Predicting Immune Escape with Pretrained Protein Language Model Embeddings

Kyle Swanson
Stanford University
swansonk@stanford.edu

Howard Chang*
Stanford University
howchang@stanford.edu

James Zou*
Stanford University
jamesz@stanford.edu

Abstract

Assessing the severity of new pathogenic variants requires an understanding of which mutations enable escape of the human immune response. Even single point mutations to an antigen can cause immune escape and infection by disrupting 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 baseline models 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 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

*Denotes co-senior author.

[1]Our code, data, embeddings, and results are available at `https://github.com/swansonk14/escape_embeddings`

Machine Learning for Structural Biology Workshop, NeurIPS 2022.

underlying statistics of naturally occurring protein sequences and can predict the likelihood that a given amino acid appears at a position in a protein. Prior work has shown that the relative likelihoods of a mutated and wildtype amino acid at a given position are predictive of the effect of that mutation in a zero-shot manner (i.e., without additional training) [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. 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 single point mutations on the binding ability of antibodies. The input to the model is the amino acid sequence of the wildtype antigen, the site on the antigen that is mutated, the new amino acid that replaces the wildtype amino acid at that site, and the amino acid sequences of the antibody's heavy and light chains. The output of the model is an escape score, which represents the degree to which antibody binding is reduced by the mutated antigen.

Mutation Model. The mutation baseline model computes the average escape score for each pair of `(wildtype, mutant)` amino acids and uses that average to predict escape for new antigens with the same amino acid mutation pair. This model ignores the antibody sequences and the entire antigen sequence except for the site that is mutated, and it assumes that escape depends solely on the identities of the wildtype and mutant amino acids. Since there are 20 amino acids, this model has $20 \times 20 = 400$ parameters. See Figure S.1 for a visualization of the model.

Site Model. The site baseline model computes the average escape score for each antigen site and uses that average to predict escape for new antigens with the same mutation site. Like the mutation baseline, this model ignores the antibody sequences and the entire antigen sequence except for the mutated site. Additionally, it ignores the identities of the wildtype and mutant amino acids. The model has one parameter for each site of the antigen. See Figure S.1 for a visualization of the model.

Likelihood Model. For our likelihood model, we adopt the zero-shot mutation prediction framework of Meier et al. [7]. In this framework, the antigen sequence is input to a pretrained protein language model model with the mutated site replaced by a mask token, and the escape score is predicted as the model's log odds ratio of the mutated amino acid versus the wildtype amino acid at that site. The likelihood model does not require any additional training and it does not incorporate the antibody sequences.

Embedding Models. Models that use pretrained protein language model embeddings 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, and some additionally use an embedding of the wildtype antigen and/or embeddings of the antibody heavy and light chains. The embedding variants are described below and are illustrated in Figure S.4.

Antigen Sequence Mutant. The language model is given the mutated antigen sequence and computes embeddings for each amino acid. Each embedding encodes the identity of the amino acid as well as its role in the context of the antigen sequence. The amino acid embeddings are averaged to form an embedding for the full antigen sequence. We refer to this embedding as Antigen Seq Mut.

Antigen Residue Mutant. As above, embeddings are computed for each amino acid in the mutated antigen. Here, the embedding of the mutated residue is used instead of the sequence average. We refer to this embedding as Antigen Res Mut.

Antigen Difference. Antigen embeddings for the mutated sequence and the wildtype sequence are computed, either both at the sequence level or both at the residue level. The difference between the embeddings (mutant minus wildtype) is computed. These embeddings are called Antigen Seq Diff and Antigen Res Diff for the difference of sequence or residue embeddings, respectively.

Antibody. The language model computes embeddings for the heavy and light chains of the antibody. For any of the antigen embeddings above, the antigen and antibody heavy and light chain embeddings are concatenated. We refer to these embeddings as the name of the antigen embedding + Antibody.

3 Experiments

Here, we describe the data, data splits, tasks, metrics, and models that we use in our experiments.

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 bind the original strain of SARS-CoV-2 by binding to the receptor binding domain (RBD) of the spike protein. The binding 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 binding compared to the wildtype antigen (see Figure S.1). Of the 943,293 escape scores in the dataset, 30,658 (3.2%) are non-zero, all in the range (0, 1] except for 74 outliers above 1 with a max of 3.6 (see Figure S.2). 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 on how the data is split. Below we describe the data splits we use.

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.

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.

Antibody. Antibodies are randomly split between train and test. This assumes that we do not know any escape scores for a new antibody, but that antibody may have a similar pattern of escape to antibodies in the train set.

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 no antibody in the train set has a similar pattern of escape to that antibody.

The latter two splits are more practically useful because they demonstrate the effectiveness of escape prediction for antibodies that have not undergone any experimental escape measurements. Models that are effective under these data splits could be used to guide antibody selection or design.

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, we also train and test separate models for each of the 247 antibodies (per-antibody setting) since each antibody can appear in train and test.

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 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 [15]. The embeddings produced by this model have a dimensionality of 1,280.

3.5 Multilayer Perceptron

For all the embedding models, we train a multilayer perceptron (MLP) to predict escape score from the embedding. The MLP has two hidden layers with 100 neurons in each layer and ReLU activation followed by a single linear output with sigmoid activation for classification. The model is trained with mean squared error loss for regression and binary cross entropy for classification using the Adam optimizer [16]. Per-antibody models were trained for 50 epochs while cross-antibody models were trained for one epoch. We implemented the MLP using PyTorch version 1.12.1 [17]. Due to the lightweight nature of the model, we were able to train each model in under an hour on a single CPU.

4 Results

In this section, we highlight some of the key results from our experiments (see Figure 1). 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 settings is in Appendix D.

Mutation Model. The mutation baseline 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 binding antibodies.

Site Model. The site baseline 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.

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.

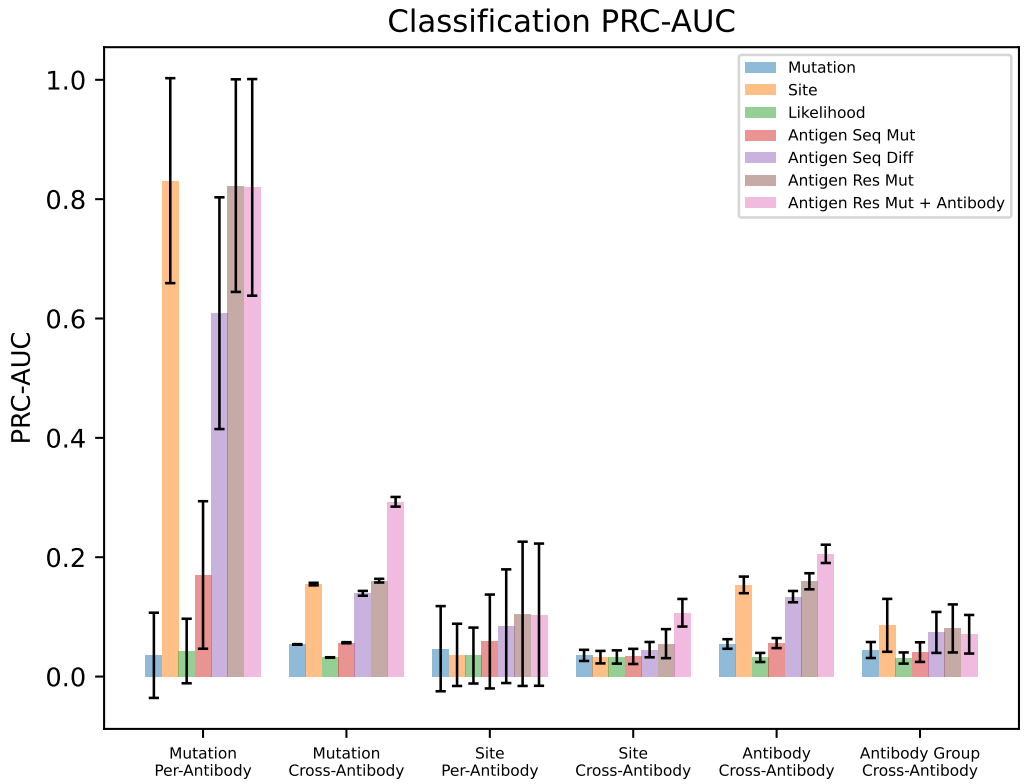

Figure 1: Classification model results with the PRC-AUC metric across data splits (x-axis) and models (color-coded bars). Error bars indicate the standard deviation across 247 antibodies for per-antibody splits and across five-fold cross-validation for cross-antibody splits.

Embedding Models. The embedding models generally match or exceed the performance of the two baseline models and the likelihood model across all data splits. This indicates that pretrained protein language models 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.

These results provide several interesting takeaways regarding how best to use pretrained embeddings to predict escape. 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 Antigen Res Mut embedding outperforms the Antigen Seq embeddings (Mut or Diff), perhaps because the sequence embeddings contain largely irrelevant information from the non-mutated residues. Interestingly, using embedding differences does not improve performance at the residue level (see Appendix D). As seen from the Antigen Res Mut + Antibody embedding, including antibody embeddings provides a benefit in all cross-antibody data splits except for the antibody group split, indicating that the antibody embedding is useful only in cases with the same or a similar antibody but not with a very different antibody.

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 these language models are much more effective at predicting escape than zero-shot likelihoods. Notably, the Antigen Res Mut + Antibody embeddings performed best, indicating that escape should be modeled at the residue level with both antigen and antibody embeddings. Although these results are promising, the relatively strong performance of the simple site baseline model and the overall poor performance of all models across most splits demonstrate that significant future work is needed to make accurate and useful escape predictions.

Acknowledgments

We would like to thank Mert Yuksekgonul, Mirac Suzgun, Jeremy Wohlwend, and Kirk Swanson 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. 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).
16. 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.
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).

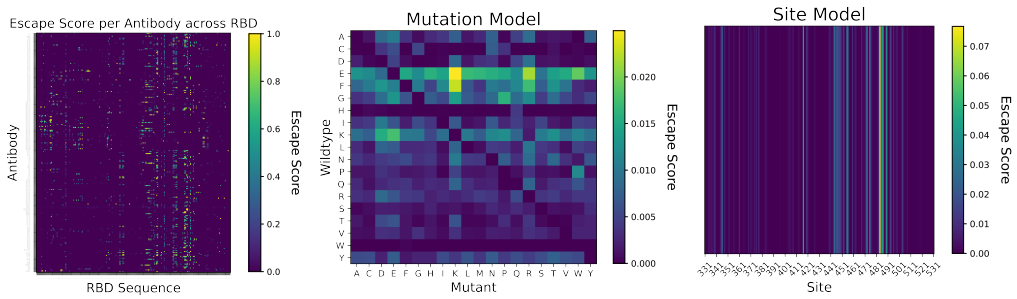

Figure S.1: SARS-CoV-2 immune escape data from Cao et al. [5] and associated statistical models. (Left) The average escape score across all amino acid mutations for each antibody and each antigen site in the receptor binding domain (RBD) of the SARS-CoV-2 spike protein. (Middle) A mutation model fit on the full dataset, showing the escape score for each wildtype to mutant amino acid change averaged across all antigen sites and antibodies. (Right) A site model fit on the full dataset, showing the escape score for each antigen site averaged across all amino acid mutations and antibodies. Note: In all figures, the 74 escape scores greater than 1 (max 3.6) are truncated to 1.

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, data, embeddings, and results are available at `https://github.com/swansonk14/escape_embeddings`.

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]. Figure S.1 shows the escape scores across antibodies and antigen sites along with the mutation and site models fitted on the whole dataset. Figure S.2 shows a histogram of the non-zero escape scores. Figure S.3 shows the escape scores across antibodies and antigen sites with antibodies divided into groups according to the escape-based clustering performed by Cao et al. [5].

C Embedding Model Figure

Figure S.4 shows an illustration of the embedding models used in this paper.

D Complete Results

The remaining figures in the appendix show the complete set of results for all combinations of data splits, models, and tasks that we ran. These results are also available in tabular form along with the data and embeddings at `https://github.com/swansonk14/escape_embeddings`.

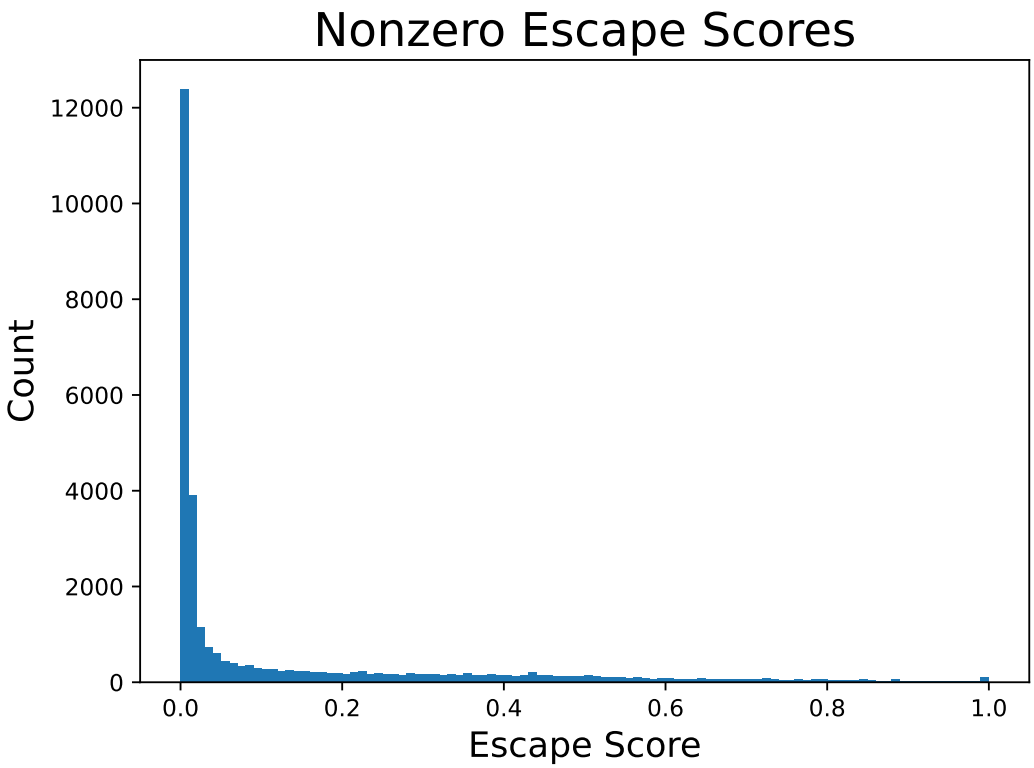

Figure S.2: The distribution of the 30,658 non-zero escape scores in the SARS-CoV-2 deep mutational scanning data from Cao et al. [5]. Note: The 74 escape scores greater than 1 (max 3.6) are truncated to 1.

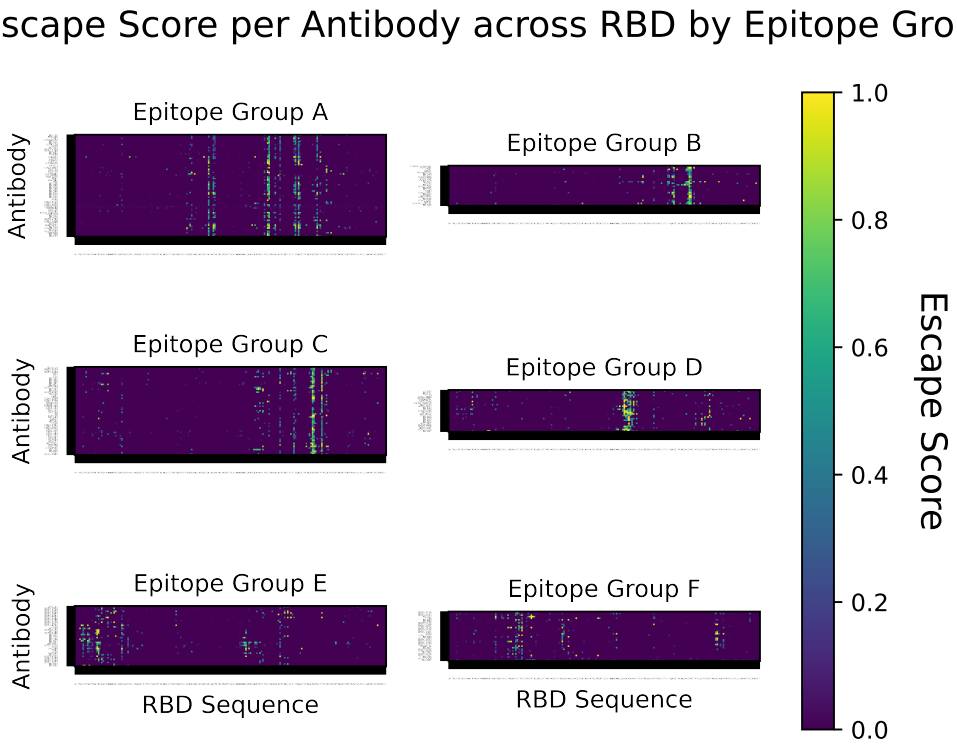

Figure S.3: The average escape score across all amino acid mutations for each antibody and each antigen site in the receptor binding domain (RBD) of the SARS-CoV-2 spike protein, grouped according to the antibody clusters defined by Cao et al. [5]. Note: The 74 escape scores greater than 1 (max 3.6) are truncated to 1.

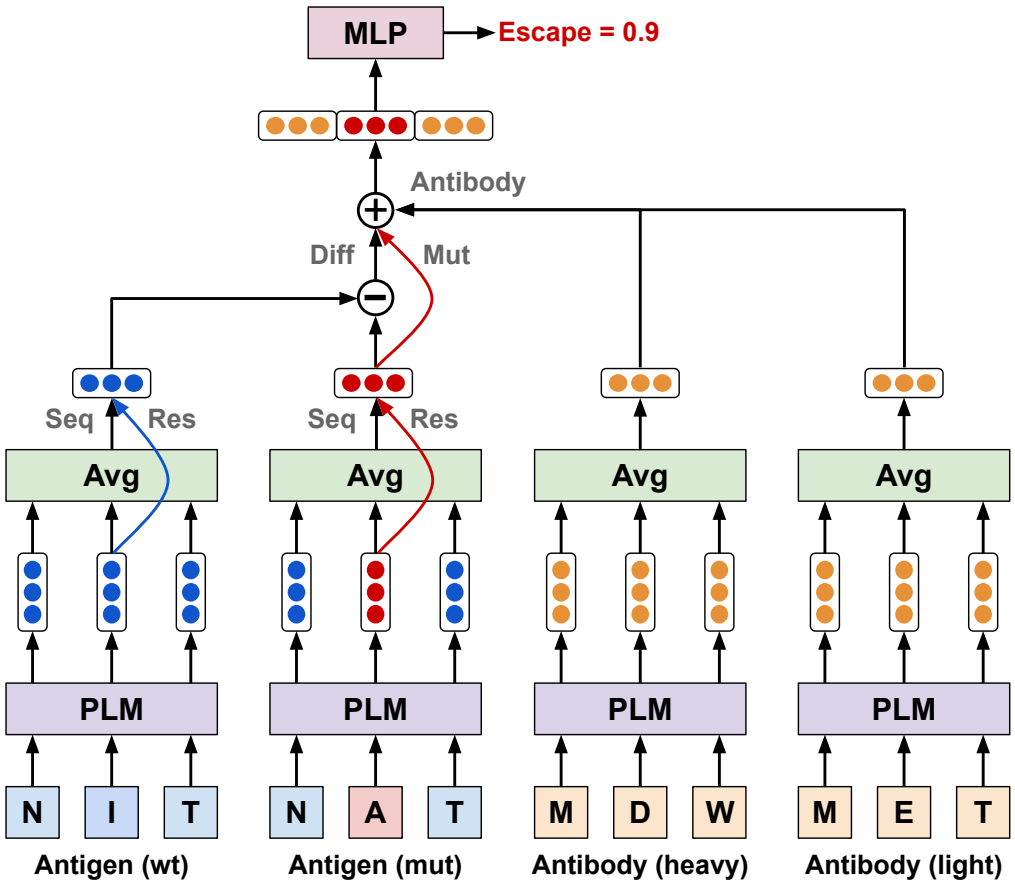

Figure S.4: An illustration of the various PLM embedding models for predicting immune escape. The different embedding types are described in detail in Section 2. Note that $\oplus$ indicates concatenation and $\ominus$ indicates elementwise difference.

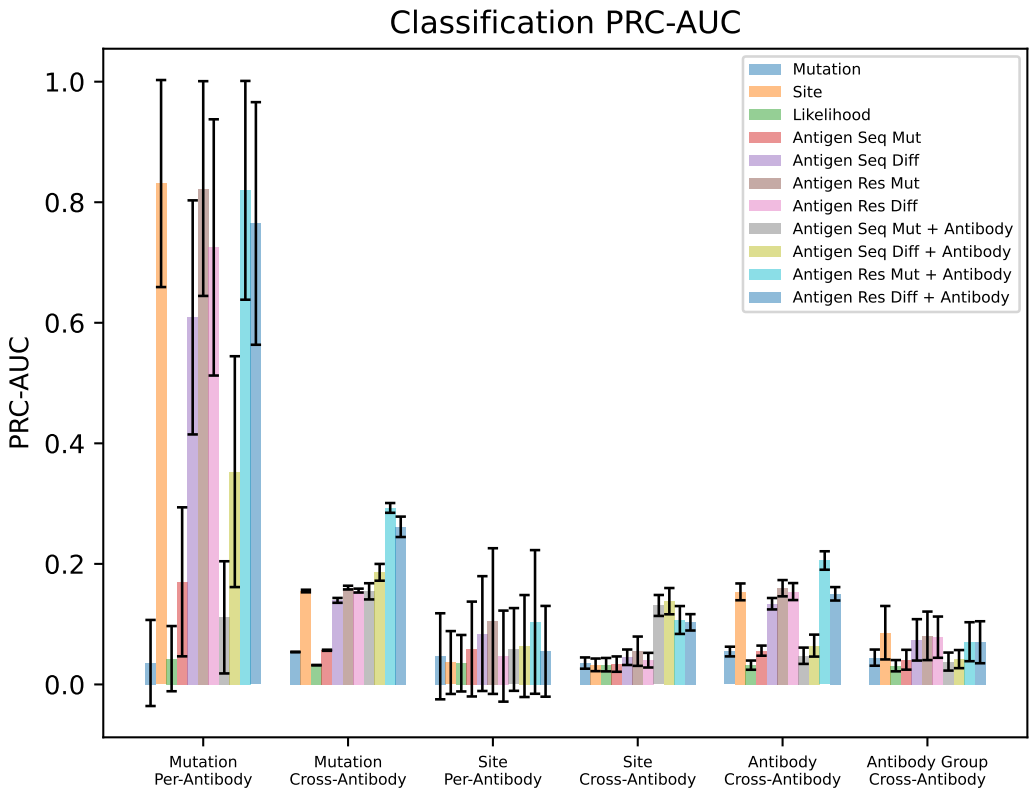

Figure S.5: Classification model results using the PRC-AUC metric.

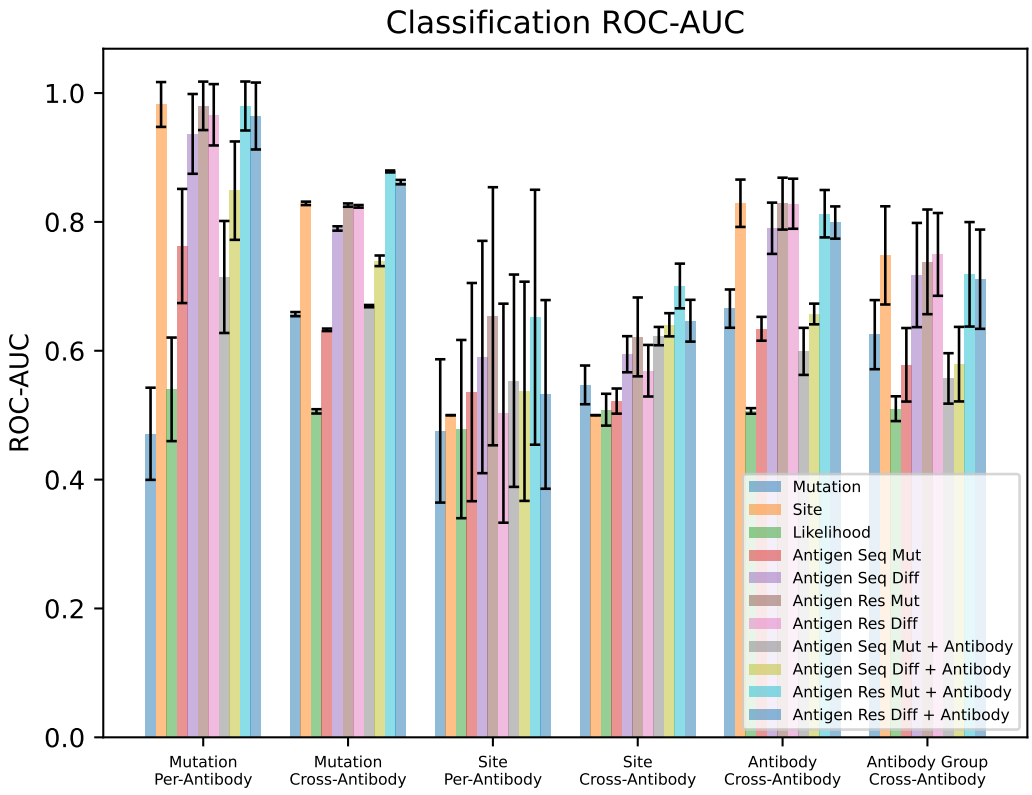

Figure S.6: Classification model results using the ROC-AUC metric.

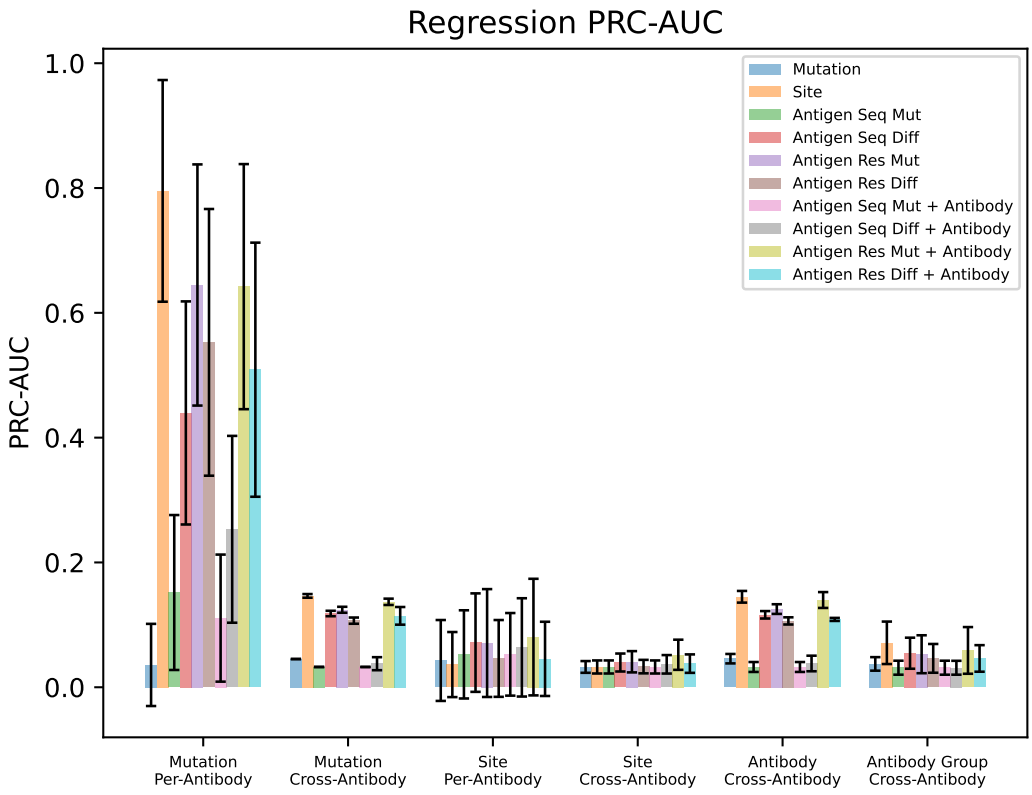

Figure S.7: Regression model results using the PRC-AUC metric.

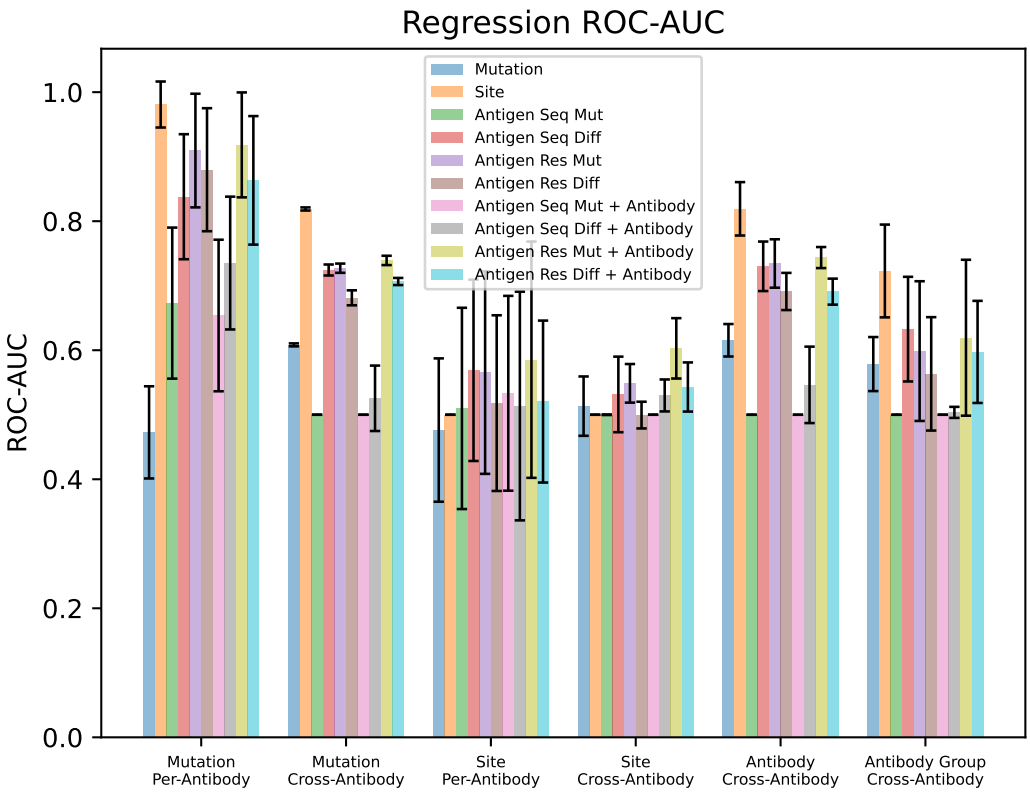

Figure S.8: Regression model results using the ROC-AUC metric.

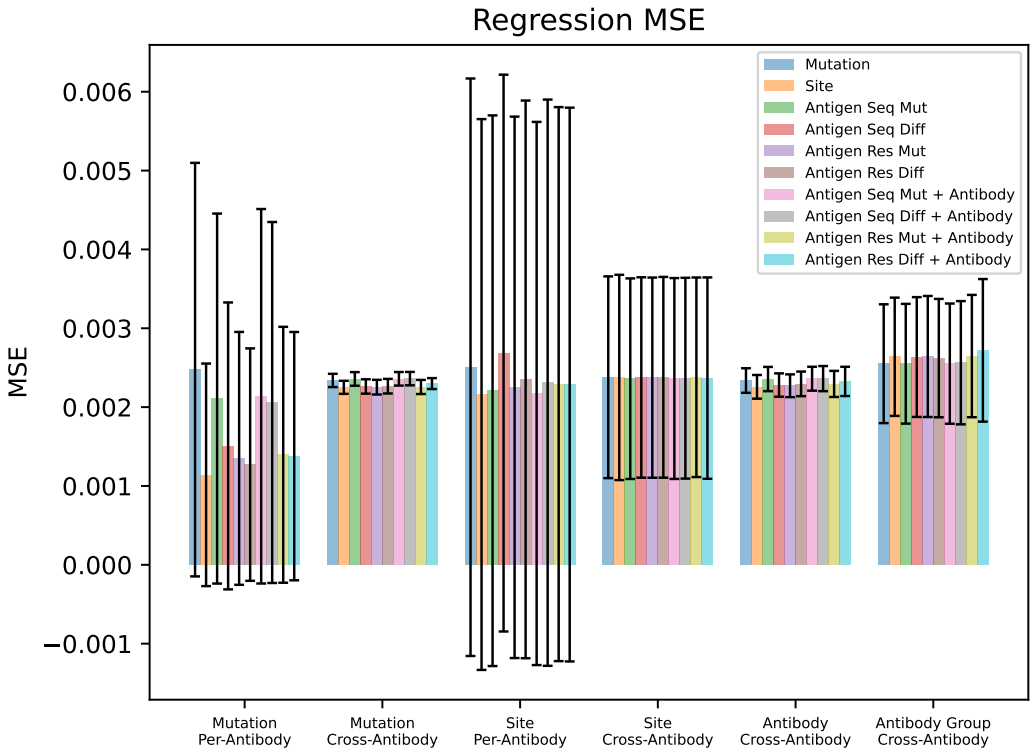

Figure S.9: Regression model results using the MSE metric.

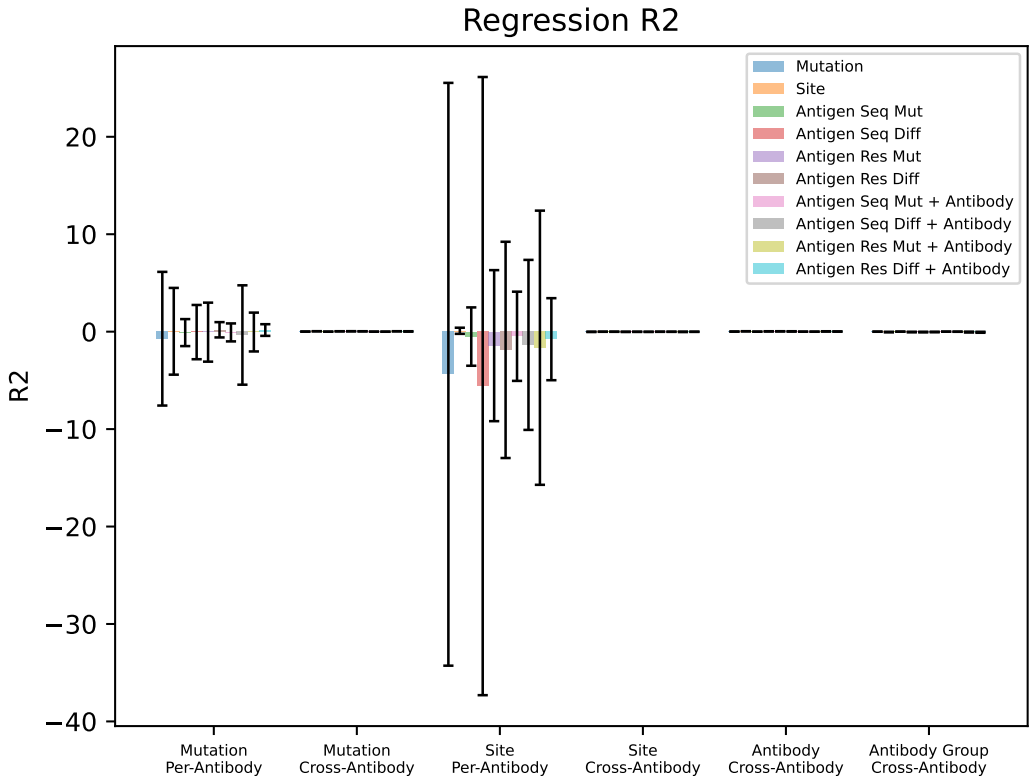

Figure S.10: Regression model results using the R^2 metric.