INTEGRATING PROTEIN LANGUAGE MODEL AND ACTIVE LEARNING FOR FEW-SHOT VIRAL VARIANT DETECTION

Marian Huot[1,2,⋆], **Dianzhuo Wang**[1,3,†,⋆], **Jiacheng Liu**[4], **Eugene I. Shakhnovich**[1,†]
[1]Department of Chemistry and Chemical Biology, Harvard University
[2]Laboratory of Physics of the Ecole Normale Supérieure, CNRS UMR 8023 and PSL Research
[3]John A. Paulson School of Engineering and Applied Sciences, Harvard University
[4]University of Washington, Seattle, WA
⋆Equal contribution
†johnwang@g.harvard.edu, shakhnovich@chemistry.harvard.edu

ABSTRACT

Early detection of high-fitness SARS-CoV-2 variants is crucial for pandemic response, yet limited experimental resources hinder timely identification. We propose an active learning framework that integrates a protein language model, a Gaussian process with uncertainty estimation, and a biophysical model to predict the fitness of novel receptor-binding domain (RBD) variants in a few-shot learning setting. Our approach prioritizes the most informative variants for experimental characterization, accelerating high-fitness variant detection by up to 5× compared to random sampling while testing fewer than 1% of all possible variants. Benchmarking on deep mutational scans, we show that our method identifies evolutionarily significant sites, particularly those facilitating antibody escape. We systematically compare different acquisition strategies and demonstrate that incorporating uncertainty-driven exploration enhances coverage of the mutational landscape, enabling the discovery of evolutionarily distant yet high-risk variants. Our results suggest that this framework could serve as an efficient early warning system for identifying concerning SARS-CoV-2 variants before they achieve widespread circulation.

1 INTRODUCTION

The relentless ascent in the number of SARS-CoV-2 infection cases has catalyzed an unprecedented protein evolution, yielding a profusion of novel mutations. Mutations in the receptor-binding domain (RBD) of the viral spike protein are particularly consequential, as they can increase binding affinity to the ACE2 receptor Ozono et al. (2021); Barton et al. (2021), facilitating more efficient host cell entry, or reduce susceptibility to neutralization by monoclonal antibodies (mAbs) and convalescent sera Tuekprakhon et al. (2022); Nabel et al. (2022); Wang et al. (2023). These adaptations enable the emergence of lineages with higher fitness Carabelli et al. (2023), which are better equipped to spread within populations or evade immune responses, thereby posing significant challenges to public health interventions and vaccine efficacy. As we navigate the aftermath of this crisis, it is imperative that we strengthen our preparedness for future pandemic threats.

Supervised machine learning has emerged as a valuable tool for predicting viral fitness (infectivity) and identifying high-risk viral mutations. These models leverage fitness labels collected from sequencing databases such as GISAID (Elbe & Buckland-Merrett, 2017a) or integrate laboratory measurements with epidemiological data. For example, Obermeyer et al. Obermeyer et al. (2022) introduced a non-epistatic approach that infers fitness from GISAID, allows estimation of fitness for combination of observed mutations. Ito et al. Ito et al. (2024) and Maher et al. Maher et al. (2022) integrate laboratory measurements of binding affinities with epidemiological data to train machine learning models that forecast the fitness of the SARS-CoV-2 variants and identify variants susceptible to widespread transmission. However, these models were rare in the early stages of the

pandemic due to the limited availability of labeled fitness data. Fitness labels are difficult to obtain unless the virus has been propagating in the population for a sufficient period.

Given these limitations, biophysical models offer a complementary approach by linking experimentally determined binding constants (K_D) to fitness as demonstrated by these worksWang et al. (2024); Chéron et al. (2016); Rotem et al. (2018). A key advantage of biophysical modeling is its ability to significantly reduce the functional space compared to machine learning, enabling accurate predictions with minimal experimental data while maintaining as a powerful fitness predictor. The underlying intuition is that high-fitness variants are usually variants that evade antibody binding while maintaining strong cellular receptor affinity. These experimental values can be obtained through low-throughput methods such as surface plasmon resonance Zhang et al. (2024) and isothermal titration calorimetryUpadhyay et al. (2023) or high-throughput methods such as deep mutational scanning (DMS) Starr et al. (2020) or combinatorial mutagenesis (CM) Moulana et al. (2023). However, the time and cost associated with comprehensive experimental measurements can limit its availability during the early stages of a pandemic.

Active learning addresses the challenge of limited labeled data by prioritizing the most promising variants for experimental characterization and has been successfully used for discovery of high fitness proteins Hie et al. (2020) as well as chemical space exploration Khalak et al. (2022); Graff et al. (2021). Active learning framework can efficiently be combined with Gaussian Process (GP), as they excel at handling scarce labeled data while providing uncertainty estimates Hie et al. (2020); Romero et al. (2013); Gessner et al. (2024).

This paper explores the integration of active learning and biophysical modeling to enhance early pandemic response capabilities, including potential variants of concern (pVOC) detection, and identifying sites under-pressure. We combine a biophysical model with a machine learning pipeline that integrates active learning with GP decoder on embeddings acquired from ESM3, a state-of-the-art protein language model. ESM3 generates structure-aware sequence embeddings, which the GP uses to smoothly predict how mutations affect binding affinities to cell receptors and antibodies. Then these binding affinities are piped into a pre-trained biophysical model Wang et al. (2024). This approach efficiently identifies potentially dangerous mutations by prioritizing the most informative variants based on acquisition functions.

We validated this pipeline using DMS, CM and GISAID sequencing data, demonstrating that our few-shot learning approach can effectively substitute for high-throughput screening in early pandemic surveillance.

2 METHODS

Datasets We utilize two distinct datasets in our research. The first is the combinatorial mutagenesis (CM) K_D measurements from the work of Moulana et al. (2022; 2023). In their study, they systematically examined the interactions between all possible combinations of 15 mutations in the RBD of BA.1 relative to the Wuhan Hu-1 strain(totaling 32,768 genotypes) and ACE2, as well as four monoclonal antibodies (LY-CoV016, LY-CoV555, REGN10987, and S309).

The second dataset we analyzed is a Deep Mutational Scanning (DMS) study by Starr et al. Starr et al. (2020; 2021a), which characterizes all possible single mutations in the RBD. This dataset provides ACE2 binding affinity (K_d values) and antibody escape ratios ($\epsilon(mut)$) for four therapeutic antibodies: LY-CoV016, LY-CoV555, REGN10987, and S309. (See SI for the conversion between $\epsilon(mut)$ and K_d). We focused our analysis on residues 334-526 of the RBD. Due to incomplete biophysical measurements, 28 sites were excluded from the dataset. To ensure unbiased estimation when computing AUC for comparison with GISAID data, we also removed these sites from the corresponding labels.

ESM ESM models are Transformer-based protein language models designed to extract meaningful representations from protein sequences, enabling tasks such as structure prediction and large-scale protein characterization.

We use these models to obtain semantic representation of protein sequences. For a protein sequence of length L, we first describe it as a sequence of tokens $\mathbf{x} \stackrel{\text{def}}{=} (x_1, \ldots, x_L)$. In the base of RBD,

$L = 201$. We then run a forward pass of ESM and obtain the hidden representations of the final layer, $(\mathbf{h}_1, \ldots, \mathbf{h}_L)$, where each $\mathbf{h}_i \in \mathbb{R}^K$. Then, we use mean pooling of these vectors to obtain a representation of the entire sequence, $\mathbf{z} = f_{\text{esm}}(\mathbf{x}) = \frac{1}{L}\sum_{i=1}^{L} \mathbf{h}_i \in \mathbb{R}^K$. Finally, we re normalized the embeddings by mean and variance.

We benchmarked different models: ESM1v Meier et al. (2021) (K=1280), ESM2 Lin et al. (2023) (K=1280) , ESM3 Hayes et al. (2024) sequence only, and ESM3 sequence with structure information from PDB 6XF5 Zhou et al. (2020) (K=1536). Main results are obtained using ESM3 with structure encoding.

Gaussian Process and Kernel Selection To model protein binding using Gaussian processes, we defined a kernel function that captures the notion of similarity between different variants in the sequence embedding space. We employed a Rational Quadratic kernel, which is well-suited for modeling functions with varying degrees of smoothness.

From a Bayesian perspective, the kernel defines the prior covariance between data points, ensuring that variants that are closer in embedding space ($\|z - z'\|^2$ small) exhibit similar binding values, while more distant variants remain weakly correlated. This reflects the fact that mutations with similar physicochemical properties and structural contexts tend to have correlated effects on protein binding.

When training the Gaussian process on a training set (Z_1, Y_1) of size N and making predictions for a new point (z_2, y_2), we used the analytical solutions for the posterior distribution:

$$P(y_2|Y_1, Z_1, z_2) \sim N(\mu_{\text{pred}}, \sigma^2_{\text{pred}})$$

with

$$\mu_{\text{pred}} = K(z_2, Z_1) \cdot [K(Z_1, Z_1)]^{-1} \cdot Y_1$$

and

$$\sigma^2_{\text{pred}} = k(z_2, z_2) - K(z_2, Z_1) \cdot [K(Z_1, Z_1)]^{-1} \cdot K(Z_1, z_2)$$

where $K(z_2, Z_1) = K(Z_1, z_2)^T$ is a vector of dimension N, $K(Z_1, Z_1)$ is the covariance matrix of dimension $N \times N$, and $k(z_2, z_2)$ is the kernel function evaluated at the test point.

In our specific application, z represents the embedding of a sequence, and $y(z) = \Delta \log K_D$ denotes the variation of log-transformed dissociation constant compared to wildtype. The kernel hyperparameters were optimized by maximizing the marginalized likelihood function using *sklearn.gaussianprocess* library.

Active Learning The Bayesian optimization approach used in this study incorporates active learning principles to efficiently explore a discrete set of candidate sequences, referred to as the "pool." This strategy selects and evaluates sequences in batches, progressively refining the optimization process. The iterative procedure consists of several key steps:

1. A random batch, denoted as S_0, is initially selected from the candidate set D. The labels for these points is calculated (e.g., experimental dissociation constants), forming the initial training dataset D_{train}. Unexplored dataset is then $D' \leftarrow D \backslash S_0$
2. A surrogate model $\hat{f}(x)$ is then trained on the dataset D_{train}, in order to predict fitness of unknown points in D'. Here, it first predicts dissociation constants, the latest being fed to a biophysical model which convert them into a fitness proxy. Each fitness prediction has value $\hat{\mu}(x)$ and variance $\hat{\sigma}^2(x)$.
3. An acquisition function $\alpha(x; \hat{f})$ is used to determine the utility of acquiring a given point x. This function considers various metrics, such as the predicted values for fitness, as well as the associated uncertainties.
4. The points with the highest utility, denoted as an ensemble X^*, are selected or "acquired". We obtain the ground truth value for their dissociation constants and add them to the training set ($D'_{train} \leftarrow D_{train} \cup \{X^*\}$) while removing it from the unexplored dataset ($D' \leftarrow D' \backslash \{X^*\}$).

Steps 2-4 are repeated iteratively until a stopping criterion is met, such as a fixed number of iterations or insufficient improvement.

In our study, we adopt following active learning parameters. For the DMS benchmark we chose initial training size of 174 mutations (randomly sampled, one mutation per RBD site) corresponding to $\sim 5\%$ of dataset and 10 more rounds acquiring 50 points each. For the Combinatorial benchmark, we chose initial training size of 20 variants (randomly sampled among single and double mutants) corresponding to $\sim 0.06\%$ of dataset and 10 more rounds acquiring 10 points each. Active learning acquisition was repeated 10 times for each benchmark, using different initial training sets sampled as described.

Acquisition Functions Various acquisition functions for active learning are considered in this study, each influencing the point selection process differently. These include:

- Random($x, \hat{f}$): Randomly selecting points from the candidate set.
- Greedy($x, \hat{f}$): Selecting points based on the surrogate model's mean ($\hat{\mu}(x)$).
- UCB(x): Employing the Upper Confidence Bound strategy, combining the mean and a scaled standard deviation ($\hat{\mu}(x) + \beta\hat{\sigma}(x)$).

To normalize the contribution of the variance term relative to the fitness values in UCB, we chose the coefficient β as: $\beta = 0.2 \times \frac{\mathrm{std(fitnesses)}}{\mathrm{std}\sqrt{\mathrm{vars}}}$. This ensures that the uncertainty term $\sqrt{\mathrm{vars}}$ is scaled to have a comparable range to the fitness values, adjusted by a scaling.

3 RESULTS

3.1 OVERVIEW OF VIRAL

We introduce VIRAL (Viral Identification via Rapid Active Learning), a framework that combines GP with active learning to predict the binding specificity of SARS-CoV-2 Receptor Binding Domain (RBD) variants to ACE2 and various antibodies (Figure 1a). The model operates in a structured pipeline: first, the RBD structure and mutant sequence are input into ESM3, a protein language model, to generate sequence embeddings that incorporate both sequence and structural context. These embeddings are then fed into a GP trained on a limited set of experimental dissociation constants to predict binding affinities. The predicted dissociation constants, as well as predicted uncertainties, are subsequently passed into into a biophysical model, following Wang et al. (2024), to infer the infectivity of each variant.

A key strength of our approach is its ability to rank variants effectively even with very few labeled data points, making it particularly powerful in early variant detection. Notably, including structural information from the wildtype RBD (PDB 6XF5 Zhou et al. (2020)) significantly enhances predictive accuracy across both deep mutational scanning and combinatorial mutation benchmarks (Figures 1b, 1c). In particular, our model achieves a surprising spearman coefficient of 0.53 on combinatorial dataset while being trained on only 20 points (0.06% of dataset). This advantage of structure-aware embeddings is consistent with Loux et al. (2024). The predictions from this pipeline guide an active learning strategy: variants with high predicted infectivity are selected for validation, and new measurements are used to iteratively retrain the GP, refining its accuracy over time. This iterative process enables the model to rapidly identify high-risk variants while minimizing the number of required measurements, making it a scalable approach for early warning systems in viral surveillance.

3.2 ACTIVE LEARNING IDENTIFIES TOP VARIANTS

Our objective is to maximize the identification of potential Variants of Concern (pVOC), defined as those ranking in the top p=10% across the mutational landscape. To rigorously evaluate our approach, we start with a retrospective study using existing Combinatorial Mutagenesis (CM) datasets containing experimentally measured binding constants (see Methods). This retrospective analysis is crucial as it provides the only systematic way to benchmark our model's performance: by simulating the sequential selection of variants from a completely characterized fitness landscape, we

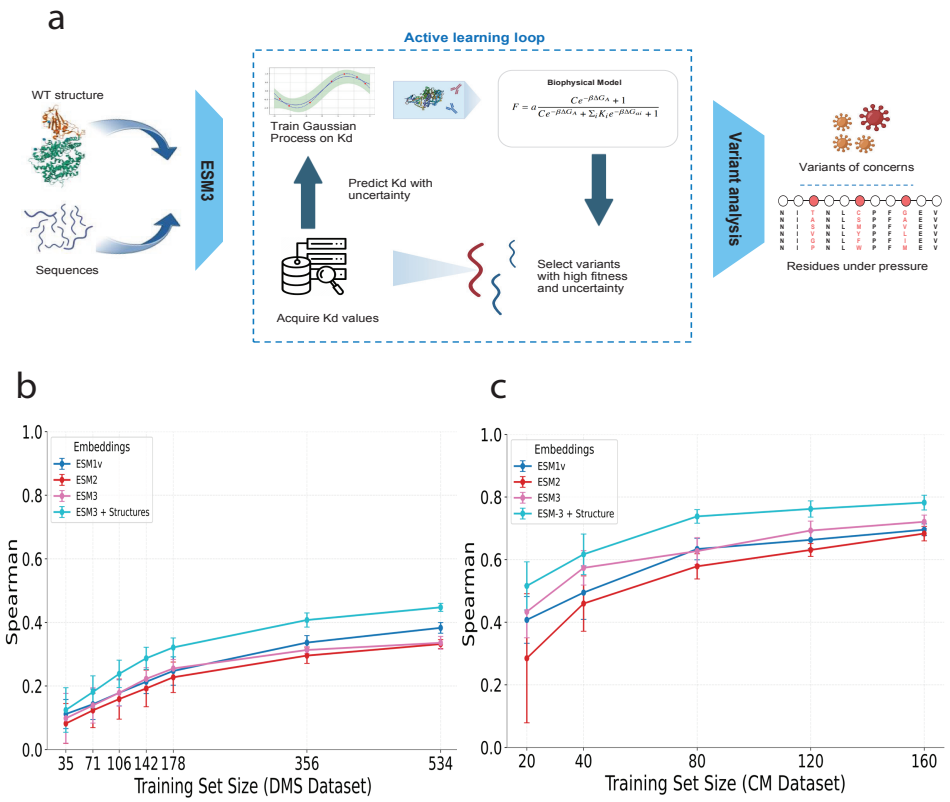


Figure 1: Overview of active learning framework for detecting high-fitness SARS-CoV-2 RBD variants. (a) The pipeline begins by using ESM3 and RBD structure to generate embeddings for RBD sequences. These embeddings serve as inputs to a combined Gaussian Process and biophysical model that predicts variant fitness. The framework operates in an iterative cycle where: (1) The model predicts fitness and uncertainty for untested variants, (2) Based on these predictions, the most promising variants are selected for experimental testing using various acquisition strategies (greedy, UCB, or random sampling), (3) An oracle representing experiments provides binding constants (Kd values), and (4) These new measurements are used to retrain the GP model, improving its predictive power. (b) Spearman correlation of predicted fitness for different training sized on DMS dataset. Error bars show std. Esm3_coord refers to ESM3 with wildtype structure. (c) Similar to (b) but on CM dataset.

can precisely quantify how efficiently our framework identifies high-fitness variants compared to alternative strategies. Such comprehensive validation would be impossible in a prospective setting, where the fitness of untested variants remains unknown.

Our main analysis focuses on two acquisition strategies: a greedy strategy that selects variants based solely on predicted fitness, and an Upper Confidence Bound (UCB) approach that combines predicted fitness with model uncertainty (see Method). While the greedy strategy excels at exploiting regions of known high fitness, UCB balances exploitation with exploration of uncertain regions in the sequence landscape, potentially uncovering novel fitness peaks. Random sampling constitutes the baseline, mimicking what we could achieve using brute force searching—screening every RBD in a library indiscriminately. We define the enrichment factor (EF) as the ratio of the percentage of top variants found by the model-guided search to the percentage of top variants found by a random search. An EF greater than 1 indicates superior performance compared to brute-force screening.

We use an initial training set of variants with a maximum number of 2 mutations, to model the first observed variants in a pandemic. Given that the most concerning variants, such as BA.1, can

accumulate up to 15 mutations in the Receptor Binding Domain (RBD), an effective early warning system must efficiently evaluate variants with increasingly complex mutation combinations to assess their potential for complete antibody escape. This translates into UCB acquisition metric achieving a final EF of 5 after 10 rounds of acquisition corresponding to a total of 120 points and 0.4% of the dataset (Figure 2b), while the greedy exploration struggles to beat random baseline: lack of exploration translates into lower capacity to identify dangerous variants (Figure 2a), quantified by Area Under the Receiver Operating Characteristic Curve (AUC).

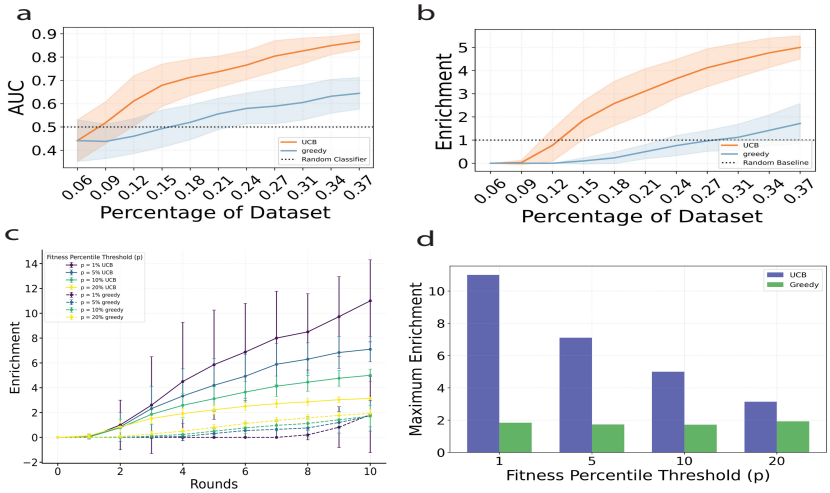


Figure 2: Active learning performance on combinatorial dataset. (a) Area Under the Curve (AUC), representing the model's ability to identify top fitness variants, is shown for different strategies (UCB and greedy). An AUC above 0.5 indicates effective identification of high-fitness variants. Each round corresponds to acquiring a new batch of variants, improving the predictor. Shaded regions represent standard deviations across runs. (b) Enrichment in top variants across acquisition runs for each strategy. Values above 1 represent improvement over a random acquisition. (c) Enrichment across acquisitions runs for different fitness thresholds p defining top variants. (d) Maximum enrichment obtained during active learning for different fitness thresholds p defining top variants.

The performance of the model improves when we increase the stringency of our definition for dangerous variants by decreasing the threshold p (Figures 2c and 2d). For instance, when defining dangerous variants as those in the top 1% of fitness scores rather than the top 10%, the model achieves a maximum enrichment factor of 11, demonstrating particularly strong performance in identifying the most concerning variants.

3.3 VIRAL IDENTIFIES HIGHLY MUTABLE SITES

Our framework successfully identifies key sites in the viral genome that emerged as mutation hotspots during natural evolution, with particular accuracy in detecting positions associated with antibody escape. We define a site as "highly mutable" when it has undergone substitutions to more than 9 of the 20 possible amino acids over the course of the pandemic. Through systematic evaluation of our active learning algorithm on the DMS dataset using multiple initial training sets, we achieved robust performance metrics (maximal enrichment of 3.1 and AUC of 0.81, see SI). Most notably, we discovered a strong correlation between sites frequently selected by our algorithm and those that exhibited high mutability during natural viral evolution.

Across 10 independent runs with different initial training sets, the AUC values ranged from 0.62 to 0.78. When acquisition scores were averaged across multiple runs to reduce sampling noise, we obtained a robust AUC of 0.76 (Figure 3a). These results remain consistent across different thresholds used to define "highly mutable" sites in natural evolution (see SI).

We bench-marked our methodology against alternative approaches to predict mutation-prone sites in Figure 3 b. Our integrated approach significantly outperforms baseline predictions derived from

variants observed before end of 2020, while performing slightly better than to predictions based on variants until end of 2022. This competitive performance is particularly noteworthy because baseline predictors have an inherent advantage—they are trained directly on pandemic-era mutations that successfully emerged in the population, while our method identifies relevant sites without this prior knowledge. Precisely, our pipeline enhances predictive capability by up to two years in identifying sites under pressure, at the cost of a limited exploration of the mutational landscape.

Figures 3c and 3d provide sequence and structural representation on the sites prioritized by our framework. The sampling frequency across the RBD sequence is visualized in Figure 3c, while Figure 3d maps the top 20 most frequently sampled sites onto the structure of the RBD-ACE2-S309 complex (PDB: 8FXCAddetia et al. (2023)). The spatial distribution of these prioritized sites reveals a immunological pattern: they cluster in critical regions, the ACE2 binding interface (which overlaps with class 1 and 2 antibody epitopes) and regions recognized by class 3 and 4 antibodies. Figures 3e–j further analyze the functional impact of mutations at these oversampled sites by plotting the change in log K_D ($\Delta \log K_D$) at each site relative to the wild type. The red dashed line represents the global average $\Delta \log K_D$ across the entire RBD. In Figures 3e–j, each of these sites is selected either for its ability to escape one or more antibodies while maintaining ACE2 binding, as highlighted in Figure 3e. In particular, despite being trained on only a few percent of single-variant data, our model successfully identified critical antibody escape mutations that later emerged in major SARS-CoV-2 variants. These key positions include residue 484 (enabling LY-CoV016 and LY-CoV555 escape) in Omicron BA.1; residue 493 (enabling CoV016 and CoV555 escape) in Omicron BA.1, BA.2, and BA.5; positions 340 (enabling S309 escape) and 356 (enabling S309 escape) in Omicron BA.2; residue 445 (enabling REGN10987 escape) in variants B.1, BF.8, and XD; and position 486 (enabling CoV016 and CoV555 escape) in numerous variants ranging from N.6 to Omicron BA.5 to BL1. We also identified position 507 as an oversampled site. Although mutations at this position allow escape from the LY-CoV016 and REGN10987 antibodies, they significantly compromise the binding affinity of ACE2. This trade-off between immune escape and receptor binding likely explains why mutations at position 507 have not been widely observed in naturally circulating variants.

4 DISCUSSION

In this study, we developed an integrated framework VIRAL for identifying high-fitness SARS-CoV-2 RBD variants and predicting mutation-prone sites using minimal experimental data. Our approach combines three key components: pLM for sequence representation, Gaussian processes for efficient learning and uncertainty quantification, and biophysical modeling for fitness prediction. This integration demonstrates several significant advantages in the context of viral surveillance and variant prediction.

First, leveraging pLM embeddings, Gaussian processes, and active learning framework, our model achieves high efficiency in identifying dangerous variants, obtaining up to 5-fold enrichment over random sampling while requiring experimental characterization of less than 1% of possible variants in the combinatorial mutagenesis landscape. This significant reduction in the number of required experiments could substantially accelerate the identification of concerning variants during early pandemic stages, when experimental resources are often limited. While initializing a model with some training data (as in this study) is advantageous, it is also feasible to start with zero training data, where zero-shot predictions initially carry equal uncertainty. Indeed, large language models have been proved to be efficient at zero-shot protein functions, provided they are trained on large and diverse protein sequence databases Meier et al. (2021). As more data are gathered, a sample-efficient model leveraging uncertainty can iteratively improve its predictions and confidence. This iterative cycle of computation and experimentation has been central to experiment prioritization, especially in drug discovery Eisenstein (2020).

Second, our framework shows high accuracy in identifying biologically relevant mutation sites and is interpretable as fitness prediction are driven by predicted antibody escape or good ACE2 binding. The systematic oversampling of positions that persisted in variants of concern, particularly at sites 356, 484, 486 and 493, demonstrates that our model could sample evolutionarily important sites using extremely limited data.

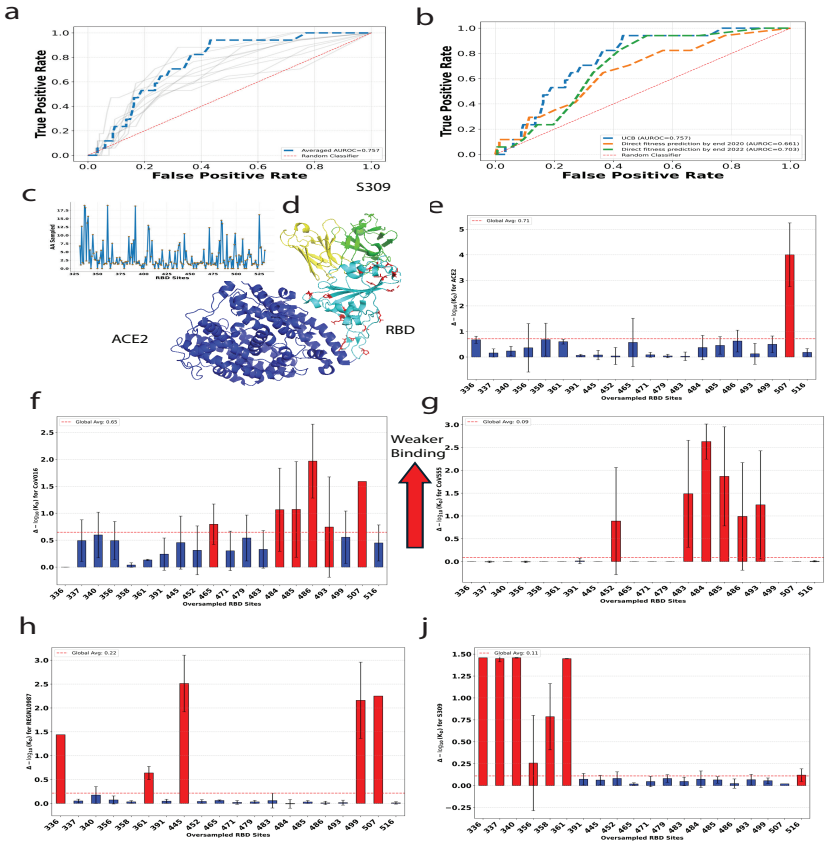


Figure 3: Active learning identifies residues under selection pressure. (a) ROC curves illustrating the performance of VIRAL in identifying frequently mutated sites in the GISAID database. Individual experimental runs are shown as grey lines. Blue curve is obtained when averaging the acquisition scores across multiple runs to reduce sampling noise. (b) Comparative ROC analysis of VIRAL benchmarked against one-batch acquisition using single-mutation variants observed by the end of 2020 (orange) or 2022 (green). (c) Number of sampled amino acid per site, plotted against the RBD sequence. (d) Structural representation of the RBD (light blue) in complex with ACE2 (dark blue) and the S309 antibody (green and yellow, representing the heavy and light chains, respectively; PDB: 8FXC). Red highlights indicate the top 20 sampled sites by VIRAL . (e-j) The KD distribution for oversampled sites identified by VIRAL is shown for ACE2, LY-Cov016, LY-CoV555, REGN10987, and S309. Higher values indicate binding loss. Red-highlighted sites exhibiting a greater average mutational binding loss than the protein-wide average.

Finally, a fundamental assumption underpinning our work is that our biophysical model accurately approximates real viral fitness. This assumption is critical because if incorrect, our active learning strategy might enrich for variants that don't represent biologically plausible evolutionary trajectories. However, there is strong precedent for biophysical modeling in viral fitness prediction. For instance, prior work by Chéron et al. (2016). and Rotem et al. (2018) demonstrated that fitness in RNA viruses can be quantitatively predicted from molecular properties. These successes in linking biophysical characteristics to viral fitness support our integrated approach for predicting SARS-CoV-2 variant fitness, which combines molecular binding measurements with epidemiological data to create a comprehensive fitness model.

CODE AVAILABILITY

The code used in this study will be made publicly available at `https://github.com/m-huot/VIRAL` upon manuscript acceptance. For access requests prior to publication, please contact the corresponding authors.

MEANINGFULNESS STATEMENT

A meaningful representation of life should capture how molecular-level changes in proteins propagate to system-level biological impacts. Our work demonstrates how integrating structural information with sequence embeddings enables accurate prediction of viral fitness from minimal experimental data. By combining machine learning with biophysical modeling in an active learning framework, we bridge multiple biological scales - from protein structure to viral epidemiology. This multiscale approach connects fundamental protein interactions to population-level outcomes, advancing both our theoretical understanding of viral evolution and providing practical value for pandemic preparedness through early identification of concerning variants.

REFERENCES

Amin Addetia, Luca Piccoli, James Brett Case, Young-Jun Park, Martina Beltramello, Barbara Guarino, Ha Dang, Guilherme Dias de Melo, Dora Pinto, Kaitlin Sprouse, et al. Neutralization, effector function and immune imprinting of omicron variants. *Nature*, 621(7979):592–601, 2023.

Michael I Barton, Stuart A MacGowan, Mikhail A Kutuzov, Omer Dushek, Geoffrey John Barton, and P Anton Van Der Merwe. Effects of common mutations in the SARS-CoV-2 Spike RBD and its ligand, the human ACE2 receptor on binding affinity and kinetics. *eLife*, 10:e70658, August 2021. ISSN 2050-084X. doi: 10.7554/eLife.70658. URL https://elifesciences.org/articles/70658.

Alessandro M. Carabelli, Thomas P. Peacock, Lucy G. Thorne, William T. Harvey, Joseph Hughes, COVID-19 Genomics UK Consortium, Thushan I. De Silva, Sharon J. Peacock, Wendy S. Barclay, Thushan I. De Silva, Greg J. Towers, and David L. Robertson. SARS-CoV-2 variant biology: immune escape, transmission and fitness. *Nature Reviews Microbiology*, January 2023. ISSN 1740-1526, 1740-1534. doi: 10.1038/s41579-022-00841-7. URL https://www.nature.com/articles/s41579-022-00841-7.

Nicolas Chéron, Adrian W R Serohijos, Jeong-Mo Choi, and Eugene I Shakhnovich. Evolutionary dynamics of viral escape under antibodies stress: A biophysical model. *Protein Sci.*, 25(7):1332–1340, July 2016.

Michael Eisenstein. Active machine learning helps drug hunters tackle biology. *Nature Biotechnology*, 38(5):512–514, May 2020. ISSN 1087-0156, 1546-1696. doi: 10.1038/s41587-020-0521-4. URL https://www.nature.com/articles/s41587-020-0521-4.

Stefan Elbe and Gemma Buckland-Merrett. Data, disease and diplomacy: Gisaid's innovative contribution to global health. *Global challenges*, 1(1):33–46, 2017a.

Stefan Elbe and Gemma Buckland-Merrett. Data, disease and diplomacy: GISAID's innovative contribution to global health: Data, Disease and Diplomacy. *Global Challenges*, 1(1):33–46, January 2017b. ISSN 20566646. doi: 10.1002/gch2.1018. URL https://onlinelibrary.wiley.com/doi/10.1002/gch2.1018.

Alexandra Gessner, Sebastian W. Ober, Owen Vickery, Dino Oglić, and Talip Uçar. Active learning for affinity prediction of antibodies, 2024. URL https://arxiv.org/abs/2406.07263.

David E. Graff, Eugene I. Shakhnovich, and Connor W. Coley. Accelerating high-throughput virtual screening through molecular pool-based active learning. *Chem. Sci.*, 12:7866–7881, 2021. doi: 10.1039/D0SC06805E. URL http://dx.doi.org/10.1039/D0SC06805E.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J. Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q. Tran, Jonathan Deaton, Marius Wiggert, Rohil Badkundri, Irhum Shafkat, Jun Gong, Alexander Derry, Raul S. Molina, Neil Thomas, Yousuf Khan, Chetan Mishra, Carolyn Kim, Liam J. Bartie, Matthew Nemeth, Patrick D. Hsu, Tom Sercu, Salvatore Candido, and Alexander Rives. Simulating 500 million years of evolution with a language model, July 2024. URL http://biorxiv.org/lookup/doi/10.1101/2024.07.01.600583.

Brian Hie, Bryan D. Bryson, and Bonnie Berger. Leveraging Uncertainty in Machine Learning Accelerates Biological Discovery and Design. *Cell Systems*, 11(5):461–477.e9, November 2020. ISSN 24054712. doi: 10.1016/j.cels.2020.09.007. URL https://linkinghub.elsevier.com/retrieve/pii/S2405471220303641.

Jumpei Ito, Adam Strange, Wei Liu, Gustav Joas, Spyros Lytras, The Genotype to Phenotype Japan (G2P-Japan) Consortium, and Kei Sato. A Protein Language Model for Exploring Viral Fitness Landscapes, March 2024. URL http://biorxiv.org/lookup/doi/10.1101/2024.03.15.584819.

K. Katoh and D. M. Standley. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, 30(4):772–780, April 2013. ISSN 0737-4038, 1537-1719. doi: 10.1093/molbev/mst010. URL https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/mst010.

Yuriy Khalak, Gary Tresadern, David F. Hahn, Bert L. De Groot, and Vytautas Gapsys. Chemical Space Exploration with Active Learning and Alchemical Free Energies. *Journal of Chemical Theory and Computation*, 18(10):6259–6270, October 2022. ISSN 1549-9618, 1549-9626. doi: 10.1021/acs.jctc.2c00752. URL https://pubs.acs.org/doi/10.1021/acs.jctc.2c00752.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574. URL https://www.science.org/doi/abs/10.1126/science.ade2574.

Thomas Loux, Dianzhuo Wang, and Eugene I. Shakhnovich. More structure, less accuracy: Esm3's binding prediction paradox. *bioRxiv*, 2024. doi: 10.1101/2024.12.09.627585. URL https://www.biorxiv.org/content/early/2024/12/09/2024.12.09.627585.

M. Cyrus Maher, Istvan Bartha, Steven Weaver, Julia Di Iulio, Elena Ferri, Leah Soriaga, Florian A. Lempp, Brian L. Hie, Bryan Bryson, Bonnie Berger, David L. Robertson, Gyorgy Snell, Davide Corti, Herbert W. Virgin, Sergei L. Kosakovsky Pond, and Amalio Telenti. Predicting the mutational drivers of future SARS-CoV-2 variants of concern. *Science Translational Medicine*, 14(633):eabk3445, February 2022. ISSN 1946-6234, 1946-6242. doi: 10.1126/scitranslmed.abk3445. URL https://www.science.org/doi/10.1126/scitranslmed.abk3445.

Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alexander Rives. Language models enable zero-shot prediction of the effects of mutations on protein function, July 2021. URL http://biorxiv.org/lookup/doi/10.1101/2021.07.09.450648.

Alief Moulana, Thomas Dupic, Angela M Phillips, Jeffrey Chang, Serafina Nieves, Anne A Roffler, Allison J Greaney, Tyler N Starr, Jesse D Bloom, and Michael M Desai. Compensatory epistasis maintains ACE2 affinity in SARS-CoV-2 omicron BA.1. *Nat. Commun.*, 13(1):7011, November 2022.

Alief Moulana, Thomas Dupic, Angela M Phillips, Jeffrey Chang, Anne A Roffler, Allison J Greaney, Tyler N Starr, Jesse D Bloom, and Michael M Desai. The landscape of antibody binding affinity in SARS-CoV-2 omicron BA.1 evolution. *Elife*, 12, February 2023.

Katherine G. Nabel, Sarah A. Clark, Sundaresh Shankar, Junhua Pan, Lars E. Clark, Pan Yang, Adrian Coscia, Lindsay G. A. McKay, Haley H. Varnum, Vesna Brusic, Nicole V. Tolan, Guohai Zhou, Michaël Desjardins, Sarah E. Turbett, Sanjat Kanjilal, Amy C. Sherman, Anand Dighe, Regina C. LaRocque, Edward T. Ryan, Casey Tylek, Joel F. Cohen-Solal, Anhdao T. Darcy, Davide Tavella, Anca Clabbers, Yao Fan, Anthony Griffiths, Ivan R. Correia, Jane Seagal, Lindsey R. Baden, Richelle C. Charles, and Jonathan Abraham. Structural basis for continued antibody evasion by the SARS-CoV-2 receptor binding domain. *Science*, 375(6578):eabl6251, January 2022. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.abl6251. URL https://www.science.org/doi/10.1126/science.abl6251.

Fritz Obermeyer, Martin Jankowiak, Nikolaos Barkas, Stephen F Schaffner, Jesse D Pyle, Leonid Yurkovetskiy, Matteo Bosso, Daniel J Park, Mehrtash Babadi, Bronwyn L MacInnis, et al. Analysis of 6.4 million sars-cov-2 genomes identifies mutations associated with fitness. *Science*, 376 (6599):1327–1332, 2022.

Seiya Ozono, Yanzhao Zhang, Hirotaka Ode, Kaori Sano, Toong Seng Tan, Kazuo Imai, Kazuyasu Miyoshi, Satoshi Kishigami, Takamasa Ueno, Yasumasa Iwatani, Tadaki Suzuki, and Kenzo Tokunaga. SARS-CoV-2 D614G spike mutation increases entry efficiency with enhanced ACE2-binding affinity. *Nature Communications*, 12(1):848, February 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-21118-2. URL https://www.nature.com/articles/s41467-021-21118-2.

Philip A. Romero, Andreas Krause, and Frances H. Arnold. Navigating the protein fitness landscape with Gaussian processes. *Proceedings of the National Academy of Sciences*, 110(3), January

2013. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1215251110. URL https://pnas.org/doi/full/10.1073/pnas.1215251110.

Assaf Rotem, Adrian W R Serohijos, Connie B Chang, Joshua T Wolfe, Audrey E Fischer, Thomas S Mehoke, Huidan Zhang, Ye Tao, W Lloyd Ung, Jeong-Mo Choi, João V Rodrigues, Abimbola O Kolawole, Stephan A Koehler, Susan Wu, Peter M Thielen, Naiwen Cui, Plamen A Demirev, Nicholas S Giacobbi, Timothy R Julian, Kellogg Schwab, Jeffrey S Lin, Thomas J Smith, James M Pipas, Christiane E Wobus, Andrew B Feldman, David A Weitz, and Eugene I Shakhnovich. Evolution on the biophysical fitness landscape of an RNA virus. *Mol. Biol. Evol.*, 35(10):2390–2400, October 2018.

Tyler N. Starr, Allison J. Greaney, Sarah K. Hilton, Daniel Ellis, Katharine H.D. Crawford, Adam S. Dingens, Mary Jane Navarro, John E. Bowen, M. Alejandra Tortorici, Alexandra C. Walls, Neil P. King, David Veesler, and Jesse D. Bloom. Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell*, 182 (5):1295–1310.e20, September 2020. ISSN 00928674. doi: 10.1016/j.cell.2020.08.012. URL https://linkinghub.elsevier.com/retrieve/pii/S0092867420310035.

Tyler N. Starr, Nadine Czudnochowski, Zhuoming Liu, Fabrizia Zatta, Young-Jun Park, Amin Addetia, Dora Pinto, Martina Beltramello, Patrick Hernandez, Allison J. Greaney, Roberta Marzi, William G. Glass, Ivy Zhang, Adam S. Dingens, John E. Bowen, M. Alejandra Tortorici, Alexandra C. Walls, Jason A. Wojcechowskyj, Anna De Marco, Laura E. Rosen, Jiayi Zhou, Martin Montiel-Ruiz, Hannah Kaiser, Josh R. Dillen, Heather Tucker, Jessica Bassi, Chiara Silacci-Fregni, Michael P. Housley, Julia Di Iulio, Gloria Lombardo, Maria Agostini, Nicole Sprugasci, Katja Culap, Stefano Jaconi, Marcel Meury, Exequiel Dellota Jr, Rana Abdelnabi, Shi-Yan Caroline Foo, Elisabetta Cameroni, Spencer Stumpf, Tristan I. Croll, Jay C. Nix, Colin Havenar-Daughton, Luca Piccoli, Fabio Benigni, Johan Neyts, Amalio Telenti, Florian A. Lempp, Matteo S. Pizzuto, John D. Chodera, Christy M. Hebner, Herbert W. Virgin, Sean P. J. Whelan, David Veesler, Davide Corti, Jesse D. Bloom, and Gyorgy Snell. SARS-CoV-2 RBD antibodies that maximize breadth and resistance to escape. *Nature*, 597(7874):97–102, September 2021a. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-021-03807-6. URL https://www.nature.com/articles/s41586-021-03807-6.

Tyler N. Starr, Allison J. Greaney, Amin Addetia, William W. Hannon, Manish C. Choudhary, Adam S. Dingens, Jonathan Z. Li, and Jesse D. Bloom. Prospective mapping of viral mutations that escape antibodies used to treat COVID-19. *Science*, 371(6531):850–854, February 2021b. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.abf9302. URL https://www.science.org/doi/10.1126/science.abf9302.

Aekkachai Tuekprakhon, Rungtiwa Nutalai, Aiste Dijokaite-Guraliuc, Daming Zhou, Helen M. Ginn, Muneeswaran Selvaraj, Chang Liu, Alexander J. Mentzer, Piyada Supasa, Helen M.E. Duyvesteyn, Raksha Das, Donal Skelly, Thomas G. Ritter, Ali Amini, Sagida Bibi, Sandra Adele, Sile Ann Johnson, Bede Constantinides, Hermione Webster, Nigel Temperton, Paul Klenerman, Eleanor Barnes, Susanna J. Dunachie, Derrick Crook, Andrew J. Pollard, Teresa Lambe, Philip Goulder, Neil G. Paterson, Mark A. Williams, David R. Hall, Elizabeth E. Fry, Jiandong Huo, Juthathip Mongkolsapaya, Jingshan Ren, David I. Stuart, Gavin R. Screaton, Christopher Conlon, Alexandra Deeks, John Frater, Lisa Frending, Siobhan Gardiner, Anni Jämsén, Katie Jeffery, Tom Malone, Eloise Phillips, Lucy Rothwell, and Lizzie Stafford. Antibody escape of SARS-CoV-2 Omicron BA.4 and BA.5 from vaccine and BA.1 serum. *Cell*, 185(14):2422–2433.e13, July 2022. ISSN 00928674. doi: 10.1016/j.cell.2022.06.005. URL https://linkinghub.elsevier.com/retrieve/pii/S0092867422007103.

Vaibhav Upadhyay, Sudipta Panja, Alexandra Lucas, Casey Patrick, and Krishna MG Mallela. Biophysical evolution of the receptor-binding domains of sars-covs. *Biophysical Journal*, 122(23): 4489–4502, 2023.

Dianzhuo Wang, Marian Huot, Vaibhav Mohanty, and Eugene I. Shakhnovich. Biophysical principles predict fitness of sars-cov-2 variants. *Proceedings of the National Academy of Sciences*, 121 (23):e2314518121, 2024. doi: 10.1073/pnas.2314518121. URL https://www.pnas.org/doi/abs/10.1073/pnas.2314518121.

Qian Wang, Sho Iketani, Zhiteng Li, Liyuan Liu, Yicheng Guo, Yiming Huang, Anthony D. Bowen, Michael Liu, Maple Wang, Jian Yu, Riccardo Valdez, Adam S. Lauring, Zizhang Sheng, Harris H. Wang, Aubree Gordon, Lihong Liu, and David D. Ho. Alarming antibody evasion properties of rising SARS-CoV-2 BQ and XBB subvariants. *Cell*, 186(2):279–286.e8, January 2023. ISSN 00928674. doi: 10.1016/j.cell.2022.12.018. URL https://linkinghub.elsevier.com/retrieve/pii/S0092867422015318.

Qianyi E Zhang, Jared Lindenberger, Ruth J Parsons, Bhishem Thakur, Rob Parks, Chan Soo Park, Xiao Huang, Salam Sammour, Katarzyna Janowska, Taylor N Spence, et al. Sars-cov-2 omicron xbb lineage spike structures, conformations, antigenicity, and receptor recognition. *Molecular Cell*, 84(14):2747–2764, 2024.

Tongqing Zhou, I-Ting Teng, Adam S. Olia, Gabriele Cerutti, Jason Gorman, Alexandra Nazzari, Wei Shi, Yaroslav Tsybovsky, Lingshu Wang, Shuishu Wang, Baoshan Zhang, Yi Zhang, Phinikoula S. Katsamba, Yuliya Petrova, Bailey B. Banach, Ahmed S. Fahad, Lihong Liu, Sheila N. Lopez Acevedo, Bharat Madan, Matheus Olivera De Souza, Xiaoli Pan, Pengfei Wang, Jacy R. Wolfe, Michael Yin, David D. Ho, Emily Phung, Anthony DiPiazza, Lauren Chang, Olubukula Abiona, Kizzmekia S. Corbett, Brandon J. DeKosky, Barney S. Graham, John R. Mascola, John Misasi, Tracy Ruckwardt, Nancy J. Sullivan, and Lawrence Shapiro. Structure-Based Design with Tag-Based Purification and In-Process Biotinylation Enable Streamlined Development of SARS-CoV-2 Spike Molecular Probes. *SSRN Electronic Journal*, 2020. ISSN 1556-5068. doi: 10.2139/ssrn.3639618. URL https://www.ssrn.com/abstract=3639618.

A APPENDIX

A.1 METHODS

DMS Dataset processing The second dataset that we examined is a DMS from Starr et al. Starr et al. (2020; 2021a), providing for all possible RBD single mutants K_d values (ACE2) and escape ratios $\epsilon(mut)$ (LY-CoV016, LY-CoV555, REGN10987 and S309), and filtered on residues between 334 and 526 included.

Noting that the dissociation constant of RBD writes:

$$K_D = \frac{[RBD_{free}][Ab]}{[RBD_{bound}]} \propto \frac{[RBD_{free}]}{[RBD_{bound}]} \sim \epsilon$$

, we assumed log dissociation constants of single mutants could be obtained as the sum of wild type log dissociation constant and the variation of escape ratio compared to its minimum value (wildtype):

$$\log K_d(mut) = \log K_d(wt) + \log \frac{\epsilon(mut)}{\min_m \epsilon(m)}$$

Fitness and Biophysical Model We determine the fitness of the RBD based on its contribution to viral infectivity, utilizing a biophysical model established by our previous work Wang et al. (2024). This model leverages the Boltzmann distribution to map molecular phenotypes—characterized by binding energies to cell receptors and antibodies—onto a fitness landscape. The RBD's fitness is primarily determined by two factors: its binding affinity for the ACE2 receptor and its ability to evade antibody neutralization.

In our model, the RBD can exist in multiple states, each with its associated free energy: Unfolded state (G_u), folded but unbound state (G_f), folded and bound to ACE2 (G_{bA}), folded and bound to one of four distinct antibodies (G_{ai}, where i indexes the antibodies).

The fitness function can be expressed as:

$$F = a\frac{Ce^{-\beta G_{bA}} + e^{-\beta G_f}}{Ce^{-\beta G_{bA}} + \Sigma_i C_i e^{-\beta G_{ai}} + e^{-\beta G_f} + e^{-\beta G_u}} \tag{1}$$

In this equation, $C = \frac{[ACE2]}{C_0}$ and $C_i = \frac{[Ab_i]}{m_i C_0}$ represent the normalized concentrations of ACE2 and antibodies respectively, where $[...]$ denotes molar concentration and m_i is a neutralization coefficient specific to each antibody. Noting that $\beta\Delta G = \beta(G_i - G_f) = ln(K_D)/T$ for every state i, where T is a hyperparameter proportional to system temperature, fitness can be expressed as a function of measured dissociation constants.

The model parameters—including the scaling factor a and effective concentrations C and C_i—were calibrated by combining experimental measurements of dissociation constants (K_D) with variant prevalence data from the GISAID database Elbe & Buckland-Merrett (2017a). This biophysical framework can then predict the fitness (F) of any RBD variant given its binding affinities, whether measured experimentally or predicted computationally.

In this paper we did not refit the biophysical model iteratively, instead we used the following coefficients fitted from et al.Wang et al. (2024): $T = 1.6$, $a = 1.57363338$, $C = 5.4764 \times 10^{-7}$, $K_1 = 5.6015 \times 10^{-8}$, $K_2 = 4.5128 \times 10^{-8}$, $K_3 = 7.1825 \times 10^{-8}$, $K_4 = 4.7273 \times 10^{-7}$. While we did not refit these parameters iteratively, the model can be trained with very limited fitness data due to its low number of parameters and still approximate well fitness derived from population data.

Uncertainty of predictions To propagate variance from the predicted dissociation constants to the fitness function, we used standard error propagation formula, which states that the variance of a function f can be approximated as:

$$\sigma_f^2 \approx \sum_i \left(\frac{\partial f}{\partial k_i}\right)^2 \sigma_{k_i}^2, \tag{2}$$

where σ_f^2 is the variance of the function, $\frac{\partial f}{\partial k_i}$ is the partial derivative of the function with respect to the i-th variable, and $\sigma_{k_i}^2$ is the variance of the i-th variable. The variance of variables is obtained from the posterior distribution of the GP.

The derivatives $\frac{\partial f}{\partial \log_{10} K_{d,i}}$ were computed symbolically using SymPy and evaluated numerically with the respective parameter values.

Model Benchmark We evaluate the ability of VIRAL to rank variants by computing the Spearman correlation between inferred fitness scores and ground-truth labels.

For the combinatorial dataset, the model is trained on varying numbers of data points, ranging from 20 to 160, and then used to predict the fitness of the remaining variants.

For the DMS dataset, training sets are constructed by sampling an average of $n \in \{0.2, 0.4, 0.6, 0.8, 1, 2, 3\}$mutations per site. When $n < 1$, a single mutant at each site is randomly selected and included in the training set with probability n.

To ensure robustness, each experiment is repeated 10 times for both datasets and for each training size, enabling a reliable assessment of the model's predictive accuracy across different levels of data availability.

Evaluation of Active Learning Performance The primary metric used to evaluate our pipeline is the enrichment factor (EF), defined as

$$EF(\text{strategy}) = \frac{\text{\# Top variants acquired (strategy)}}{\text{\# Top variants acquired (random)}},$$

where "top variants" refers to those with fitness in the top 10% of the dataset.

Additionally, we compute the Area Under the Curve (AUC) using the active learning predicted fitness as a score to identify whether a variant belongs to the top 10% of the dataset. Notably, this metric is evaluated on the entire dataset (including both tested and untested variants) to ensure the results are not negatively biased against models that have acquired all the top variants.

Lastly, we calculate the embedding variance of the tested variants as a proxy for diversity. Variance is defined as the mean of the coordinate variance among the acquired variants.

Comparison with GISAID data We analyzed SARS-CoV-2 spike sequences from the GISAID database Elbe & Buckland-Merrett (2017b), collecting 15,371,428 sequences up to April 14, 2023. Following the methodology of Starr et al. Starr et al. (2021b), we implemented a sequence filtration process. Sequences were excluded if they were: (1) from non-human hosts, (2) outside the length range of 1,260-1,276 amino acids, (3) contained unicode errors, gaps, or ambiguous characters. The remaining sequences were aligned using MAFFT Katoh & Standley (2013).

The final dataset comprised 11,976,984 submissions, containing 25,725 unique RBD sequences. For each unique RBD sequence, we tracked its frequency of occurrence and estimated its emergence time using the 5th percentile of its temporal distribution. To ensure robustness, we excluded singleton sequences that appeared only once in the dataset.

We assessed site mutability by analyzing amino acid diversity at each position within the RBD. Sites were labeled as "highly mutable" if they exhibited at least 9 distinct amino acid variants out of the possible 20, observed in mutants with a count of 10 or more. Using this cutoff of 10 on variant prevalence helps ensure that the identified mutations provided a clear evolutionary advantage rather than occurring randomly, allowing us to identify sites under selection pressure during the pandemic. To evaluate the effectiveness of our active learning pipeline, we used site sampling frequency as a performance metric and calculated the Area Under the Receiver Operating Characteristic Curve (AUC). The AUC score quantifies our algorithm's ability to identify positions that emerged as mutation hotspots throughout the pandemic. AUC scores were also tested for mutation thresholds different from 9, see SI.

Baseline for identification of highly mutable sites We define a baseline model to identify high-fitness mutations, trained on pandemic data. The training data includes mutations observed in the

GISAID database within a variant having a minimum count of 1000, under the assumption that variants with counts below this threshold lack reliable fitness estimates. Specifically, mutations in variants with occurrences prior December 2020 / December 2022 and counts exceeding 1000 were selected. The baseline was trained on 14 mutations for 2021 deadline and 64 mutations for 2022 deadline. Single-batch acquisition size is $674-$training size, ensuring the total number of variants acquired by the active learning process and the baseline is the same. We then computed acquisition score for every site, based on this single batch acquisition, and compared it to GISAID data.

A.2 Active learning on DMS

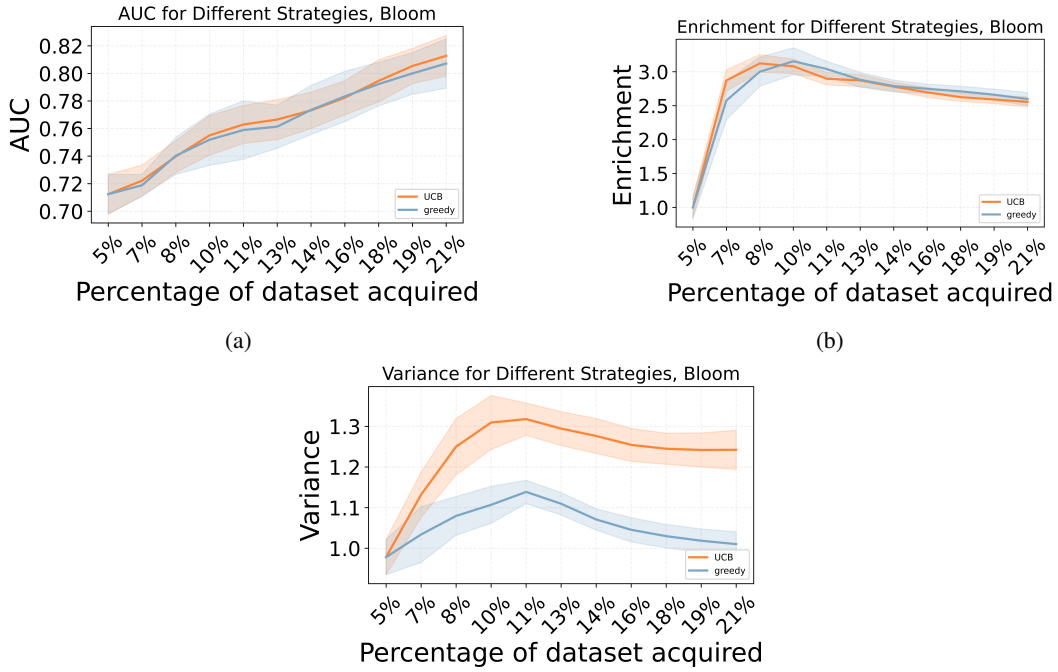


Figure 4: Active learning on DMS. (a) Area Under the Curve (AUC), representing the model's ability to identify top fitness variants, is shown for different strategies (UCB and greedy). An AUC above 0.5 indicates effective identification of high-fitness variants. Each round corresponds to acquiring a new batch of variants, improving the predictor. Shaded regions represent standard deviations across runs. (b) Enrichment in top variants across acquisition runs for each strategy. Values above 1 represent improvement other a random acquisition. (c) Comparison of embedding variance of points acquired across rounds between UCB (orange) and greedy (blue) strategies

We assess the capacity of our pipeline combined with active learning to identify high fitness single variants in the DMS dataset. This task requires to model the impact of approximately $193 * 19$ mutations, with initial knowledge of only one mutation per site. Interestingly, the UCB acquisition perform best in early rounds, while greedy acquisition gets better as the predictor gets more accurate on the longer term. This is explained by the high accuracy of our model to identify top variants after a few round of acquisitions, with AUC above 0.80.

A.3 INFLUENCE OF UNCERTAINTY WEIGHT IN UCB ACQUISITION

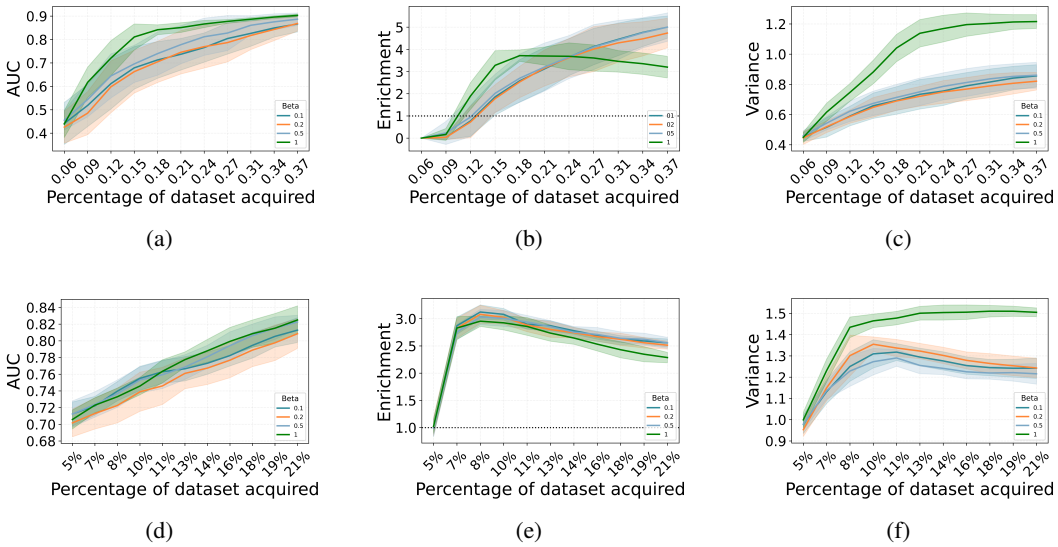
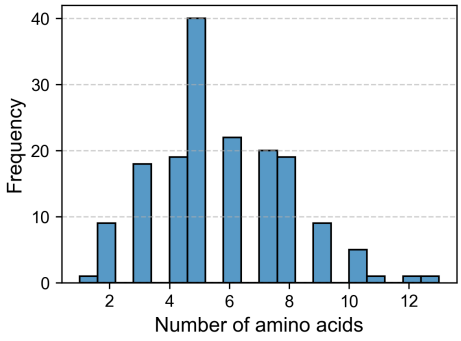


Figure 5: Influence of uncertainty weight β in UCB acquisition. (a-c) AUC, enrichment, and embedding variance on the CM dataset. (d-f) Similar analysis on the DMS dataset. Different values of β control the balance between exploration and exploitation, affecting model performance in identifying high-fitness variants.

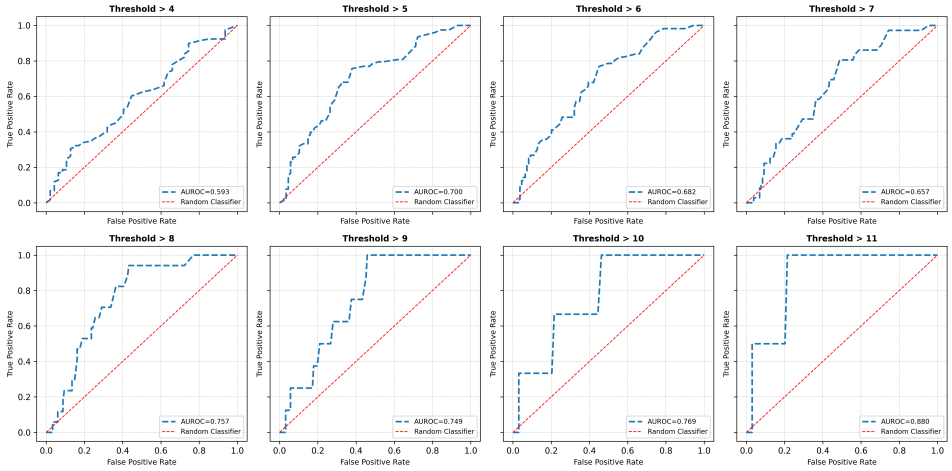
Increasing the uncertainty weight β in UCB acquisition enhances embedding variance and higher diversity in the training set and potentially improving predictive performance (higher AUC). However, in the long run, as the predictor becomes more accurate, it is preferable to reduce the emphasis on uncertainty, allowing the model to rely more on its predictions for optimal decision-making. Consequently, we selected $\beta = 0.2$ as a reasonable trade-off between exploitation and exploration of the mutational landscape. However, it is important to note that the impact of β on overall performance remains relatively limited, and one can choose any reasonable β *a priori* for the purpose of pandemic prevention.

β is then scaled by $\frac{\text{std(fitnesses)}}{\text{std}(\sqrt{\text{vars}})}$ to ensure that the uncertainty term $\sqrt{\text{vars}}$ has a comparable range to the fitness values

A.4 IDENTIFICATION OF MUTATION-PRONE SITES



(a)



(b)

Figure 6: Active learning performance for identifying mutation-prone sites across different thresholds. (a) Histogram showing the distribution of SARS-CoV-2 RBD sites by the number of distinct amino acid substitutions observed during the pandemic. The data reveals most sites experienced between 4-8 different amino acid changes, with a peak at 6 amino acid variants. (b) ROC curves demonstrating VIRAL's ability to identify highly mutable sites in the GISAID database across various thresholds (from > 4 to > 11 mutations). The averaged acquisition scores from 10 independent runs show consistently strong predictive performance, with AUC values ranging from 0.593 to 0.890. Higher thresholds ($> 9, > 10, > 11$) yield the best performance ($AUC > 0.75$), indicating VIRAL is particularly effective at identifying the most highly mutable positions in the viral genome.

We assessed site mutability by analyzing amino acid diversity at each position within the RBD. In the main text, sites were labeled as "highly mutable" if they exhibited at least 9 distinct amino acid variants out of the possible 20, observed in mutants with a count of 10 or more. The performance of VIRAL in identifying highly mutated sites remains robust across different thresholds defining mutation frequency (from > 4 to > 11 mutations), confirming the consistency of our approach regardless of the specific cutoff chosen for analysis.