

Here, we extend a supervised learning framework for application to interrogating the biochemical phenotypes relevant to the virulence of SARS-CoV-2[16]. We utilized different neural network architectures including linear regression, multilayer perceptron (fully connected), sequence convolutional neural network, and graph convolutional neural network. These architectures can be viewed as directed computational graphs organized as nodes in a series of hidden layers $h_{(k)}$

$$h_{(k)} = \varphi(\mathbf{W}_{(k)} h_{(k-1)} + b_{(k)})$$

with a matrix of weights $\mathbf{W}_{(k)}$ and biases $b_{(k)}$. The network architectures are implemented as compositions of functions (equation 1), reflecting the universal approximation properties of the architectures with respect to the compact convergence (compact-open) topology[33].

Convolutional neural networks perform convolution, activation, pooling, and flattening to learn features from input data. Briefly, convolution is a mathematical operation involving a convolution operator m and input s:

$$(m * s)(t) = \int_{-\infty}^{\infty} m(x).s(t-x)dx$$

This one-dimensional definition can be extended to arbitrary dimensions. This formulation has a discrete version:

$$(m * s)(t) = \sum_{x=1}^{N} m[x]\,.\,s[t+x-1]$$

The convolution operation can be viewed as a weight assignment scheme. Different operators or *filters* can be applied to the input data to derive an optimally predictive set of features. Activation introduces non-linearity and pooling helps to reduce overfitting. A fully-connected neural network is then applied to the resulting feature vector (after flattening) to obtain the final prediction.

Minimization of the empirical loss across the training samples

$$\hat{g} = \underset{g}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} \mathrm{L}(\mathrm{y}_i, \mathrm{g}(\mathrm{x}_i))$$

yields an estimate of the mutation- or genetically- determined $\hat{g}(x)$, whose properties can be explored (and compared with the original g) and which may enable causal inference on a biochemical phenotype. We trained the models using mean squared error loss and the Adam optimizer with dropout as the

regularization method (in particular, dropout rate of 20%). For activation function φ (equation 1), a leaky ReLU was applied, which permits a small, positive gradient for a non-active unit

$$\varphi(x) = \begin{cases} x & \text{if } x>0 \\ 0.01x & \text{otherwise} \end{cases}$$

Neural networks were implemented using TensorFlow. We set the maximum possible number of epochs to 400 with early stopping.

The uncertainty associated with the neural network model can be quantified. Let ω be the set of all model parameters, i.e., the weights $\mathbf{W}_{(k)}$ and biases $b_{(k)}$. Then the posterior probability $P(\omega \mid X_{\text{train}}, Y_{\text{train}})$ of the set of parameters conditional on the training sequence data X_{train} and the training output data Y_{train} is, according to Bayes' rule, given by

$$P(\omega \mid X_{\text{train}}, Y_{\text{train}}) = P(Y_{\text{train}} \mid X_{\text{train}}, \omega)P(\omega) \Big/ \int P(Y_{\text{train}} \mid X_{\text{train}}, \omega')P(\omega')d\omega'$$

Here, $P(\omega)$ is the *prior* distribution of the set of parameters and $P(Y_{\text{train}} \mid X_{\text{train}}, \omega)$ may be assumed to follow

$$\mathrm{P}(\mathrm{Y} \mid \mathrm{X}, \omega) = \begin{cases} \mathrm{N}(\hat{\mathrm{g}}(\mathrm{x}), \mathrm{V}), & \text{for continous Y} \\ \exp\left(\widehat{\mathrm{g}_{\mathrm{a}}}(\mathrm{x})\right) \Big/ \sum_{\mathrm{a}'} \exp\left(\widehat{\mathrm{g}_{\mathrm{a}'}}(\mathrm{x})\right), & \mathrm{Y} = \mathrm{a} \end{cases}$$

That is, in the continuous case, the likelihood is Gaussian with mean $\hat{g}(x)$ and covariance matrix V whereas for classification, the conditional probability is a softmax likelihood (assuming the output layer is a softmax layer and the cross-entropy is the loss function). The posterior probability $P(\omega \mid X_{\text{train}}, Y_{\text{train}})$ can be calculated using Markov Chain Monte Carlo (MCMC), as the integral (i.e., the so-called *model evidence* $P(Y_{\text{train}} \mid X_{\text{train}})$) is known to be typically intractable and have no analytical formulation. Prediction of the phenotype y for *any* given input x can then be quantified through the following inference

$$P(y \mid x, X_{\text{train}}, Y_{\text{train}}) = \int P(y \mid x,\ \omega')P(\omega' \mid X_{\text{train}}, Y_{\text{train}})d\omega'$$

Note the left hand side of this equation can also be sampled using the neural network.

Joint model consisting of the neural network derived antibody-escape phenotypes

The spike RBD binding affinity was jointly modeled using the convolutional neural network derived (i.e., estimated, mutation-determined) antibody-escape phenotype for the ten Abs. The ten "escape fraction" features were scaled and randomly split into 80% training and 20% testing subsets for the Random Forest regressor implemented using sckit-learn[34]. We applied hyper parameter tuning to ensure the best performance.

AAindex shuffling experiment

We performed a shuffling experiment on the AAindex to assess the extent to which the selected amino acid physicochemical properties impact the performance of the neural network architectures. All elements of the AAindex matrix were re-sampled in-place so that the original covariances among rows or columns were broken. As a result, we generated 100 new sets of (pseudo) AAindex tables. The p-value, i.e., the probability, if the null hypothesis H_0 (of no gain) is true, that the Spearman correlation ρ is at least as great as the value ρ_{actual} from the observed data is calculated as follows

$$p := P(\rho \geq \rho_{\text{actual}} \mid \mathrm{H}_0) \approx (|\rho_{\text{sim}} \geq \rho_{\text{actual}}| + 1)/(n_{\text{sims}} + 1)$$

where ρ_{sim} is the correlation from a shuffled dataset, n_{sims} is the total number of shuffled datasets, and | | is the count operator.

References

1. Zhang, H., Penninger, J. M., Li, Y., Zhong, N. & Slutsky, A. S. Angiotensin-converting enzyme 2 (ACE2) as a SARS-CoV-2 receptor: molecular mechanisms and potential therapeutic target. *Intensive Care Med.* **46**, 586–590 (2020).
2. Lan, J. *et al.* Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature* **581**, 215–220 (2020).
3. Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273 (2020).
4. Chan, K. K. *et al.* Engineering human ACE2 to optimize binding to the spike protein of SARS coronavirus 2. *Science* **369**, 1261–1265 (2020).
5. Han, Y. & Král, P. Computational Design of ACE2-Based Peptide Inhibitors of SARS-CoV-2. *ACS Nano* **14**, 5143–5147 (2020).
6. Procko, E. Deep mutagenesis in the study of COVID-19: a technical overview for the proteomics community. *Expert Rev. Proteomics* **0**, 1–6 (2020).
7. 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).
8. Blanco, J. D., Hernandez-Alias, X., Cianferoni, D. & Serrano, L. In silico mutagenesis of human ACE2 with S protein and translational efficiency explain SARS-CoV-2 infectivity in different species. *PLOS Comput. Biol.* **16**, e1008450 (2020).
9. Calcagnile, M. *et al.* Molecular docking simulation reveals ACE2 polymorphisms that may increase the affinity of ACE2 with the SARS-CoV-2 Spike protein. *Biochimie* **180**, 143–148 (2021).
10. Hussain, M. *et al.* Structural variations in human ACE2 may influence its binding with SARS-CoV-2 spike protein. *J. Med. Virol.* **92**, 1580–1586 (2020).

11. Pietzner, M. *et al.* Genetic architecture of host proteins involved in SARS-CoV-2 infection. *Nat. Commun.* **11**, 6397 (2020).

12. Fowler, D. M. *et al.* High-resolution mapping of protein sequence-function relationships. *Nat. Methods* **7**, 741–746 (2010).

13. Fowler, D. M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nat. Methods* **11**, 801–807 (2014).

14. Stein, A., Fowler, D. M., Hartmann-Petersen, R. & Lindorff-Larsen, K. Biophysical and Mechanistic Models for Disease-Causing Protein Variants. *Trends Biochem. Sci.* **44**, 575–588 (2019).

15. Angermueller, C., Pärnamaa, T., Parts, L. & Stegle, O. Deep learning for computational biology. *Mol. Syst. Biol.* **12**, 878 (2016).

16. Gelman, S., Romero, P. A. & Gitter, A. Neural networks to learn protein sequence-function relationships from deep mutational scanning data. *bioRxiv* 2020.10.25.353946 (2020) doi:10.1101/2020.10.25.353946.

17. Li, B., Yang, Y. T., Capra, J. A. & Gerstein, M. B. Predicting changes in protein thermodynamic stability upon point mutation with deep 3D convolutional neural networks. *PLOS Comput. Biol.* **16**, e1008291 (2020).

18. Heinzelman, P. & Romero, P. A. Discovery of human ACE2 variants with altered recognition by the SARS-CoV-2 spike protein. *bioRxiv* (2020) doi:10.1101/2020.09.17.301861.

19. Greaney, A. J. *et al.* Complete Mapping of Mutations to the SARS-CoV-2 Spike Receptor-Binding Domain that Escape Antibody Recognition. *Cell Host Microbe* (2020) doi:10.1016/j.chom.2020.11.007.

20. Elbe, S. & Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob. Chall.* **1**, 33–46 (2017).

21. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data – from vision to reality. *Eurosurveillance* **22**, 30494 (2017).

22. Sarkisyan, K. S. *et al.* Local fitness landscape of the green fluorescent protein. *Nature* **533**, 397–401 (2016).

23. Gray, V. E., Hause, R. J., Luebeck, J., Shendure, J. & Fowler, D. M. Quantitative Missense Variant Effect Prediction Using Large-Scale Mutagenesis Data. *Cell Syst.* **6**, 116-124.e3 (2018).

24. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv180203426 Cs Stat* (2020).

25. Addetia, A. *et al.* Neutralizing Antibodies Correlate with Protection from SARS-CoV-2 in Humans during a Fishery Vessel Outbreak with a High Attack Rate. *J. Clin. Microbiol.* **58**, (2020).

26. Leek, J. T. & Storey, J. D. Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. *PLOS Genet.* **3**, e161 (2007).

27. Listgarten, J., Kadie, C., Schadt, E. E. & Heckerman, D. Correction for hidden confounders in the genetic analysis of gene expression. *Proc. Natl. Acad. Sci.* **107**, 16465–16470 (2010).

28. Ali, A. & Vijayan, R. Dynamics of the ACE2–SARS-CoV-2/SARS-CoV spike protein interface reveal unique mechanisms. *Sci. Rep.* **10**, 14214 (2020).

29. Hou, Y. J. *et al.* SARS-CoV-2 D614G variant exhibits efficient replication ex vivo and transmission in vivo. *Science* **370**, 1464–1468 (2020).

30. Lauring, A. S. & Hodcroft, E. B. Genetic Variants of SARS-CoV-2—What Do They Mean? *JAMA* (2021) doi:10.1001/jama.2020.27124.

31. Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations. *Virological* https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563 (2020).

32. Kawashima, S. *et al.* AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res.* **36**, D202–D205 (2008).

33. Warner, S. The topology of compact convergence on continuous function spaces. *Duke Math. J.* **25**, 265–282 (1958).

34. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

Author contributions

E.R.G. and B.W. designed the study and wrote the manuscript. B.W. performed the analyses. E.R.G. supervised and acquired funding for the study.

Competing interests

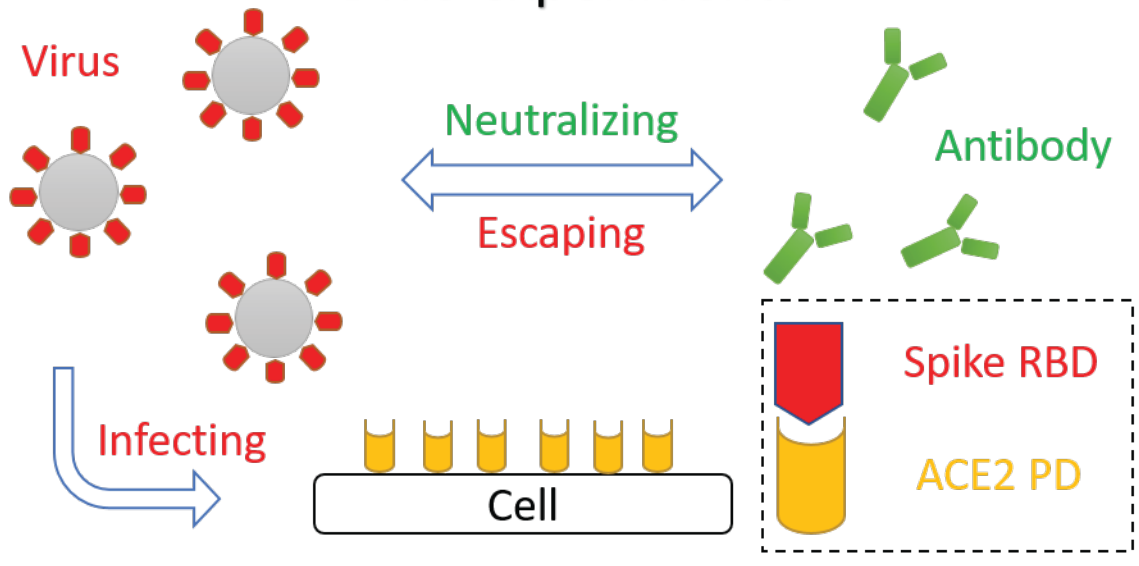
E.R.G. receives an honorarium from the journal *Circulation Research* of the American Heart Association, as a member of the Editorial Board.

Funding

This research is supported by the National Institutes of Health (NIH) Awards R35HG010718 and R01HG011138.

A

DMS experiments



B

DL pipeline

Input
- protein sequence
- mutant-score profile

Encoding
- AAindex
- one-hot

Networks
- Linear Regression
- Multilayer Perceptron
- Convolutional Networks

C

Convolutional Networks

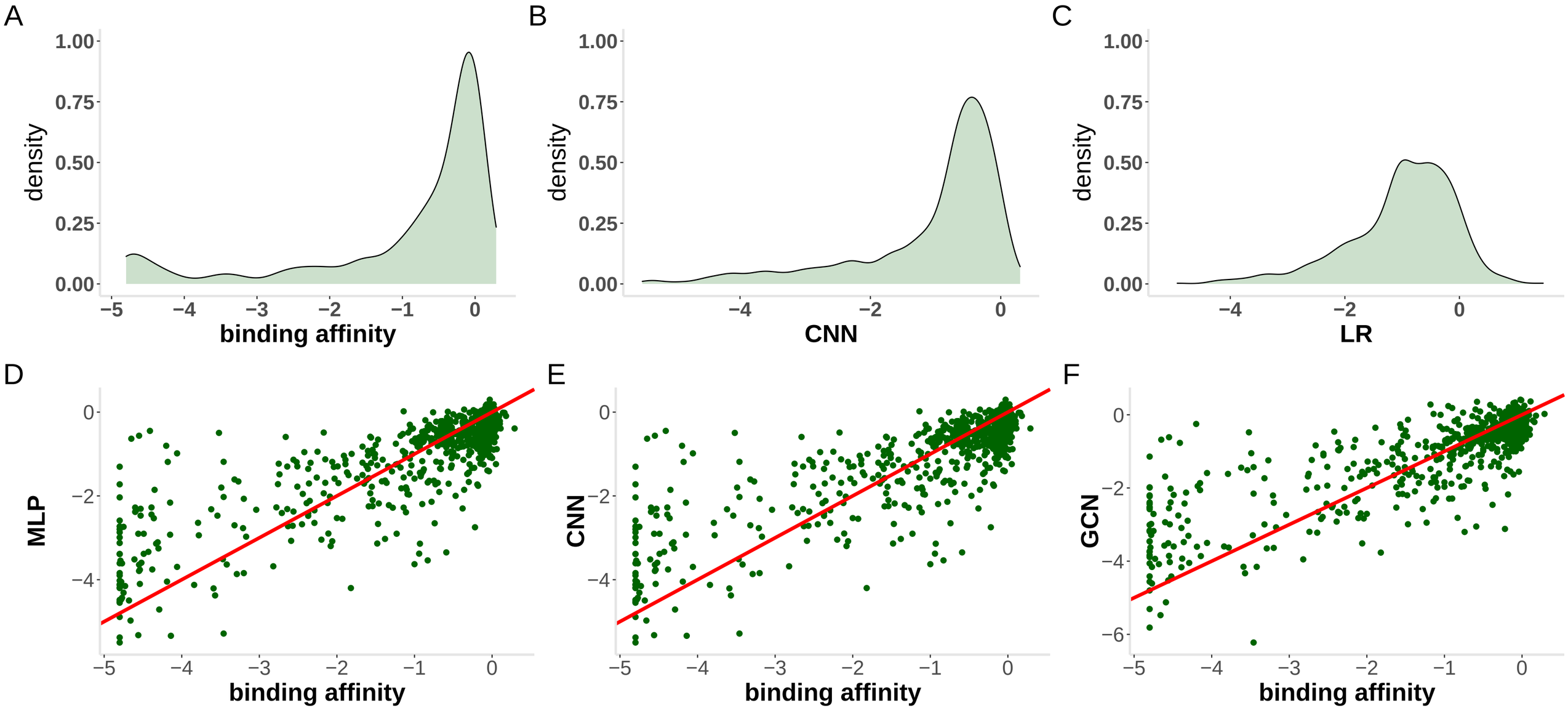
TNLCPFGEVFNATRFASVYAWN

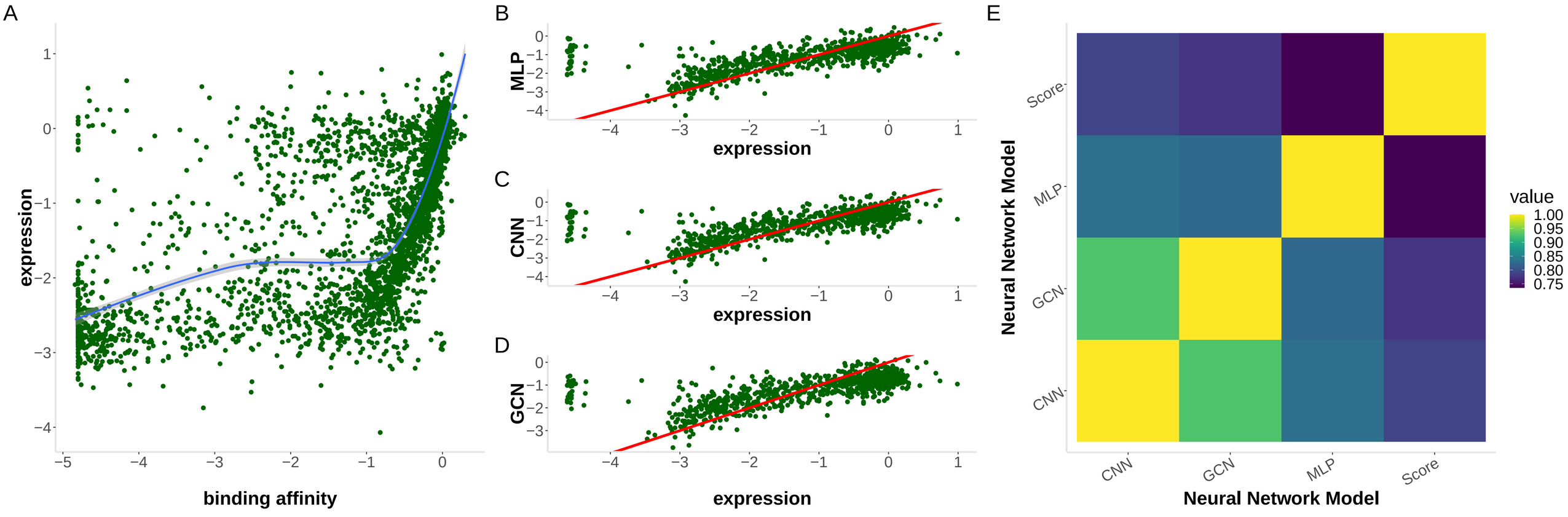
RKRISNCVADYSVLYNSASFSTFK

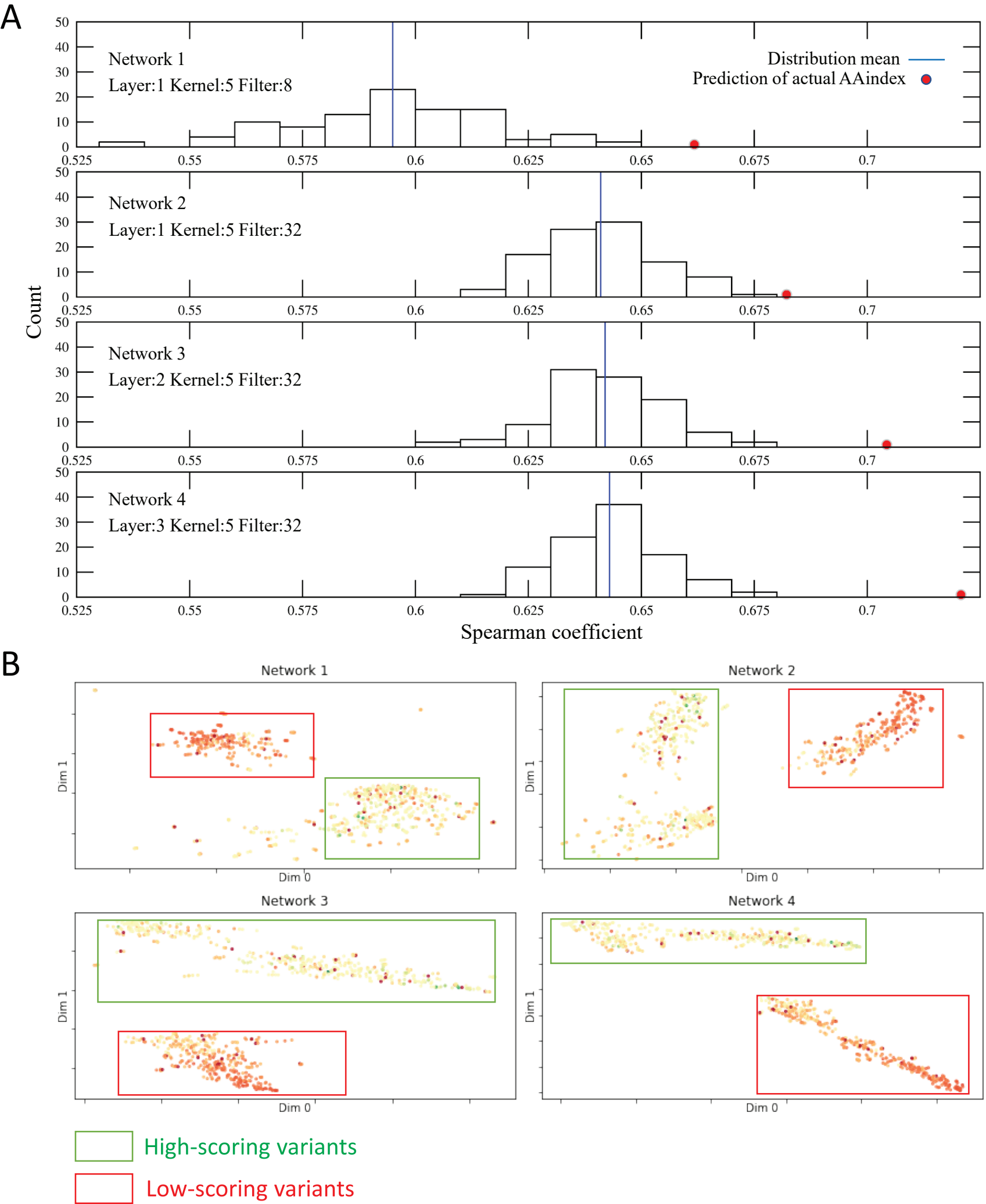
CYGVSPTKLNDLCFTNVYADSFVI....

Filter on Sequence

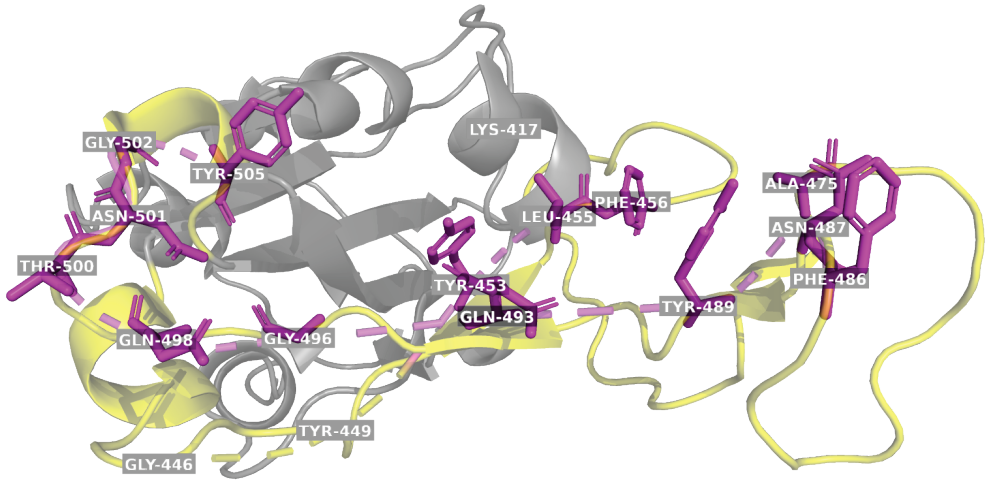
Filter on Structural Graph


A
B
C
D
E
F
density
1.00
0.75
0.50
0.25
0.00
−5
−4
−3
−2
−1
0
binding affinity
CNN
LR
MLP
GCN
−6

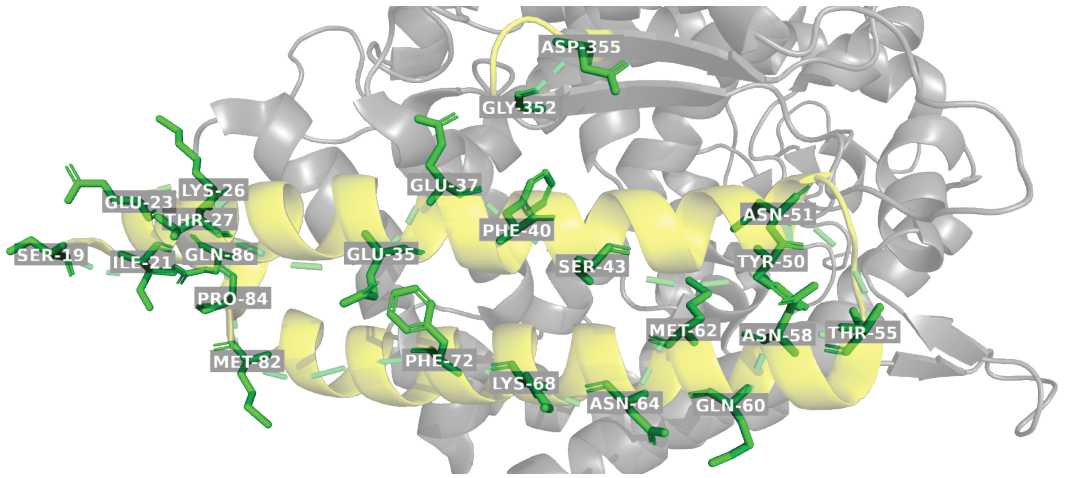

A
expression
binding affinity
B
MLP
expression
C
CNN
expression
D
GCN
expression
E
Neural Network Model
Score
MLP
GCN
CNN
value
1.00
0.95
0.90
0.85
0.80
0.75


A
Network 1
Layer:1 Kernel:5 Filter:8
Distribution mean
Prediction of actual AAindex
Network 2
Layer:1 Kernel:5 Filter:32
Network 3
Layer:2 Kernel:5 Filter:32
Network 4
Layer:3 Kernel:5 Filter:32
Count
Spearman coefficient
B
Network 1
Network 2
Network 3
Network 4
Dim 0
Dim 1
High-scoring variants
Low-scoring variants

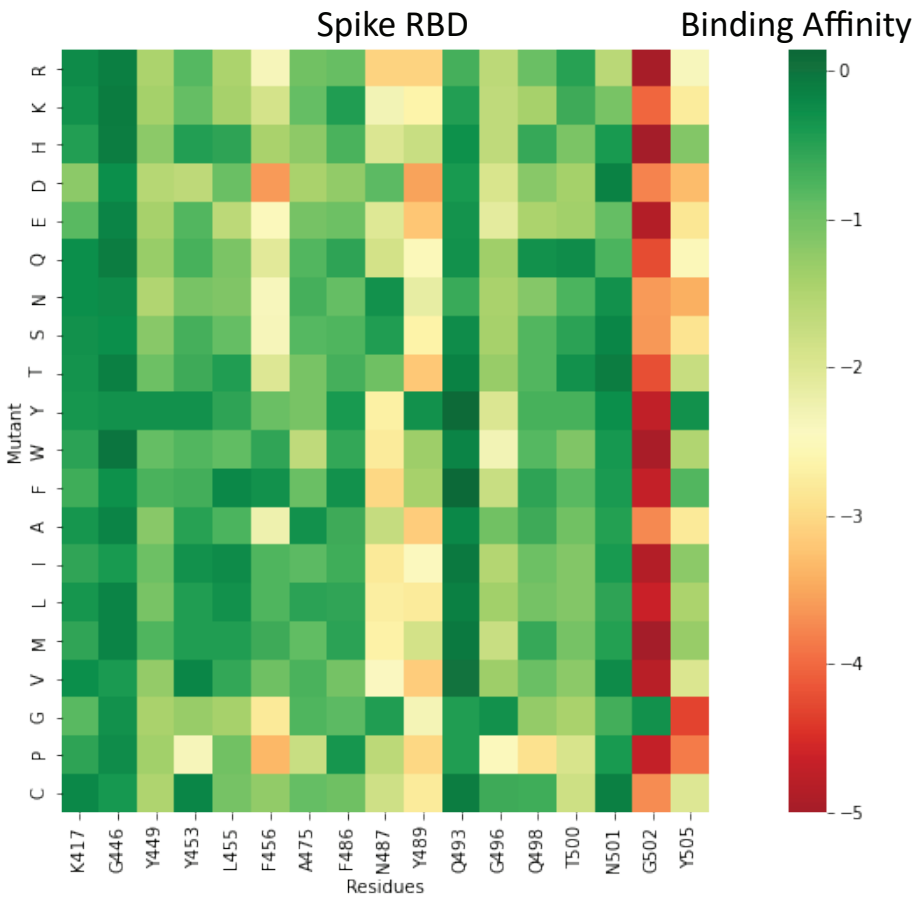
A


GLY-502
TYR-505
LYS-417
ASN-501
LEU-455
PHE-456
ALA-475
ASN-487
THR-500
TYR-453
PHE-486
GLN-493
TYR-489
GLN-498
GLY-496
TYR-449
GLY-446

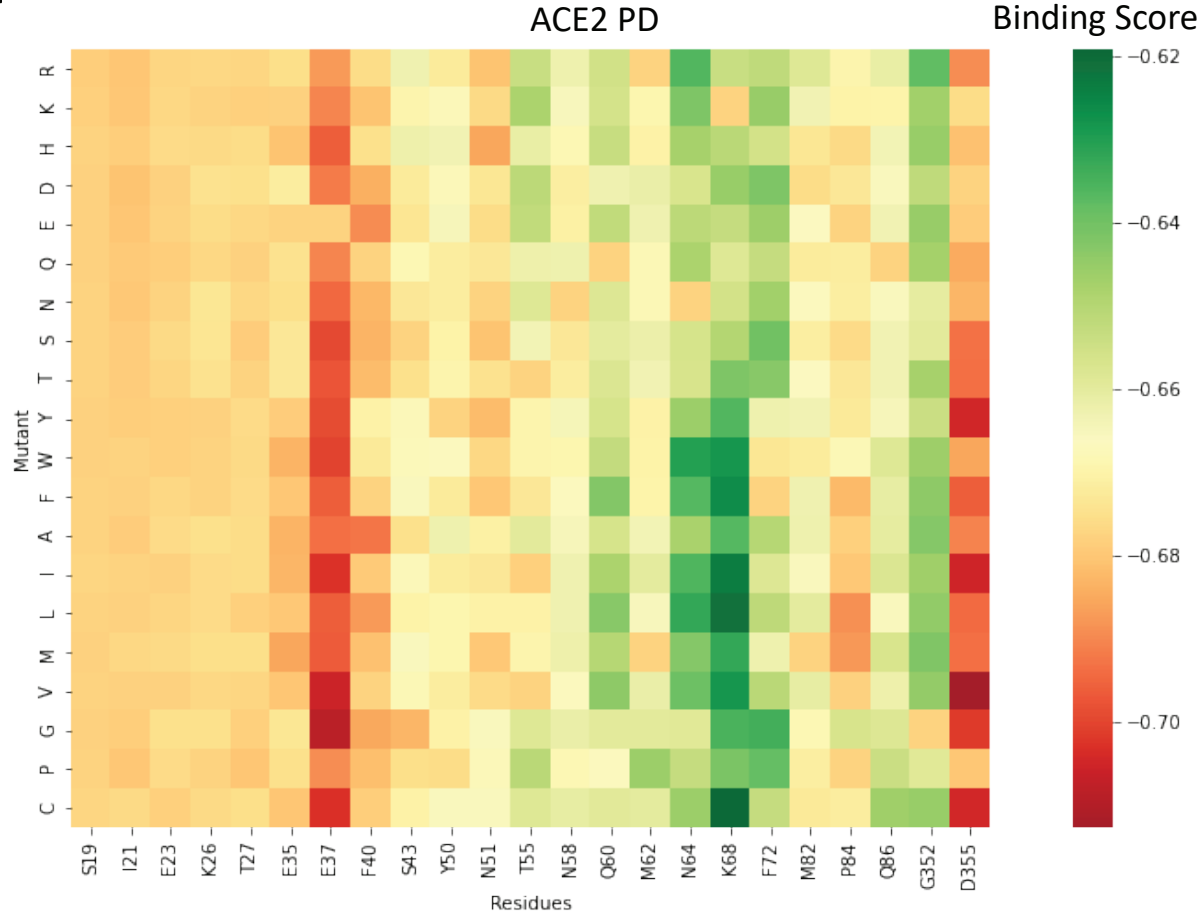
D


ASP-355
GLY-352
GLU-37
LYS-26
GLU-23
THR-27
PHE-40
ASN-51
SER-19
ILE-21
GLN-86
GLU-35
SER-43
TYR-50
PRO-84
MET-62
ASN-58
THR-55
MET-82
PHE-72
LYS-68
ASN-64
GLN-60

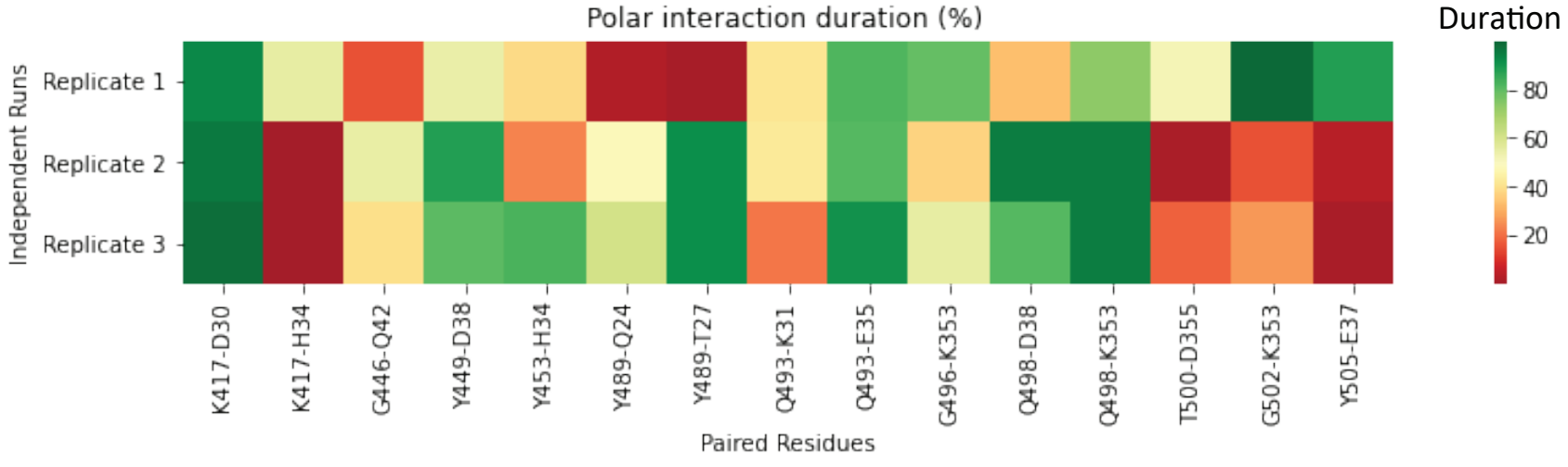
B

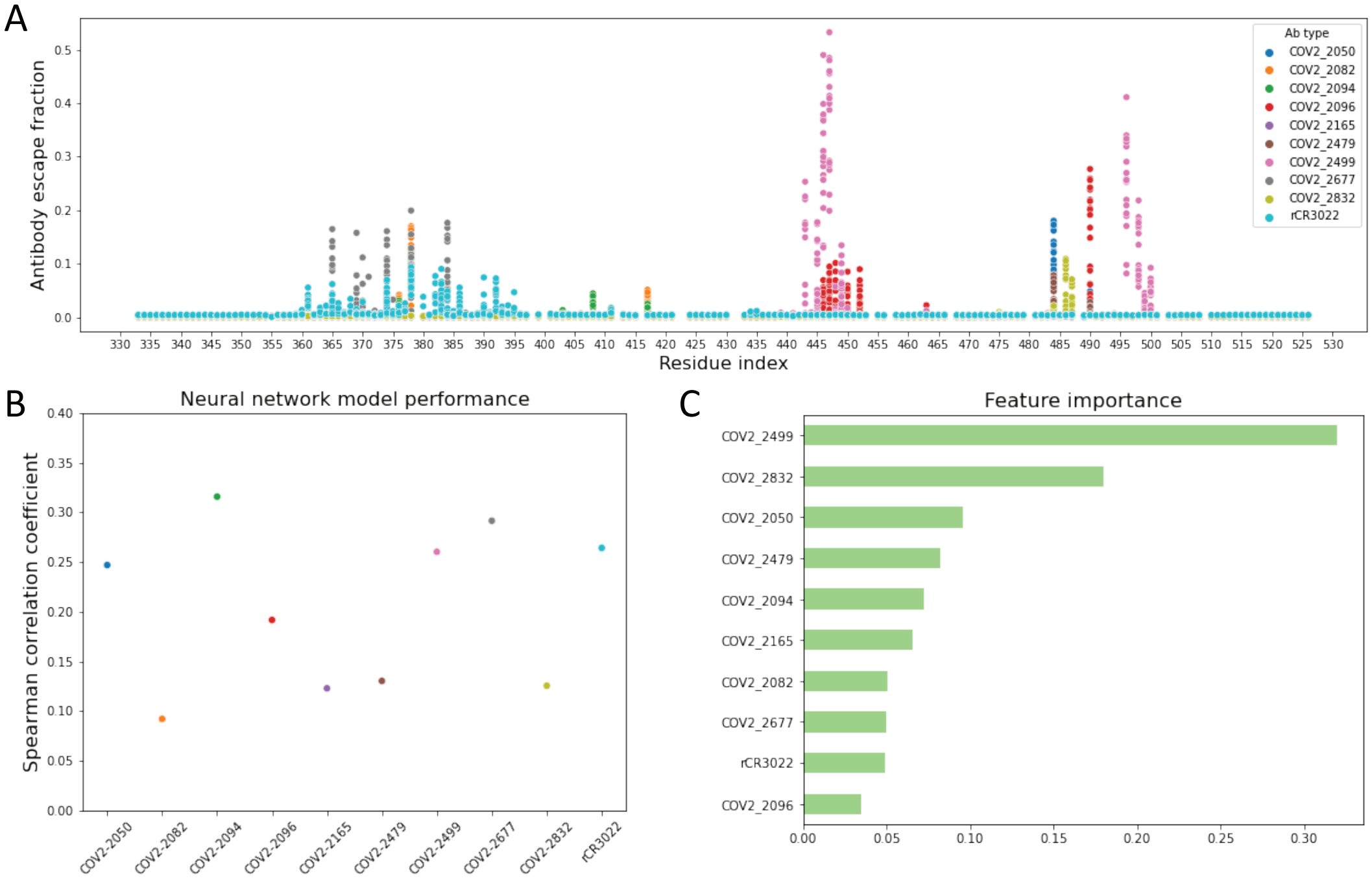

Spike RBD
Binding Affinity
Mutant
R K H D E Q N S T Y W F A I L M V G P C
K417 G446 Y449 Y453 L455 F456 A475 F486 N487 Y489 Q493 G496 Q498 T500 N501 G502 Y505
Residues
0
−1
−2
−3
−4
−5

E


ACE2 PD
Binding Score
Mutant
R K H D E Q N S T Y W F A I L M V G P C
S19 I21 E23 K26 T27 E35 E37 F40 S43 Y50 N51 T55 N58 Q60 M62 N64 K68 F72 M82 P84 Q86 G352 D355
Residues
−0.62
−0.64
−0.66
−0.68
−0.70

C


Polar interaction duration (%)
Duration
Independent Runs
Replicate 1
Replicate 2
Replicate 3
K417-D30
K417-H34
G446-Q42
Y449-D38
Y453-H34
Y489-Q24
Y489-T27
Q493-K31
Q493-E35
G496-K353
Q498-D38
Q498-K353
T500-D355
G502-K353
Y505-E37
Paired Residues
80
60
40
20


A
Antibody escape fraction
Residue index
Ab type
COV2_2050
COV2_2082
COV2_2094
COV2_2096
COV2_2165
COV2_2479
COV2_2499
COV2_2677
COV2_2832
rCR3022
B
Neural network model performance
Spearman correlation coefficient
C
Feature importance
COV2_2499
COV2_2832
COV2_2050
COV2_2479
COV2_2094
COV2_2165
COV2_2082
COV2_2677
rCR3022
COV2_2096