Predicting Immune Escape with Pretrained Protein Language Model Embeddings

Kyle Swanson
Department of Computer Science
Stanford University
swansonk@stanford.edu

Howard Chang
Center for Personal Dynamic Regulomes
Howard Hughes Medical Institute
Stanford University
howchang@stanford.edu

James Zou
Department of Biomedical Data Science
Stanford University
jamesz@stanford.edu

Abstract

Assessing the severity of new pathogenic variants requires an understanding of which mutations will escape the human immune response. Even single point mutations to an antigen can cause immune escape and infection via abrogation of antibody binding. Recent work has modeled the effect of single point mutations on proteins by leveraging the information contained in large-scale, pretrained protein language models (PLMs). PLMs are often applied in a zero-shot setting, where the effect of each mutation is predicted based on the output of the language model with no additional training. However, this approach cannot appropriately model immune escape, which involves the interaction of two proteins—antibody and antigen—instead of one protein and requires making different predictions for the same antigenic mutation in response to different antibodies. Here, we explore several methods for predicting immune escape by building models on top of embeddings from PLMs. We evaluate our methods on a SARS-CoV-2 deep mutational scanning dataset and show that our embedding-based methods significantly outperform zero-shot methods, which have almost no predictive power. We also highlight insights gained into how best to use embeddings from PLMs to predict escape. Despite these promising results, simple statistical and machine learning baseline models that do not use pretraining perform comparably, showing that computationally expensive pretraining approaches may not be beneficial for escape prediction. Furthermore, all models perform relatively poorly, indicating that future work is necessary to improve escape prediction with or without pretrained embeddings[1].

1 Introduction

Pathogens are constantly evolving in their search to evade the immune system and infect host organisms [1]. In many organisms, including humans, this evolutionary battle occurs in the context of antibody-antigen interactions [2]. Antibodies are proteins produced by the immune system that are designed to bind to antigens, which are pathogenic proteins that induce an immune response. Antibodies that effectively bind to an antigen and neutralize the pathogen put evolutionary pressure

[1]Our code is available at https://github.com/swansonk14/escape_embeddings and all data, embeddings, and results are available at https://drive.google.com/drive/folders/18heVMWK46ExHkSeixyrNiJLovnIbZ4jg?usp=share_link

NeurIPS 2022 AI for Science Workshop.

on the pathogen to mutate its antigen in a process known as immune escape [3]. Predicting which mutations cause escape is crucial to identifying dangerous pathogenic variants that can cause infection and disease even in the presence of antibodies from prior infection, vaccination, or therapies [3–5].

Machine learning models have been developed that can predict the effect of protein mutations on various protein functions [4, 6–9]. Recent approaches to mutation effect prediction have leveraged large protein language models (PLMs) that have been trained in an unsupervised manner on huge databases with hundreds of millions to billions of protein sequences [10, 11]. PLMs learn the underlying statistics of naturally occurring protein sequences and can predict the likelihood that a given amino acid appears at a position in a protein. These likelihoods can be used to predict the effect of a mutation in a zero-shot manner (i.e., without additional training) by comparing the likelihood of the mutated amino acid to that of the wildtype amino acid at a given position [6, 7, 12, 13].

However, a major limitation of the zero-shot likelihood approach is that it predicts the same likelihood for a mutation regardless of the protein function in question [7]. Since proteins can have multiple functions that are affected differently by the same mutation, one likelihood cannot model the effect of a mutation on all of these functions simultaneously. Additionally, the likelihood only accounts for the protein that is mutated, which means that it ignores any interacting proteins such as antibodies.

We propose to overcome these limitations by modeling immune escape using antibody and antigen embeddings produced by a PLM. These embeddings encode information about the protein, including aspects of 3D structure, that can inform the effect of protein mutations [14]. We build a lightweight neural model that learns to extract information from the embeddings to predict escape in an antibody-dependent manner. We develop several variants of this embedding-based approach and evaluate them on a SARS-CoV-2 deep mutational scanning dataset from Cao et al. [5]. We show that embeddings significantly outperform zero-shot likelihoods, which have almost no predictive power. We discuss insights gained from our experiments about how best to use embeddings from PLMs to predict escape. We also develop two statistical baseline models and a machine learning model that do not rely on the pretrained models. These models perform comparably to the embedding models, indicating that pretrained embeddings may not be beneficial for predicting escape. Furthermore, the relatively poor performance of all models demonstrates that future work is necessary to improve escape prediction with or without pretrained embeddings.

2 Methods

Our goal is to design a model that can predict the effect of antigenic mutations on the neutralization ability of antibodies. Below, we first outline notation used throughout this section. Then, we describe several models to predict this escape effect either using simple statistics, a machine learning model trained from scratch, or machine learning models built on top of PLMs.

2.1 Notation

An antigen is a sequence of amino acids $A = \{A_1, A_2, \ldots, A_n\}$ with each $A_i \in \mathbf{P}$ where $\mathbf{P}$ is the set of 20 naturally occurring amino acids. The antigen is the subsequence of a protein in a pathogen that antibodies bind to. Each location $s \in \{1, 2, \ldots, n\}$ in the antigen is called a *site*. The original, unmutated sequence of antigen amino acids is referred to as the *wildtype* sequence. Here, we consider single point mutations, where a single site s in the antigen has its amino acid mutated from the wildtype amino acid A_s to the mutant amino acid $M \in \mathbf{P} \setminus A_s$, which is one of the other 19 possible amino acids. The mutated antigen sequence then becomes $A^{s,M} = \{A_1, \ldots, A_{s-1}, M, A_{s+1}, \ldots, A_n\}$. In this paper, we consider a single antigen A and all possible $n \times 19$ single point mutations across the n antigen sites.

An antibody B is a protein that consists of four chains, where each chain is a sequence of amino acids. Among the four chains, two are identical chains called the heavy chain with the sequence $B^H = \{B_1^H, B_2^H, \ldots, B_h^H\}$ and two are identical chains called the light chain with the sequence $B^L = \{B_1^L, B_2^L, \ldots, B_l^L\}$. The antibody as a whole is represented by the pair of unique chains, $B = (B^H, B^L)$. Here, we consider many different antibodies $B \in \mathbf{B}$ where $\mathbf{B}$ is a set of antibodies, all of which bind to the same antigen A.

When the antibody comes into contact with the antigen, and interactions between the antibody and antigen amino acids can lead to binding and subsequent neutralization of the pathogen. Mutations

to the antigen may inhibit antibody binding. The degree to which antibody binding is reduced is represented by a value called the escape score. For antigen A with amino acid M at site s in the presence of antibody B, the experimentally determined escape score is $E(A, s, M, B) \in [0, 1]$, with larger numbers indicating more escape (less antibody binding after the mutation). If M is the wildtype amino acid at site s, i.e., $M = A_s$, then the escape score is zero since the antigen is unchanged. If there is a mutation so that $M \neq A_s$, then the escape score may be zero or non-zero depending on whether and to what degree the mutation affects antibody binding.

Given a set of training data points $T = \{(A, s, M, B)\}_{s\in[1,n],M\in\mathbf{P},B\in\mathbf{B}}$ consisting of one fixed antigen with many site, mutation, and antibody combinations, along with their known escape scores given by E, our goal is to build a model that can predict the escape score for a site, mutation, and antibody combination not in the training set.

2.2 Mutation Model

The mutation model MM models escape as a function of the change in amino acid from the wildtype to the mutant. This assumes that amino acid changes have consistent escape effects regardless of the site and antibody. The model is fitted by computing the average escape score in the training set for each pair of wildtype (wt) and mutant (mut) amino acids across all sites and antibodies, i.e.,

$$\texttt{MM}(M^{wt}, M^{mut}) = \frac{1}{Z}\sum_{s=1}^{n}\sum_{B\in\mathbf{B}} E(A, s, M^{mut}, B) \cdot \mathbb{1}_{A_s=M^{wt}} \cdot \mathbb{1}_{(A,s,M^{mut},B)\in T} \tag{1}$$

where n is the number of antigen sites, $\mathbb{1}$ is an indicator variable, and

$$Z = \sum_{s=1}^{n}\sum_{B\in\mathbf{B}} \mathbb{1}_{A_s=M^{wt}} \cdot \mathbb{1}_{(A,s,M^{mut},B)\in T} \tag{2}$$

is the total number of data points in the training set where site s is mutated from M^{wt} to M^{mut}. To make a prediction for a new site s, mutation M, and antibody B, the model simply outputs $\texttt{MM}(A, s, M, B) = \texttt{MM}(A_s, M)$, thereby ignoring the site, the rest of the antigen sequence, and the antibody sequences. Since there are 20 amino acids, this model has $20 \times 20 = 400$ parameters.

2.3 Site Model

The site model SM models escape as a function of the antigen site. This assumes that sites have consistent escape effects regardless of the wildtype and mutant amino acids and the antibody. The model is fitted by computing the average escape score in the training set for each antigen site across all mutant amino acids and across all antibodies, i.e.,

$$\texttt{SM}(s) = \frac{1}{Z}\sum_{M\in\mathbf{P}}\sum_{B\in\mathbf{B}} E(A, s, M, B) \cdot \mathbb{1}_{(A,s,M,B)\in T} \tag{3}$$

where $\mathbb{1}$ is an indicator variable and

$$Z = \sum_{M\in\mathbf{P}}\sum_{B\in\mathbf{B}} \mathbb{1}_{(A,s,M,B)\in T} \tag{4}$$

is the total number of data points in the training set for site s. To make a prediction for a new site s, mutation M, and antibody B, the model simply outputs $\texttt{SM}(A, s, M, B) = \texttt{SM}(s)$, thereby ignoring the entire antigen sequence, including wildtype and mutant amino acids, and the antibody sequences. The model has one parameter for each antigen site for a total of n parameters.

2.4 RNN

As a non-pretrained baseline embedding model, we train a recurrent neural network (RNN) from scratch. Given antigen A, site s, and mutated amino acid M at that site, we construct the mutated antigen sequence $A^{s,M}$ and embed it using a bidirectional LSTM [15]. We then extract an embedding in one of two ways and pass that embedding through a small multilayer perceptron to predict escape. In the model we call RNN Seq, this embedding is the average of the hidden embeddings for the n amino acids in the antigen. In the model we call RNN Res, this embedding is the output embedding corresponding to the mutated site s. Since the RNN model ignores the antibody sequences, it computes $\texttt{RNN}(A, s, M, B) = \texttt{RNN}(A, s, M)$.

2.5 Likelihood Model

For our likelihood model $\texttt{L}$, we adopt the zero-shot mutation prediction framework of Meier et al. [7]. In this framework, a PLM is applied to the antigen sequence $A^{s,\texttt{<mask>}}$, where the amino acid at site s has been replaced with a <mask> token. The escape score is predicted as the model's log odds ratio of the mutated amino acid M versus the wildtype amino acid A_s at that site s. Specifically, the model computes

$$\texttt{L}(A, s, M) = \log(p(M|A^{s,\texttt{<mask>}})) - \log(p(A_s|A^{s,\texttt{<mask>}})) \tag{5}$$

where $p(\cdot|\cdot)$ is the probability the model assigns to an amino acid at a specific site within a given sequence. The likelihood model does not require any additional training, and it does not incorporate the antibody sequences so $\texttt{L}(A, s, M, B) = \texttt{L}(A, s, M)$.

2.6 Embedding Models

Models that use PLM embeddings instead of likelihoods provide a more flexible way of predicting mutation effect. In these models, we train a small multilayer perceptron to use some form of protein embedding as input to predict the escape score. All of the models use an embedding of the mutated antigen sequence $A^{s,M}$, and some additionally use an embedding of the wildtype antigen sequence A and/or embeddings of the antibody heavy and light chains, B^H and B^L. The embedding variants are described below.

Antigen Mutant Sequence. The PLM is given the mutated antigen sequence $A^{s,M}$ and computes the embedding matrix $R = \texttt{PLM}(A^{s,M})$ with $R \in \mathbb{R}^{n \times d}$, which contains a d-dimensional embedding for each amino acid in the antigen sequence. The embedding of each site encodes the identity of the amino acid at that site as well as its role in the context of the antigen sequence. The amino acid embeddings are averaged to form a single embedding for the full antigen sequence, $R^{seq} = \frac{1}{n}\sum_{s=1}^{n} R_s$ with $R^{seq} \in \mathbb{R}^d$. We refer to this embedding as Antigen Mut Seq.

Antigen Mutant Residue. As above, the PLM computes the embedding matrix for the mutated antigen sequence $A^{s,M}$. Here, the embedding $R^{res} = R_s$ of the mutated residue at site s is used instead of the sequence average. We refer to this embedding as Antigen Mut Res.

Antigen Difference. Antigen embeddings for both the mutated antigen sequence $A^{s,M}$ and the wildtype antigen sequence A are computed, either both at the sequence level or both at the residue level. The difference between the embeddings (mutant minus wildtype) is computed. We refer to these embeddings as Antigen Seq Diff for the difference of sequence embeddings and Antigen Res Diff for the difference of residue embeddings. We also concatenate the mutant and difference embeddings to form what we call Antigen Seq MutDiff and Antigen Res MutDiff embeddings.

Antibody. We experiment with three different methods of incorporating antibody information into the antigen embedding models. First, we concatenate the antigen embedding (of any form) with a one-hot encoding of the antibody to provide the model with the antibody identity but without any embedding information (Antigen + Antibody OH model). Second, we use the PLM to embed the heavy and light chains of the antibody, B^H and B^L, and we concatenate those two embeddings with the antigen embedding (of any form) to obtain a $3d$-dimensional embedding (Antigen + Antibody Emb model). Finally, we create embeddings of combined antibody-antigen sequences. For each antibody, we create one sequence with the heavy chain B^H and mutated antigen $A^{s,M}$ and another sequence with the light chain B^L and mutated antigen $A^{s,M}$. In both cases, the antibody and antigen sequences are joined by seven repeats of a glycine-glycine-serine linker. Both combined sequences are embedded by the PLM, sequence averaged, and then concatenated to form a single $2d$-dimensional embedding for the antibody and mutated antigen. (Antigen Linker Antibody model). This linker design is inspired by the use of linkers to enable protein complex prediction from single chain protein structure prediction models like AlphaFold2 [16].

3 Experiments

Here, we describe the data we use to train and evaluate our model as well as the data splits, tasks, metrics, and models that we use.

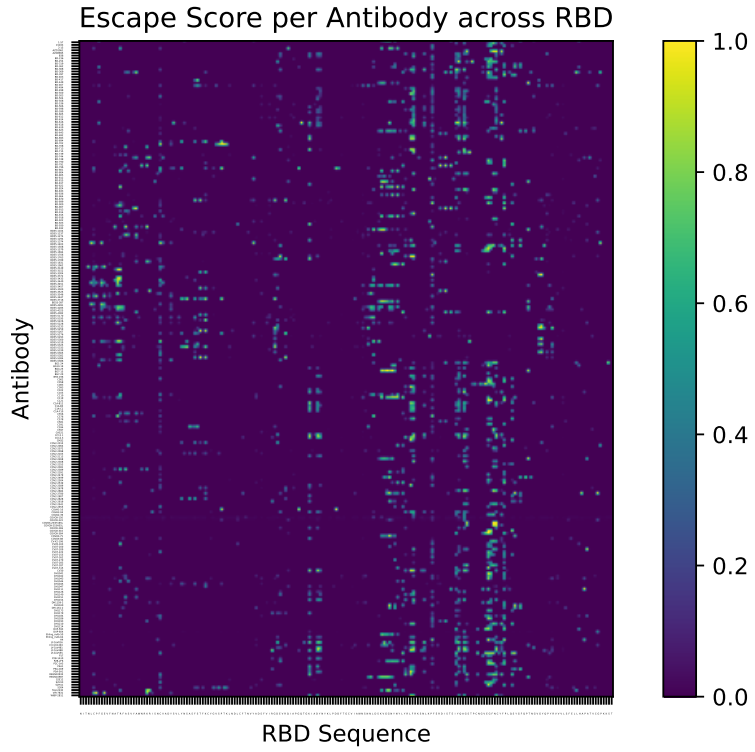

Figure 1: The average escape score across all amino acid mutations for each antibody and each antigen site in the receptor binding domain (RBD) in the SARS-CoV-2 deep mutational scanning data from Cao et al. [5]. Note: The 74 escape scores greater than 1 (max 3.6) are truncated to 1.

3.1 Data

We use SARS-CoV-2 deep mutational scanning data from Cao et al. [5]. This data consists of 247 antibodies that are known to neutralize the original strain of SARS-CoV-2 by binding to the receptor binding domain (RBD) of the spike protein. The neutralization ability of each antibody is measured for the wildtype RBD antigen as well as for all 3,819 single point mutations to the antigen (201 sites in the RBD with 19 amino acid substitutions at each site). For each antibody and each antigen mutation, an escape score is computed as a normalized measure of the reduction in antibody neutralization compared to the wildtype antigen (see Figures 1 and S.2). Of the 943,293 escape scores in the dataset, 30,658 (3.2%) are non-zero, usually in the range [0, 1] with 74 outliers above 1 with a max of 3.6 (see Figure S.1). Cao et al. [5] clustered the 247 antibodies into six groups based on their escape scores (see Figure S.3).

3.2 Data Splits

The practical usefulness of an escape prediction model, as well as the difficulty of learning such a model, depends heavily on how the data is split. Below we describe and motivate our data splits.

Mutation. Mutations are randomly split between train and test. This assumes that for a new antibody, we already know escape scores for some but not all mutations across all antigen sites. This split corresponds to a scenario in which we have a significant amount of escape data, either from laboratory mutation experiments or real-world infections by mutated pathogens, across antigen sites for a particular antibody. However, we do not have escape data for the complete set of mutations, so a model trained in this setting would be able to fill in the escape effect of any missing mutations.

Site. Antigen sites are randomly split between train and test. This assumes that for a new antibody, we already know escape scores for some but not all antigen sites. In this split, we are more conservative and assume that we only have escape data for some antigen sites and need to make predictions for other antigen sites. This still requires knowing some escape scores for the antibody, but we no longer need to know escape scores across all sites, making it possible to use any available escape data.

Antibody. Antibodies are randomly split between train and test. This assumes that we do not know any escape scores for a new antibody. This split models a situation in which some antibodies have already been experimentally evaluated and have escape data, and we want to make predictions for a new antibody for which we do not yet have any escape data.

Antibody group. Antibody groups, as defined by a clustering of escape scores, are randomly split between train and test. This assumes that we do not know any escape scores for a new antibody, and furthermore, no antibody in the train set has a similar pattern of escape to this antibody. This split is especially relevant since new groups of antibodies may continue to bind the antigen and eliminate the pathogen even in the presence of antigenic mutations that abrogate binding to other antibody groups.

In general, the antibody and antibody group splits are more practically useful because they demonstrate the effectiveness of escape prediction for antibodies that have not undergone any experimental escape measurements. This means that new antibodies can be evaluated entirely *in silico*. Escape prediction models that are effective under these data splits could thus be used to guide the selection or design of antibodies that are robust to antigenic mutations that escape other antibodies, providing an avenue for designing effective new antibody treatments against mutating pathogens.

For all four splits, we train and test the models across all antibodies (cross-antibody setting) using five-fold cross-validation. For the mutation and site splits where each antibody can appear in both the train and test sets, we also build separate models for each of the 247 antibodies (per-antibody setting). This makes it possible to compare the ability of a single model learned across antibodies to separate models learned for each antibody individually.

3.3 Tasks and Metrics

For all of the models except for the likelihood model, which doesn't require training, we train the model either for a regression task, where escape scores are real values, or for a classification task, where escape scores are binarized into zero or non-zero escape. All models are evaluated with the metrics ROC-AUC (area under the receiver operating characteristic curve) and PRC-AUC (area under the precision recall curve), and regression models are additionally evaluated with the metrics MSE (mean squared error) and R^2 (coefficient of determination).

3.4 Models

Below we describe the implementation details of the models we developed. All models were built using PyTorch version 1.12.1 [17].

3.4.1 RNN

The RNN is a bidirectional LSTM [15] with a hidden dimensionality of 100. The input to the RNN is the mutated antigen sequence with amino acids encoded using trainable embeddings with a dimensionality of 100. The output of the RNN is an embedding for each amino acid with a dimensionality of 200 (100 for each direction of the RNN). Either the sequence average hidden embedding (RNN Seq) or the output embedding of the mutated amino acid (RNN Res) is used as input to a small multilayer perceptron (see below), which makes escape predictions. The amino acid embeddings, RNN, and multilayer perceptron are trained end-to-end.

3.4.2 Protein Language Model

For the likelihood and embedding models, we use the pretrained protein language model ESM2 [14]. We specifically use the `esm2_t33_650M_UR50D` version of the model consisting of 33 layers and 650M parameters that was trained on the UniRef50 database [18]. The embeddings produced by this model have a dimensionality of 1,280 and are used as fixed input to a small multilayer perceptron.

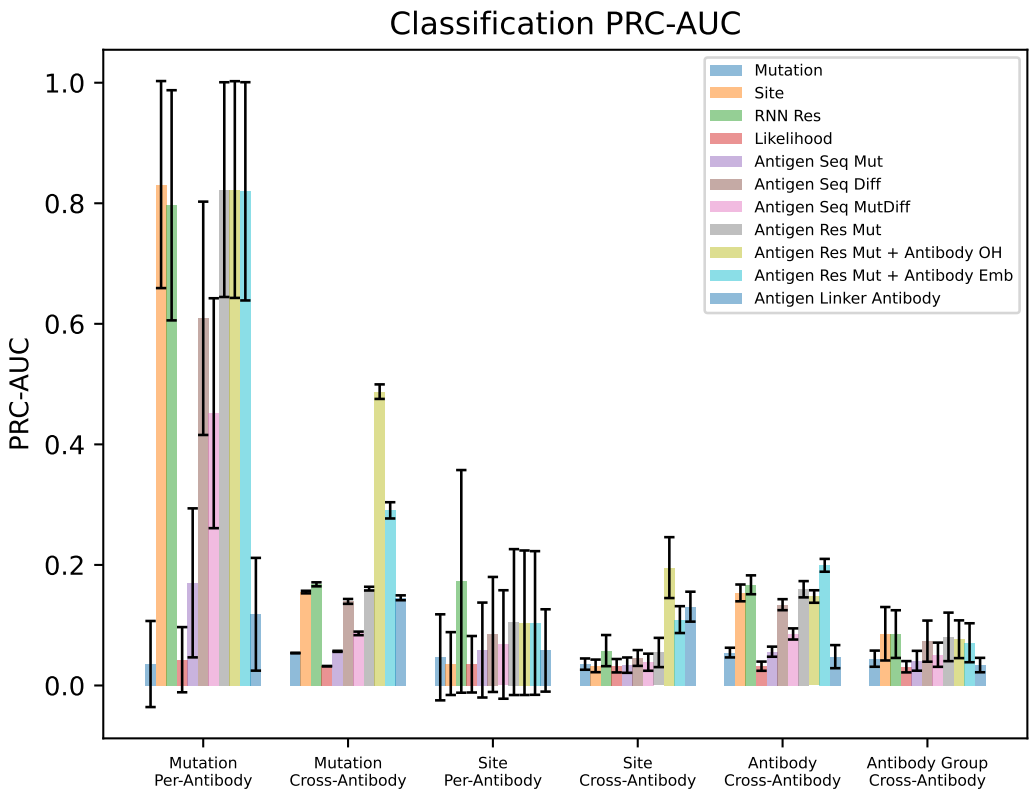

Figure 2: Classification model results with the PRC-AUC metric across data splits (x-axis) and models (color-coded bars). Error bars indicate the standard deviation across 247 antibodies for per-antibody splits and across five-fold cross-validation for cross-antibody splits.

3.4.3 Multilayer Perceptron

The multilayer perceptron (MLP) model that we use with the RNN and with all of the pretrained embeddings has two hidden layers with 100 neurons in each layer and ReLU activation followed by a single output. For classification tasks, we apply a sigmoid activation to the output.

3.4.4 Training

The RNN and the embedding models were trained with mean squared error loss for regression and binary cross entropy loss for classification using the Adam optimizer [19]. Per-antibody models were trained for 50 epochs while cross-antibody models were trained for one epoch. The RNN model was trained on a single GPU, with training taking about 3 minutes for a cross-antibody model (one fold) and about 30 seconds for each per-antibody model. The embedding models were trained on a single CPU, with training taking about 1 minute for a cross-antibody model (one fold) and about 15 seconds for each per-antibody model.

4 Results

In this section, we highlight some of the key results from our experiments (see Figure 2). We only show classification model results since the regression models performed poorly. Additionally, since the relative ranking of models was similar between ROC-AUC and PRC-AUC but the differences in PRC-AUC scores were more noticeable, we only present PRC-AUC results. We show results for all data splits and for a subset of the models, leaving out embedding models whose performance was not insightful for space. The complete set of results across all 162 experiments is in Appendix C.

4.1 Mutation Model

The mutation model is a very weak model. On the mutation and site splits in the per-antibody setting, the model has essentially no predictive power, and on all four splits in the cross-antibody setting, the model performs poorly. This is to be expected since the model ignores the mutation site even though the mutation site is very informative of immune escape due to the consistent interaction of key sites with neutralizing antibodies.

4.2 Site Model

The site model is strong across most splits, with the exception of the site split where the model has no information about unseen sites. The site model is frequently competitive with the best embedding models despite containing only 201 parameters instead of 650M parameters. The site model is significantly more effective in the per-antibody mutation split than in any of the cross-antibody splits since escape is highly consistent at a given antigen site for an antibody across amino acid mutations. Even so, the fact that the model retains some predictive power across antibodies and antibody groups indicates that patterns of escape at specific sites are conserved.

4.3 RNN

The RNN Res model performs comparatively well across all splits. It is only outperformed by the embedding models that include antibody embeddings, which is reasonable given that the RNN only processes the antigen sequence and has no knowledge of the antibody. Notably, the model performs on par with or better than most of the embedding models, even on the site split, despite training in just a couple of minutes with no expensive pretraining needed. This shows that existing pretraining methods and models may not be particularly beneficial for escape prediction, at least for this dataset. Interestingly, the RNN Seq model performs very poorly (see Appendix C), which may indicate that sequence averaging obscures the relevant information from the mutated amino acid. A similar phenomenon occurs in the Antigen Seq versus Antigen Res embeddings, meaning that it may be preferable to use residue rather than sequence averaged embeddings across model types.

4.4 Likelihood Model

The likelihood model has virtually no predictive power across all data splits. This is in contrast to examples in the literature where likelihoods achieve reasonable mutation effect prediction performance [7]. This finding demonstrates a fundamental limitation of the zero-shot prediction framework since likelihoods derived from models trained to recreate naturally occurring proteins may not be calibrated to predict the probability of antigen escape.

4.5 Embedding Models

The embedding models significantly outperform the likelihood model across all data splits. This indicates that PLMs do contain information that is useful for mutation effect prediction but require that their representations are adapted to the task rather than used in a zero-shot manner. However, the strength of the RNN model indicates that pretrained embeddings are not necessary for escape prediction. Even so, the embedding model results still provide several interesting takeaways regarding how best to use pretrained embeddings to predict escape in cases where they may be useful.

Mutant vs Difference. The Antigen Seq Diff embedding consistently outperforms the Antigen Seq Mut embedding, which indicates that the change in embedding from wildtype to mutant is more informative than the mutant embedding in isolation. The concatenation of the mutant and difference embeddings (MutDiff) does not improve performance further, indicating that the mutant embeddings do not contribute information beyond that contained in the difference embeddings.

Sequence vs Residue. The Antigen Res Mut embedding outperforms the Antigen Seq embeddings, perhaps because the sequence embeddings contain largely irrelevant information from the non-mutated residues. Interestingly, using embedding differences instead of mutant embeddings does not improve performance at the residue level (see Appendix C).

Antibody. Including antibody information alongside the antigen embeddings generally provides a benefit in all cross-antibody splits, where each model sees more than one antibody. The Antibody Res Mut + Antibody OH (one-hot) encodings provide a particularly large benefit in the mutation and site cross-antibody splits, where the same antibody can appear in train and test and the one-hot encoding makes it easy for the model to associate patterns of escape with particular antibodies. Interestingly, in the mutation splits, the one-hot antibody encoding does not allow the cross-antibody model to recover the performance of the per-antibody models, which may suggest a benefit to training separate models for each antibody, even though each model will be trained on less data.

Although the antibody embedding in the Antibody Res Mut + Antibody Emb model should also indicate the antibody identity, the model is not able to use this information as well as the one-hot embedding in the mutation and site cross-antibody splits. However, in the antibody and antibody group splits, no antibodies are shared between train and test and the one-hot encoding confers no benefit while the antibody embedding provides a small performance boost. Even so, this effect disappears in the harder antibody group split where no similar antibodies are present in the test set and the model does not learn how to extract useful information from the antibody embeddings.

The Antigen Linker Antibody embeddings do provide some benefit over antibody-agnostic models in the site cross-antibody split, but otherwise they do not help and sometimes hurt performance, as in the antibody split. This is likely because the PLM was not designed to use linkers and may not provide particularly useful embeddings for such artificially linked sequences.

5 Conclusion

We presented several methods for predicting immune escape using pretrained protein language model embeddings. We performed a comprehensive set of experiments on a SARS-CoV-2 deep mutational scanning dataset and showed that embeddings from PLMs are much more effective at predicting escape than zero-shot likelihoods. The Antigen Res Mut + Antibody Emb embeddings was particularly powerful among the embedding models, indicating that escape is best modeled at the residue level with both antigen and antibody embeddings. Although these results are promising, the relatively strong performance of the site model and the RNN, neither of which rely on computationally expensive pretraining, show that PLM embeddings may not be particularly beneficial for tasks such as escape prediction. Furthermore, the overall poor performance of all models across most splits demonstrates that significant future work is needed to make accurate and useful escape predictions. Notably, the results here are limited to a single antibody-antigen escape prediction task and dataset. The comprehensive nature of the data, which includes every possible single point mutation of the antigen for every antibody, gives our conclusions strength, but further experimentation on additional datasets is necessary to validate whether the conclusions drawn here generalize to other escape prediction tasks.

Acknowledgments

We would like to thank Mert Yuksekgonul, Mirac Suzgun, and Jeremy Wohlwend for their insightful comments, suggestions, and feedback. We would also like to thank the members of the Chang lab and the Zou lab for their helpful discussions. K.S. gratefully acknowledges the support of the Knight-Hennessy Scholarship.

References

1. Rai, K. R. *et al.* Acute Infection of Viral Pathogens and Their Innate Immune Escape. *Frontiers in Microbiology* **12.** ISSN: 1664-302X. https://www.frontiersin.org/articles/10.3389/fmicb.2021.672026 (2021).
2. Kapingidza, A. B., Kowal, K. & Chruszcz, M. in *Vertebrate and Invertebrate Respiratory Proteins, Lipoproteins and other Body Fluid Proteins* (eds Hoeger, U. & Harris, J. R.) 465–497 (Springer International Publishing, Cham, 2020). ISBN: 978-3-030-41769-7. https://doi.org/10.1007/978-3-030-41769-7_19.

3. Starr, T. N. *et al.* Prospective mapping of viral mutations that escape antibodies used to treat COVID-19. *Science* **371,** 850–854. eprint: https://www.science.org/doi/pdf/10.1126/science.abf9302. https://www.science.org/doi/abs/10.1126/science.abf9302 (2021).
4. Hie, B., Zhong, E. D., Berger, B. & Bryson, B. Learning the language of viral evolution and escape. *Science* **371,** 284–288. eprint: https://www.science.org/doi/pdf/10.1126/science.abd7331. https://www.science.org/doi/abs/10.1126/science.abd7331 (2021).
5. Cao, Y. *et al.* Omicron escapes the majority of existing SARS-CoV-2 neutralizing antibodies. *Nature* **602,** 657–663. ISSN: 1476-4687. https://doi.org/10.1038/s41586-021-04385-3 (Feb. 2022).
6. Frazer, J. *et al.* Disease variant prediction with deep generative models of evolutionary data. *Nature* **599,** 91–95. ISSN: 1476-4687. https://doi.org/10.1038/s41586-021-04043-8 (Nov. 2021).
7. Meier, J. *et al.* Language models enable zero-shot prediction of the effects of mutations on protein function. *bioRxiv.* eprint: https://www.biorxiv.org/content/early/2021/11/17/2021.07.09.450648.full.pdf. https://www.biorxiv.org/content/early/2021/11/17/2021.07.09.450648 (2021).
8. 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.* ISSN: 0092-8674. https://www.sciencedirect.com/science/article/pii/S0092867422011199 (2022).
9. Shan, S. *et al.* Deep learning guided optimization of human antibody against SARS-CoV-2 variants with broad neutralization. *Proceedings of the National Academy of Sciences* **119,** e2122954119. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.2122954119. https://www.pnas.org/doi/abs/10.1073/pnas.2122954119 (2022).
10. Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences* **118,** e2016239118. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.2016239118. https://www.pnas.org/doi/abs/10.1073/pnas.2016239118 (2021).
11. Nijkamp, E., Ruffolo, J., Weinstein, E. N., Naik, N. & Madani, A. *ProGen2: Exploring the Boundaries of Protein Language Models* 2022. https://arxiv.org/abs/2206.13517.
12. Riesselman, A. J., Ingraham, J. B. & Marks, D. S. Deep generative models of genetic variation capture the effects of mutations. *Nature Methods* **15,** 816–822. ISSN: 1548-7105. https://doi.org/10.1038/s41592-018-0138-4 (Oct. 2018).
13. Shin, J.-E. *et al.* Protein design and variant prediction using autoregressive generative models. *Nature Communications* **12,** 2403. ISSN: 2041-1723. https://doi.org/10.1038/s41467-021-22732-w (Apr. 2021).
14. Lin, Z. *et al.* Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv.* eprint: https://www.biorxiv.org/content/early/2022/07/21/2022.07.20.500902.full.pdf. https://www.biorxiv.org/content/early/2022/07/21/2022.07.20.500902 (2022).
15. Hochreiter, S. & Schmidhuber, J. Long Short-Term Memory. *Neural Computation* **9,** 1735–1780. ISSN: 0899-7667. eprint: https://direct.mit.edu/neco/article-pdf/9/8/1735/813796/neco.1997.9.8.1735.pdf. https://doi.org/10.1162/neco.1997.9.8.1735 (Nov. 1997).
16. Evans, R. *et al.* Protein complex prediction with AlphaFold-Multimer. *bioRxiv.* eprint: https://www.biorxiv.org/content/early/2022/03/10/2021.10.04.463034.full.pdf. https://www.biorxiv.org/content/early/2022/03/10/2021.10.04.463034 (2022).
17. Paszke, A. *et al.* in *Proceedings of the 33rd International Conference on Neural Information Processing Systems* (Curran Associates Inc., Red Hook, NY, USA, 2019).
18. Suzek, B. E. *et al.* UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31,** 926–932. ISSN: 1367-4803. eprint: https://academic.oup.com/bioinformatics/article-pdf/31/6/926/569379/btu739.pdf. https://doi.org/10.1093/bioinformatics/btu739 (Nov. 2014).

19. Kingma, D. P. & Ba, J. *Adam: A Method for Stochastic Optimization* in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings* (eds Bengio, Y. & LeCun, Y.) (2015). http://arxiv.org/abs/1412.6980.

A Data and Code Availability

The SARS-CoV-2 deep mutational scanning data from Cao et al. [5] is available at https://github.com/jbloomlab/SARS2_RBD_Ab_escape_maps/tree/main/data/2022_Cao_Omicron. The file data.csv contains the escape data for each antibody-antigen mutation combination, and the file antibodies.csv contains the sequences for the heavy and light chains for all the antibodies. The ESM2 pretrained protein language model [7] that we used is the esm2_t33_650M_UR50D model from https://github.com/facebookresearch/esm. Our code is available at https://github.com/swansonk14/escape_embeddings and all data, embeddings, and results are available at https://drive.google.com/drive/folders/18heVMWK46ExHkSeixyrNiJLovnIbZ4jg?usp=share_link.

B Data Visualization

Figures S.1, S.2, and S.3 visualize the SARS-CoV-2 deep mutational scanning data from Cao et al. [5].

C Complete Results

The remaining figures in the appendix section show the complete set of results for all 162 combinations of data splits, models, and tasks that we ran. These results are also available in tabular form in the results.csv file at https://drive.google.com/drive/folders/18heVMWK46ExHkSeixyrNiJLovnIbZ4jg?usp=share_link.

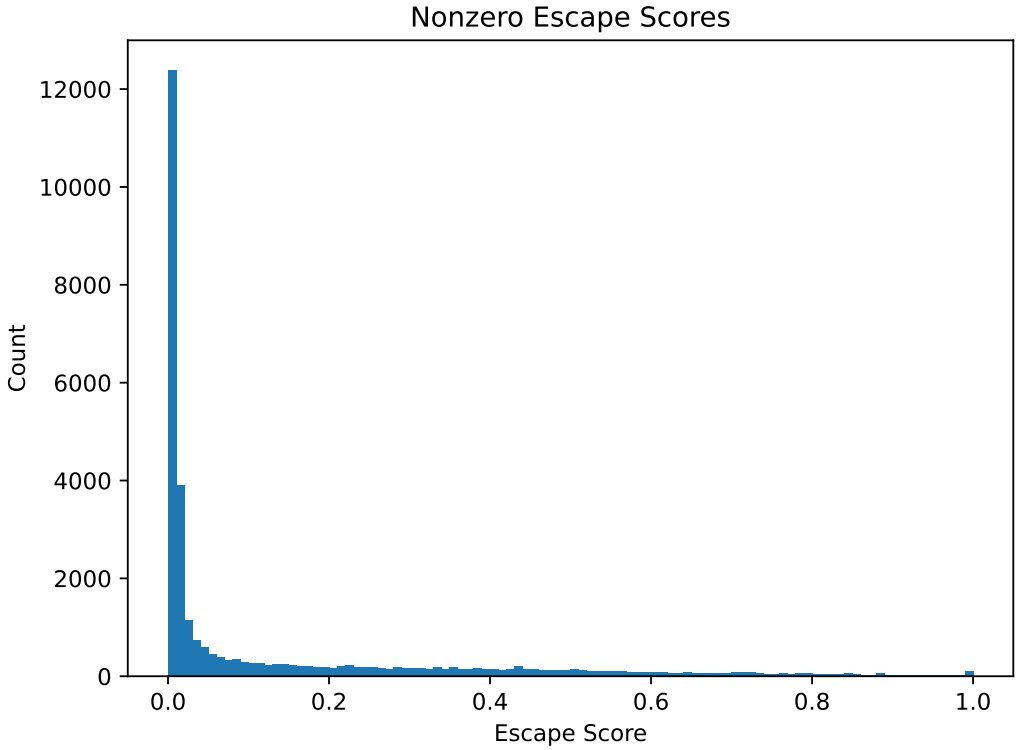

Figure S.1: The distribution of the 30,658 non-zero escape scores in the SARS-CoV-2 deep mutational scanning data from Cao et al. [5]. Note: The 74 escape scores greater than 1 (max 3.6) are truncated to 1.

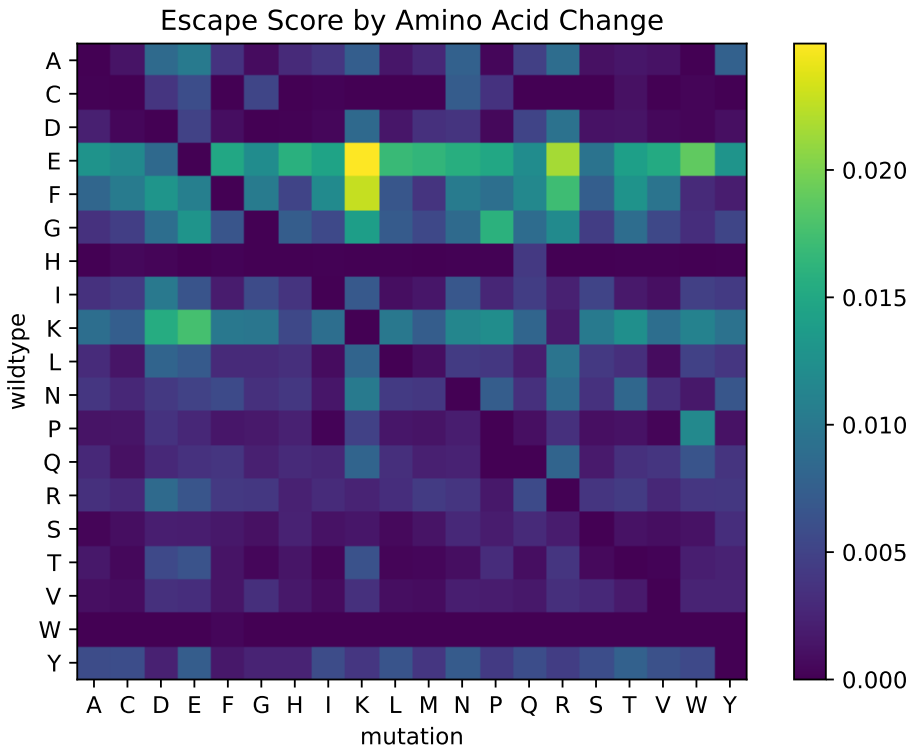

Figure S.2: The average escape score across all antibodies and all antigen sites for each wildtype to mutant amino acid change in the SARS-CoV-2 deep mutational scanning data from Cao et al. [5]. Note: The 74 escape scores greater than 1 (max 3.6) are truncated to 1.

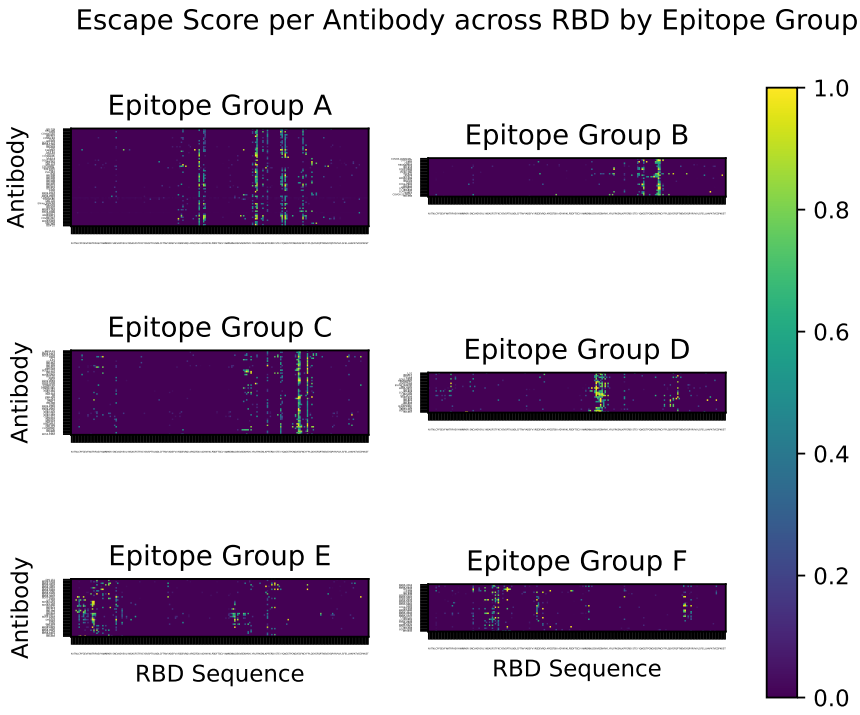

Figure S.3: The average escape score across all amino acid mutations for each antibody and each antigen site in the receptor binding domain (RBD), grouped according to the antibody clusters defined by Cao et al. [5]. Note: The 74 escape scores greater than 1 (max 3.6) are truncated to 1.

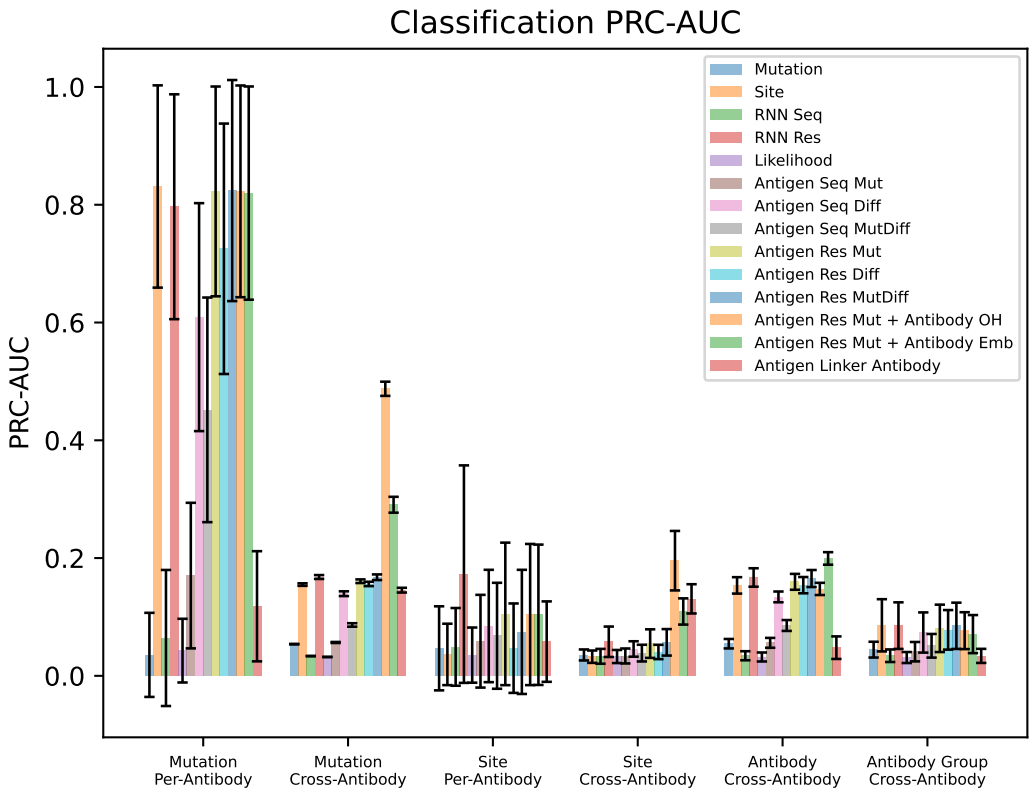

Figure S.4: Classification model results using the PRC-AUC metric.

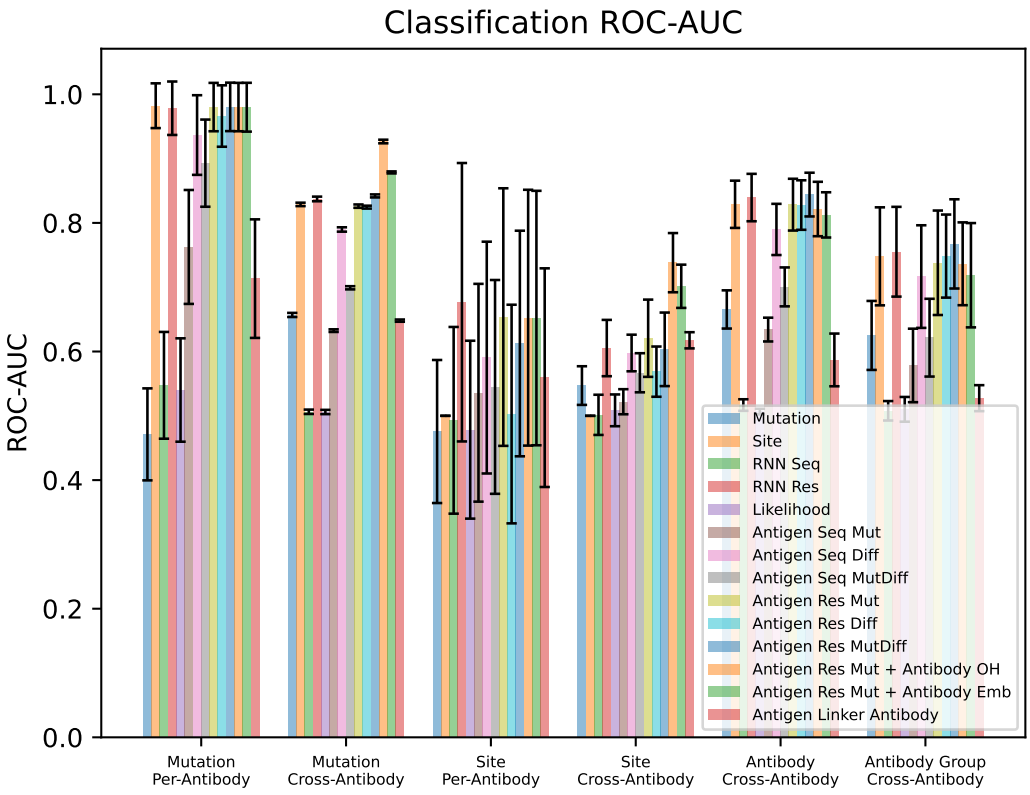

Figure S.5: Classification model results using the ROC-AUC metric.

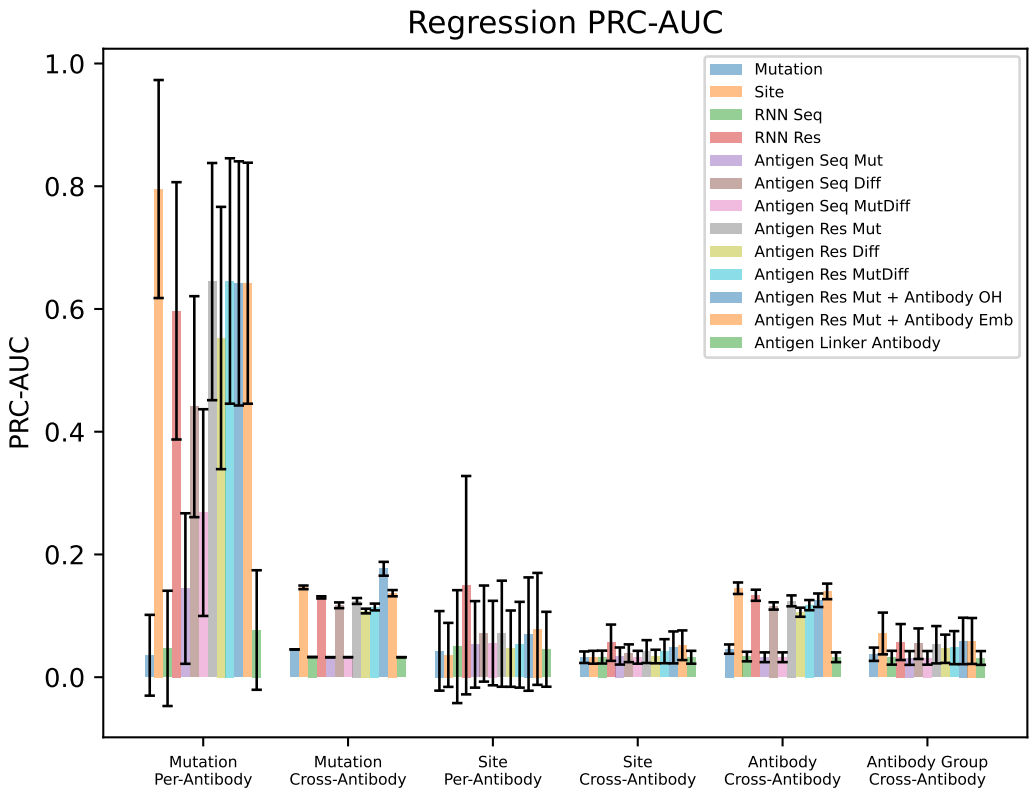

Figure S.6: Regression model results using the PRC-AUC metric.

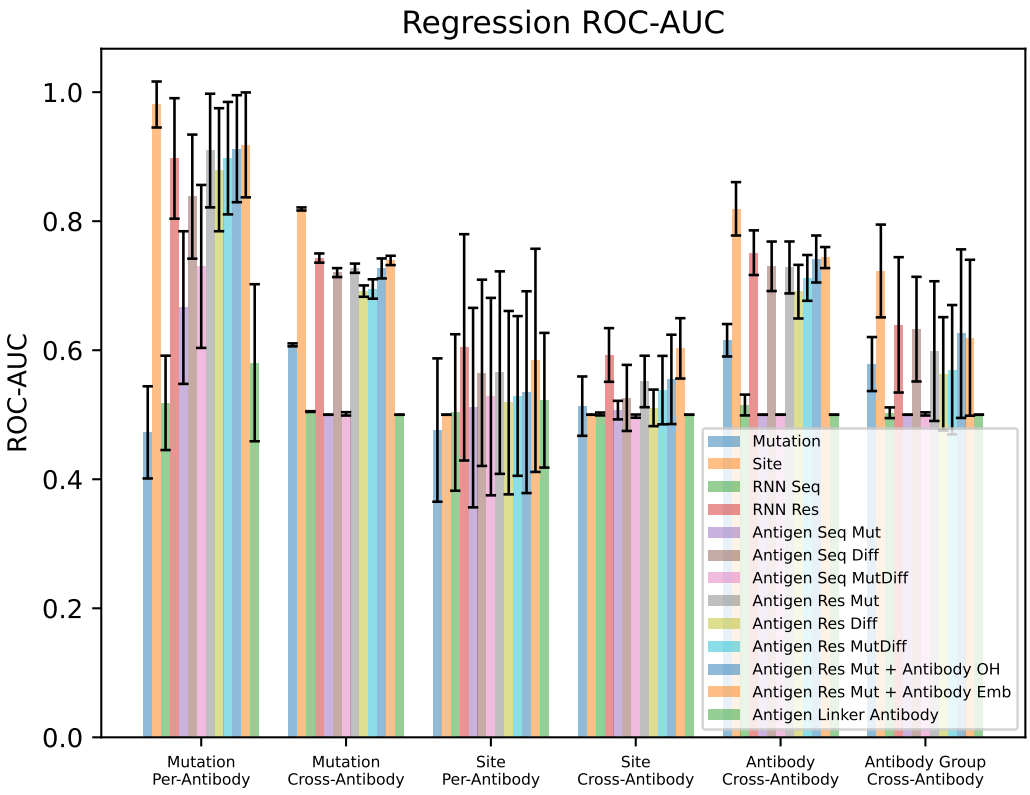

Figure S.7: Regression model results using the ROC-AUC metric.

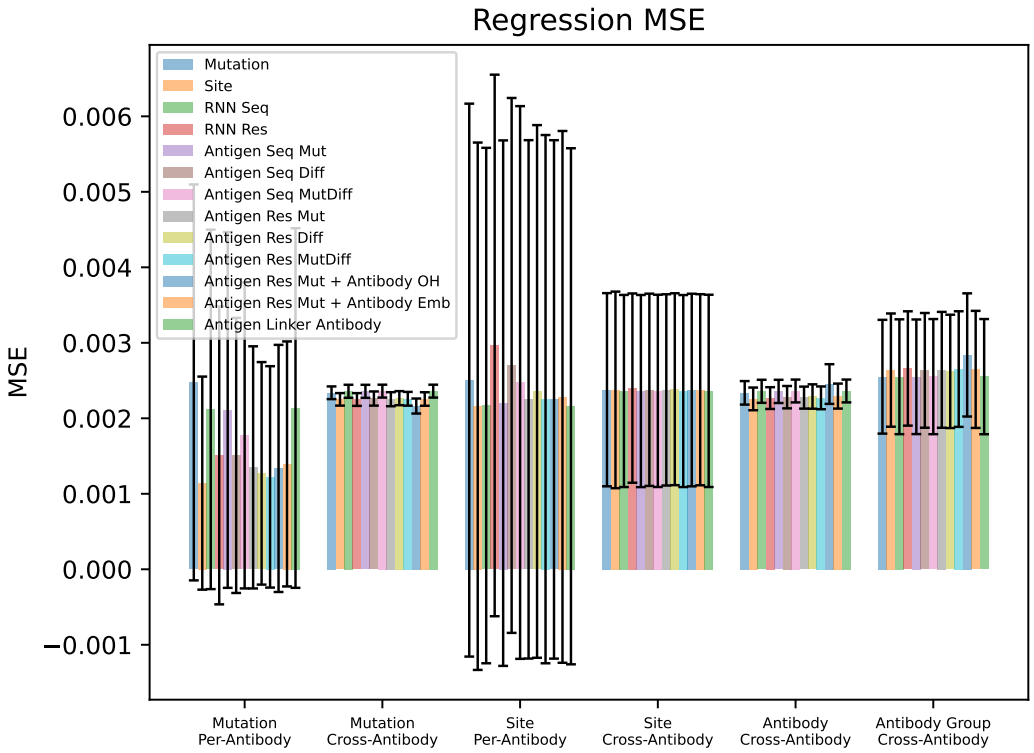

Figure S.8: Regression model results using the MSE metric.

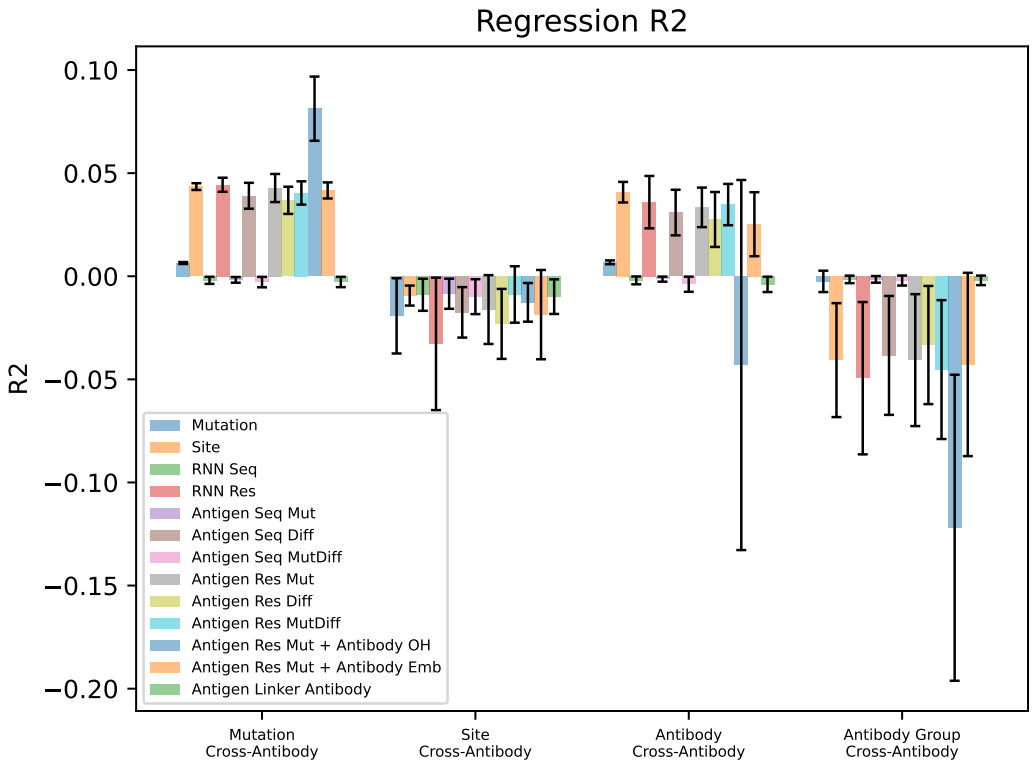

Figure S.9: Regression model results using the R^2 metric. Since the per-antibody splits have high variance, those splits are in separate plots below to make the scales legible in each plot.

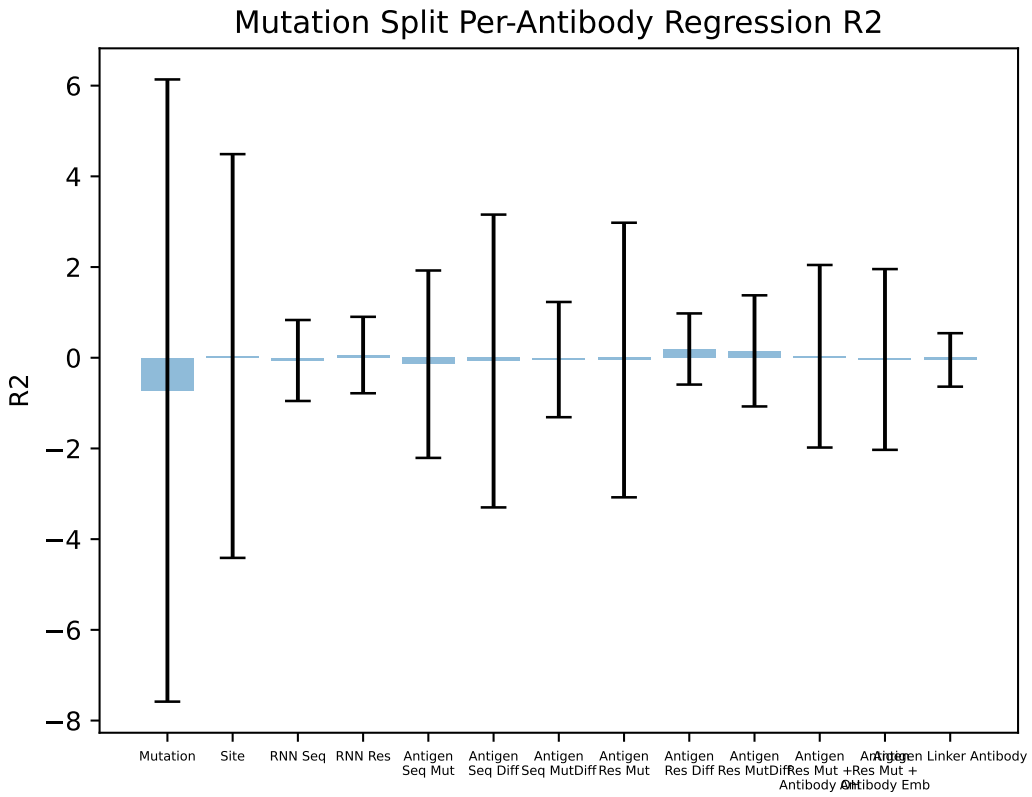

Figure S.10: Regression model results using the R^2 metric for the mutation per-antibody split.

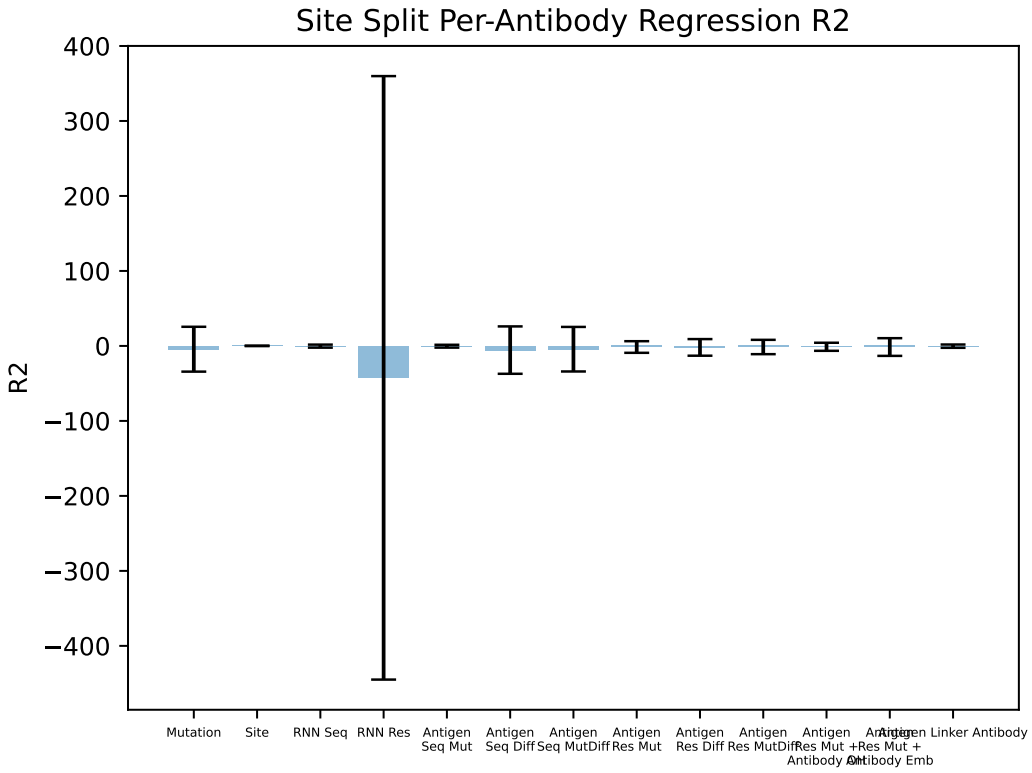

Figure S.11: Regression model results using the R^2 metric for the site per-antibody split.