

Predicting Lassa Virus Glycoprotein Mutations Affecting Viral Entry and Antibody Escape

Oluremi Israel Ajayi

Redeemer's University

Blessing Oluwatobi Olorunfemi

Redeemer's University

Jolly Amoche Adole

Redeemer's University

Amoge Chidinma Ogu

Redeemer's University

Ifeoluwa O. Bejide

Redeemer's University

Blossom Akinawo

Redeemer's University

Christian T. Happi

Redeemer's University

Chinedu A. Ugwu

`ugwu@cun.edu.ng`

Redeemer's University

Research Article

Keywords: Lassa virus, Glycoprotein Complex (GPC), Machine Learning, Antibody Escape, Viral Entry

Posted Date: August 31st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10708477/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Predicting Lassa Virus Glycoprotein Mutations Affecting Viral Entry and Antibody Escape

Oluremi I. Ajayi^{1,2}, Blessing O. Olorunfemi³, Jolly A. Adole^{1,2}, Amoge C. Ogu^{1,2}, Ifeoluwa O. Bejide^{1,2}, Blossom Akinlawo^{1,2}, Christian T. Happi^{1,2}, Chinedu A. Ugwu^{*1,2}

¹. Institute of Genomics and Global Health (IGH), Redeemer's University, Ede, Nigeria.

². Department of Biological Sciences, Faculty of Natural Sciences, Redeemer's University, Ede, Osun State, Nigeria.

³. Department of Computer Sciences, Faculty of Computing and Digital Technologies, Redeemer's University, Ede, Osun State, Nigeria.

*Correspondence:

Chinedu A. Ugwu

ugwuc@run.edu.ng

Abstract

Lassa virus (LASV) is a major cause of viral haemorrhagic fever in West Africa and remains a significant public health challenge because of its extensive genetic diversity and lack of broadly effective vaccines and antibody-based therapeutics. Mutations within the viral glycoprotein complex (GPC), the principal target of neutralizing antibodies, can alter viral entry and susceptibility to antibody-mediated neutralization. This study developed an integrated machine learning framework to predict the effects of experimentally validated amino acid substitutions within LASV GPC on viral entry competence and antibody escape. Mutational data from published pseudovirus entry and neutralization studies were mapped to a reference GPC sequence and analysed across representative LASV lineages. Viral entry was modelled as a binary classification task, while antibody escape was predicted using antibody-specific regression models based on mutation position, wild-type residue, and substituted amino acid. Tree-based ensemble methods outperformed linear and single-tree algorithms, with Extreme Gradient Boosting (XGBoost) achieving the highest predictive performance for viral entry (ROC-AUC = 0.871). Across Lineages I–III, entry-competent and poor-entry mutations occurred in near-equal proportions (approximately 53% and 47%, respectively), while 3,923 mutations were conserved across all three lineages. Antibody escape analyses revealed epitope-specific differences: 25.10C exhibited the broadest predicted escape landscape, whereas 37.2D and 37.7H displayed restricted escape profiles. Intermediate escape patterns were observed for 12.1F, 25.6A, and 8.9F. The framework provides a practical approach for prioritizing functionally important mutations, identifying conserved epitopes with limited escape potential, and supporting rational design of broadly protective vaccines, therapeutic antibody combinations, and genomic surveillance strategies.

Keywords: Lassa virus, Glycoprotein Complex (GPC), Machine Learning, Antibody Escape, Viral Entry

1. Introduction

Lassa virus (LASV) is an Old-World arenavirus that causes Lassa fever, an acute viral hemorrhagic disease endemic to West Africa. Lassa Fever is an important public health concern because of the high morbidity and mortality in Nigeria, Sierra Leone, Liberia, and Guinea (Aloke *et al.*, 2023). The primary mode of transmission involves contact with the urine or feces of infected *Mastomys* rodents, although secondary transmission has been documented from person to person, especially from household and healthcare settings, especially when infection control practices are poor (Uppala *et al.*, 2025).

The clinical presentation of Lassa fever ranges from mild febrile illness to severe disease with symptoms of hemorrhage, multi-organ failure, and neurological manifestations, of which sensorineural hearing loss is one of the common sequelae of survivors of this infection (Ficenec *et al.*, 2020). Despite the significance of Lassa fever as a public health concern in West Africa, therapeutic interventions are limited. There is currently no license vaccine or therapeutic available for the management of LF (Cheng *et al.*, 2022).

The immunopathogenesis of Lassa Fever is not fully understood. One of the most critical factors that define LASV infectivity is the entry of the virus, a process that is controlled by the viral glycoprotein complex. The GPC is the only viral protein exposed on the viral surface responsible for receptor binding and membrane fusion (Hastie *et al.*, 2017). Previous structural studies have demonstrated that LASV GPC has a metastable trimeric structure characterized by high levels of inter-subunit interactions and glycosylation, which are important for the immunoprotection of the virus Figure 1 (Perrett *et al.*, 2022).

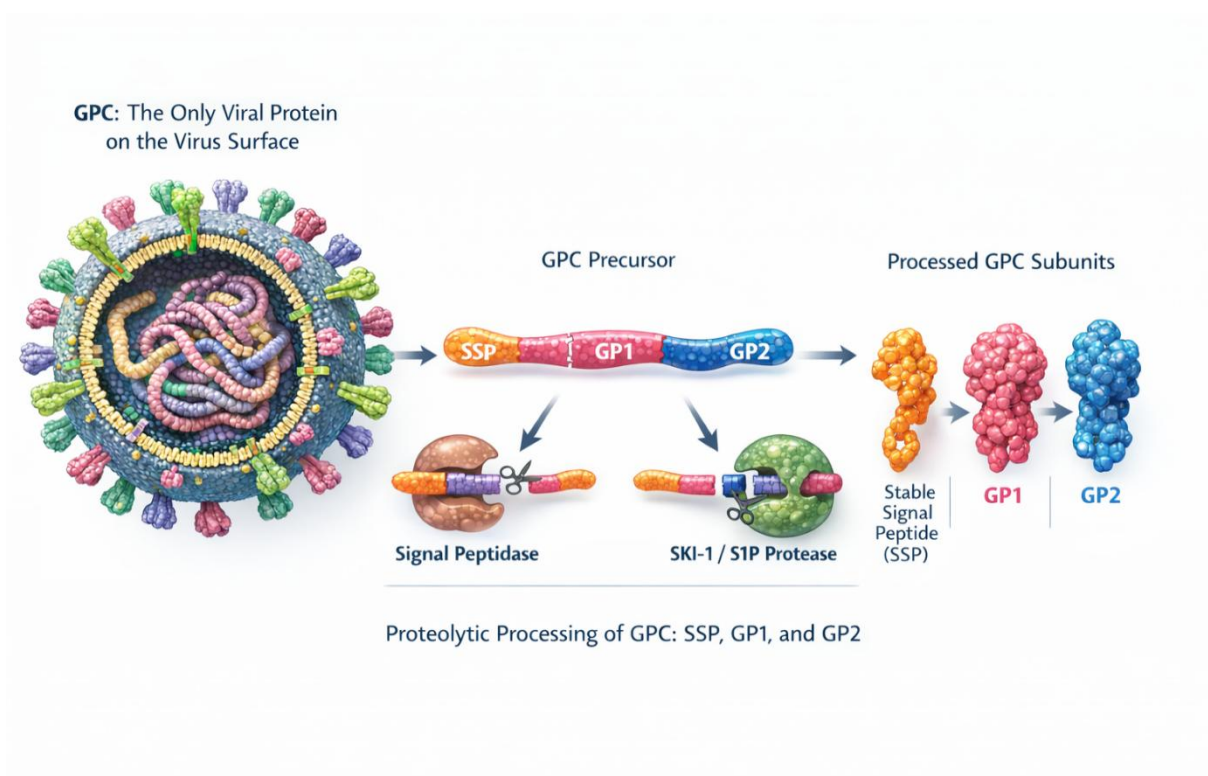


Figure 1. Proteolytic processing and surface organization of the Lassa virus glycoprotein complex (GPC) Adapted from Pennington & Lee, 2022

Because of the critical role of the protein in the infection process, GPC is an important target of neutralizing antibodies LASV GPC, being the only protein on the surface of the virus, is under immense pressure by the immune system, resulting in heightened mutations in the GPC. These mutations drive the genetic diversity seen among LASV, with up to seven different lineages distributed across West Africa (Li, 2023). Alterations in the GPC have been shown to influence protein folding, protein processing, protein receptor-binding affinity, as well as protein fusion activity, thus influencing viral entry process efficiency (Newby *et al.*, 2024). Deep mutational scanning showed that changes in the viral genome, even in one amino acid residue, significantly influence the efficiency of the viral entry process and immune response (Burton & Eyre, 2021; Dadonaite *et al.*, 2024; Carr *et al.*, 2024).

In recent times, there has been considerable progress in the application of machine learning (ML) as a tool for elucidating the link between changes in viral sequence and phenotypes (Elste *et al.*, 2024; Asnicar *et al.*, 2024). Machine learning has successfully predicted the impact of mutations on functional outputs such as protein structure, host receptor binding, viral fitness, antigenicity, and immune escape in many medically important viruses such as SARS-CoV-2, influenza virus, HIV-1, hepatitis viruses, and Ebola virus (Bowyer *et al.*, 2025; Shahin *et al.*, 2026). However, for an important genetically diverse virus like LASV, there has been limited data on the application of ML to predict the biological effects of mutations in the LASV glycoprotein.

Tree-based and ensemble machine learning methods, such as random forests and gradient-boosted trees, remain widely used for mutation effect prediction because they perform well with moderate-sized datasets and can effectively integrate different types of features, including amino acid identity, sequence position, and structural information (Ayulani *et al.*, 2023; Yang *et al.*, 2024).

Recent advances in deep learning have expanded the tools available for mutation effect prediction. In particular, protein language models can learn complex relationships and evolutionary patterns from large collections of unlabeled protein sequences (Cheng *et al.*, 2024).

In this study, we have used the mutational data sets obtained through experiments along with supervised machine learning algorithms for predicting the impact of amino acid substitutions on the entry and antibody escape capabilities of Lassa virus glycoprotein lineages. Predicting the effects of mutations in the Lassa virus glycoprotein complex (GPC) is important because these mutations can influence viral entry into host cells and enable escape from antibody-mediated immunity. Understanding these effects can improve our ability to anticipate viral evolution and identify mutations that may affect disease control strategies.

2. Methodology

2.1 Conceptual Overview of the Study

Mutations that impair viral entry can profoundly influence infectivity, transmission dynamics, and disease severity. For Lassa virus, where entry is mediated exclusively by the glycoprotein complex, identifying mutations that compromise entry competence is therefore a biologically meaningful gatekeeping step for downstream analyses. In this study, poor viral cell entry was formulated as a binary supervised learning task, with machine learning models trained to associate mutation-specific features with experimentally observed entry outcomes. Rather than committing to a single classifier a priori, we adopted a multi-model benchmarking strategy in which several established algorithms were trained and evaluated under identical experimental conditions. This design allows model selection to be driven by empirical performance rather than methodological preference (Figure 2). Throughout the workflow, particular emphasis was placed on preventing information leakage, maintaining reproducibility, and performing post-hoc diagnostics, thereby ensuring that performance estimates reflect genuine generalization rather than artifacts of data handling.

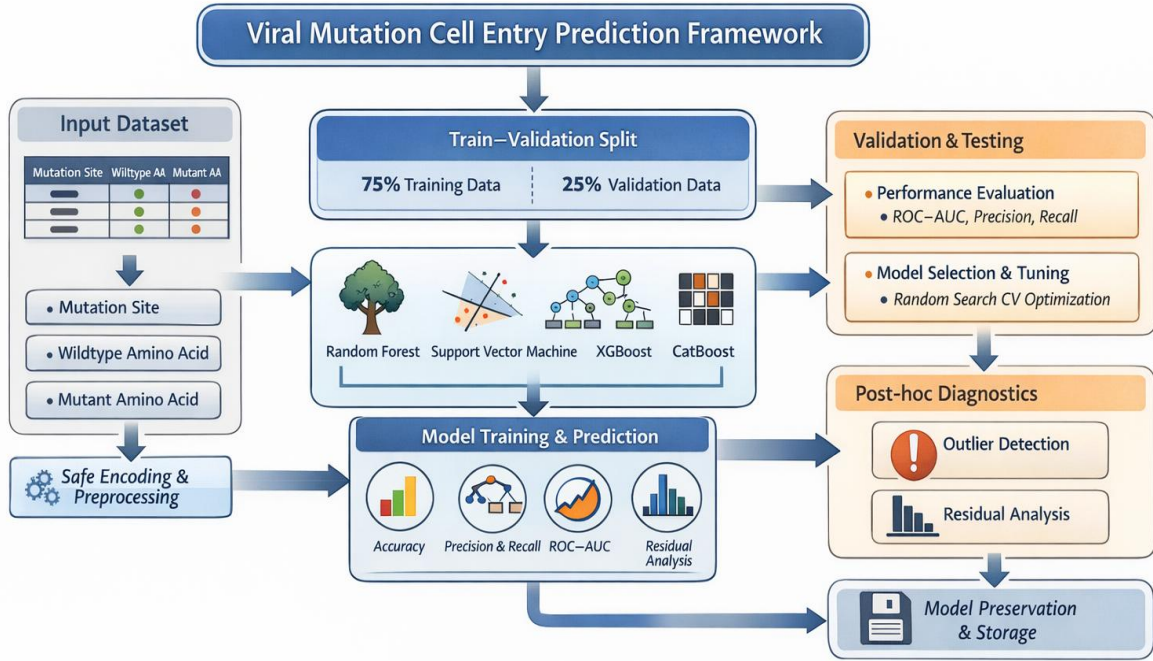


Figure 2. Workflow of the viral mutation cell entry prediction framework.

Mutation site, wild-type amino acid, and mutant amino acid features were encoded and preprocessed prior to model training. The dataset was split into training (75%) and validation (25%) sets. Multiple machine learning models (Random Forest, Support Vector Machine, XGBoost, and CatBoost) were trained and evaluated using ROC-AUC, precision, and recall. Post-hoc diagnostics, including outlier detection and residual analysis, were performed before model preservation and storage.

2.2 Problem Formulation

The dataset is formulated as a supervised learning problem consisting of input-output pairs:

$$D = \{(x_i, y_i)\}_{i=1}^N$$

where x_i represents the feature vector describing the i -th viral mutation, and y_i denotes the corresponding observed biological outcome related to viral entry efficiency.

Since the task is formulated as a binary classification problem, the target variable is defined as:

$$y_i \in \{0,1\}$$

where $y_i = 1$ indicates that the mutation leads to poor viral cell entry, while $y_i = 0$ indicates normal or unimpaired entry functionality.

The objective of the learning process is to approximate an unknown nonlinear mapping function:

$$f: \mathbb{R}^d \rightarrow \{0,1\}$$

This function estimates the probability that a given mutation disrupts viral entry mechanisms. In practice, model outputs may be interpreted either as hard class predictions or probabilistic scores, depending on the evaluation criterion used.

2.3 Dataset Description and Feature Selection

The dataset consists of experimentally validated mutation records of the glycoprotein complex (GPC), structured in tabular format as shown in Table 1. Each observation corresponds to a single amino acid substitution and its experimentally determined entry phenotype.

To ensure biological relevance and interpretability, three core mutation-specific features were selected. These include the mutation site, which captures the positional location of the amino acid change along the GPC sequence and reflects structural constraints affecting folding and receptor interaction; the wild-type amino acid, which represents the original residue prior to mutation; and the mutant amino acid, which describes the substituted residue after mutation.

These features collectively define the input representation:

$$x_i = [x_{i1}, x_{i2}, x_{i3}]$$

Only samples with complete and experimentally verified outcome labels were retained. This ensures that the learning process is grounded in reliable biological evidence while minimizing noise introduced by incomplete or ambiguous annotations.

Summary of the Lassa virus GPC mutation dataset used in this study

Category	Description
Virus	Lassa virus (LASV)
Protein	Glycoprotein complex (GPC)
Data source	NCBI GenBank
Reference	DOI: 10.1016/j.immuni.2024.06.013
Lineages included	LI, LII, LIII, LIV
Geographic origin	Nigeria, Sierra Leone, Guinea
Total mutations	7,489
Unique mutation sites	491
Mutation type	Amino-acid substitutions
Features used	Site, wild-type residue, mutant residue
Target variable	Poor viral entry (binary: TRUE/FALSE)
Downstream filtering	Entry-incompetent mutations excluded prior to escape analysis

2.4 Reproducibility and Experimental Control

Reproducibility is a fundamental requirement in computational biology, particularly when machine learning models are applied to mutation effect prediction. To ensure that results are stable and independently reproducible, all stochastic components of the pipeline were controlled using a fixed random seed.

This seed was consistently applied across dataset partitioning, model initialization, hyperparameter optimization, and cross-validation procedures. As a result, any variation in model performance can be attributed to methodological differences rather than randomness in data sampling or model training.

2.5 Train–Validation Data Partitioning

To evaluate generalization performance, the dataset was partitioned into mutually exclusive training and validation subsets:

$$D_{train} \cup D_{val} = D, \quad D_{train} \cap D_{val} = \emptyset$$

A stratified 75:25 split was adopted to ensure that both subsets preserve similar class distributions:

$$P(y = 1)_{train} \approx P(y = 1)_{val}$$

This step is critical in biological datasets where class imbalance may otherwise bias performance estimation. Importantly, the split was performed prior to any preprocessing operations to prevent information leakage.

2.6 Feature Encoding and Data Preprocessing

Categorical mutation features cannot be directly used by machine learning models and therefore require transformation into numerical representations. To avoid information leakage, encoding mappings were learned exclusively from the training dataset.

The encoding strategy is defined as:

$$g(c) = \begin{cases} k, & \text{if } c \in C_{train} \\ -1, & \text{if } c \notin C_{train} \end{cases}$$

where C_{train} represents the set of categories observed during training. Unseen categories in validation data are assigned a reserved encoding value to ensure robustness during inference.

The mutation site feature was preserved in its numerical form because it encodes biologically meaningful positional information along the protein sequence. Missing values were handled using median imputation:

$$x_{site}^* = \text{median}(x_{site}^{train})$$

This approach preserves distributional stability while avoiding distortion from extreme values.

2.7 Machine Learning Models

To ensure a fair and comprehensive comparison, four widely used machine learning classifiers were evaluated under identical experimental conditions.

2.7.1 Random Forest Classifier

Random Forest is an ensemble learning method that aggregates predictions from multiple decision trees trained on bootstrapped subsets of the data. The final prediction is computed as:

$$\hat{p}(y = 1|x) = \frac{1}{T} \sum_{t=1}^T h_t(x)$$

where T is the number of trees and $h_t(x)$ represents the prediction of the t -th tree. This aggregation reduces variance and improves generalization performance.

2.7.2 Support Vector Machine

The Support Vector Machine (SVM) employs a radial basis function kernel to capture nonlinear relationships in the feature space:

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$$

This kernel enables transformation into a higher-dimensional space where complex decision boundaries become linearly separable.

2.7.3 XGBoost

XGBoost builds an additive ensemble of decision trees in a sequential manner, where each new tree corrects errors made by previous models. The optimization objective is defined as:

$$L = \sum_{i=1}^N l(y_i, \hat{y}_i) + \sum_k \Omega(f_k)$$

where l represents the loss function (log-loss for classification), and Ω is a regularization term that penalizes model complexity to prevent overfitting.

2.7.4 CatBoost

CatBoost is a gradient boosting framework optimized for categorical data and employs an ordered boosting strategy to reduce prediction bias. Its loss function is defined as:

$$l(y, p) = -[y \log(p) + (1 - y) \log(1 - p)]$$

This formulation ensures stable optimization for binary classification tasks involving biological sequence data.

2.8 Model Training and Prediction

During inference, each trained model produces both probability estimates and class labels. The decision rule is defined as:

$$\hat{y}_i = \begin{cases} 1, & \hat{p}_i \geq 0.5 \\ 0, & \hat{p}_i < 0.5 \end{cases}$$

where:

$$\hat{p}_i = P(y = 1 | x_i)$$

These probability estimates are essential for threshold-independent evaluation and downstream diagnostic analysis.

2.9 Evaluation Metrics

Model performance was evaluated using a combination of complementary metrics.

Overall classification accuracy is defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision and recall are defined as:

223
$$Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN}$$

224 Sensitivity and specificity are captured using:

225
$$TPR = \frac{TP}{TP + FN}, \quad FPR = \frac{FP}{FP + TN}$$

226 The Area Under the ROC Curve (AUC) is defined as:

227
$$AUC = \int_0^1 TPR(t) d(FPR(t))$$

228 ROC–AUC is selected as the primary evaluation metric due to its robustness to class imbalance
229 and threshold independence.

230 2.10 Model Selection and Hyperparameter Optimization

231 To obtain a reliable estimate of generalization performance, cross-validation was applied
232 across K folds:

233
$$ROC-AUC_{cv} = \frac{1}{K} \sum_{k=1}^K AUC_k$$

234 This ensures that model selection is not biased by a single train–validation split and improves
235 robustness in performance estimation.

236 2.11 Post-hoc Diagnostic and Residual Analysis

237 To assess model calibration and identify systematic prediction errors, residuals were computed
238 as:

239
$$r_i = y_i - \hat{p}_i$$

240 Standardized residuals were defined as:

241
$$z_i = \frac{r_i - \mu_r}{\sigma_r}$$

242 Observations with $|z_i| \geq 3$ were considered potential outliers, indicating regions of poor model
243 calibration or anomalous prediction behavior.

244 2.12 Model Preservation and Reusability

245 The final trained model, along with the preprocessing pipeline and experimental configuration,
246 was serialized and stored. This ensures reproducibility of results and enables consistent
247 deployment on future unseen mutation datasets without retraining.

248 3. Results

249 3.1 Assessment of the Best Model for Viral Entry Prediction

250 The performance of the poor viral entry discrimination model was evaluated using receiver
251 operating characteristic (ROC) analysis on the validation dataset. The validation dataset
252 included experimentally characterized amino acid substitutions in the Lassa virus (LASV)

glycoprotein complex (GPC). These data were obtained from peer-reviewed studies of Carr *et al.*, 2024 that reported quantitative measurements of viral entry and antibody neutralization. Functional data were derived primarily from pseudovirus entry assays and neutralization assays conducted in mammalian cell lines. Only mutations with clearly defined residue position, amino-acid substitution, and corresponding functional measurement were included.

Residue numbering was harmonized to a curated Lassa virus glycoprotein complex (GPC) reference sequence obtained from the National Center for Biotechnology Information (NCBI) GenBank database (NC_004297, Lineage IV, Josiah strain). Lineage-specific analyses used representative complete GPC sequences: Lineage I, Pinneo 1969 (KM822128), Lineage II, Nig08_04 (GU481068), and Lineage III, NIG-CSF-Jos 2000 (GU481072), retrieved from GenBank and verified for annotation consistency. All mutations were cross-validated against the selected reference sequence to confirm wild-type residue identity and positional accuracy. Tree-based models significantly outperformed the support vector machine (SVM), suggesting that non-linear feature interactions contribute strongly to classification of mutations affecting viral entry.

3.1.1 Model Performance Comparison

The performance of XGBoost was at its best with a ROC-AUC value of 0.856, while that of CatBoost also scored relatively highly with an ROC-AUC value of 0.841. The two algorithms were better in terms of accuracy and precision than Random Forest. SVM had relatively poor performance (ROC-AUC = 0.622), while Random Forest had relatively good performance (ROC-AUC = 0.799) Table 2.

Table 2. Model Performance Comparison

Model	Accuracy	ROC_AUC	Precision	Recall
XGBoost	0.7843	0.8563	0.7931	0.7069
CatBoost	0.7683	0.8415	0.7754	0.6856
Random Forest	0.7256	0.7993	0.7075	0.6690
SVM	0.6001	0.6215	0.5484	0.6501

According to the results of the baseline comparisons as shown in figure 3, XGBoost was chosen for further fine-tuning. After the hyperparameter optimization, the ROC-AUC of the algorithm was 0.871 with accuracy 0.798, precision 0.792, and recall 0.749.

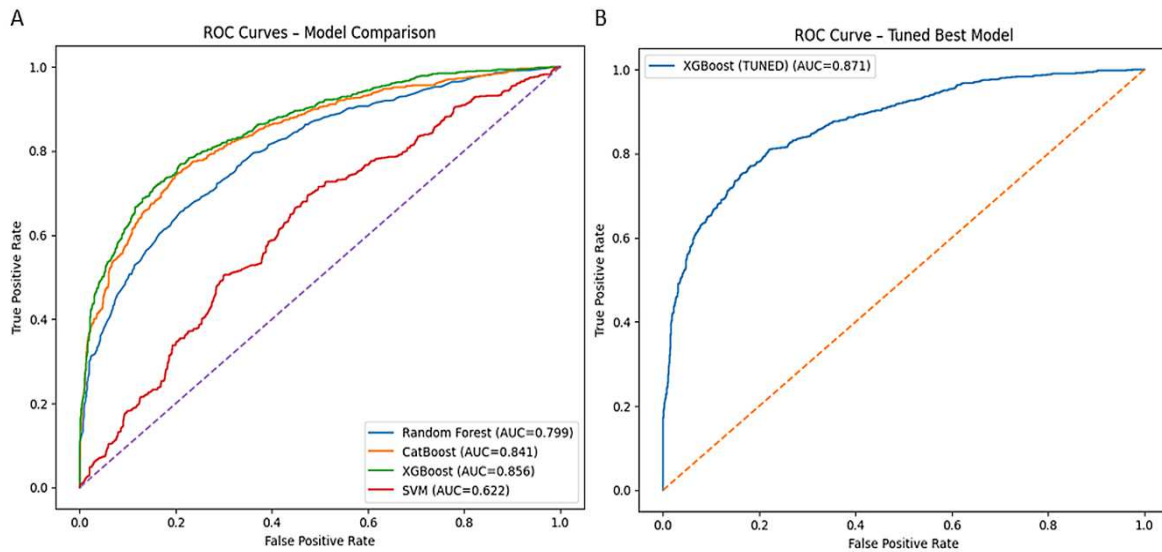


Figure 3. ROC curve comparison and tuned XGBoost for poor viral entry classification models: *A. Receiver operating characteristic (ROC) curves comparing model performance on held-out validation data. Boosting-based models (XGBoost and CatBoost) show superior discrimination across decision thresholds relative to Random Forest and SVM. The dashed diagonal indicates random classification performance (AUC = 0.5). B. ROC curve showing the performance of the tuned XGBoost model on held-out validation data. The dashed diagonal indicates random classification. Model tuning resulted in improved discrimination relative to the baseline model.*

Based on the confusion matrix (Figure 4), the optimized XGBoost classifier was able to efficiently differentiate mutations possessing the poor cell entry mutation from the ones with efficient cell entry mutations. For both classes, there was a high number of correct predictions compared to errors, which is an indication that the XGBoost classifier learnt some inherent patterns of the relation of the features of mutation and the phenotype. The high prevalence of correct negative predictions means that the classifier was efficient in identifying the mutations that had low chances of blocking Lassa virus entry into the cells. On the other hand, the correct positive predictions show the ability of classifier in identifying mutations that block the cellular entry.

False positives mean that there were some mutations that were identified by the classifier as potentially disrupting the Lassa virus entry but were normal in regard to the level of cellular entry, as seen from the experimental results, while the false negatives indicate mutations that blocked the entry but were unrecognized by the classifier. In addition, the higher proportion of the false negatives as compared to false positives means that there was a bias due to the setting of boundary by the classifier in order to achieve maximum specificity at the cost of slight compromise in sensitivity.

Overall, our result indicates that the mutation information encoded in the sequence is sufficient in training the XGBoost classifier to correctly predict the effects of mutations on the viral entry. The accuracy from our prediction demonstrated that the gradient boosting approach can the impact of amino acid substitutions on Lassa virus glycoprotein functions.

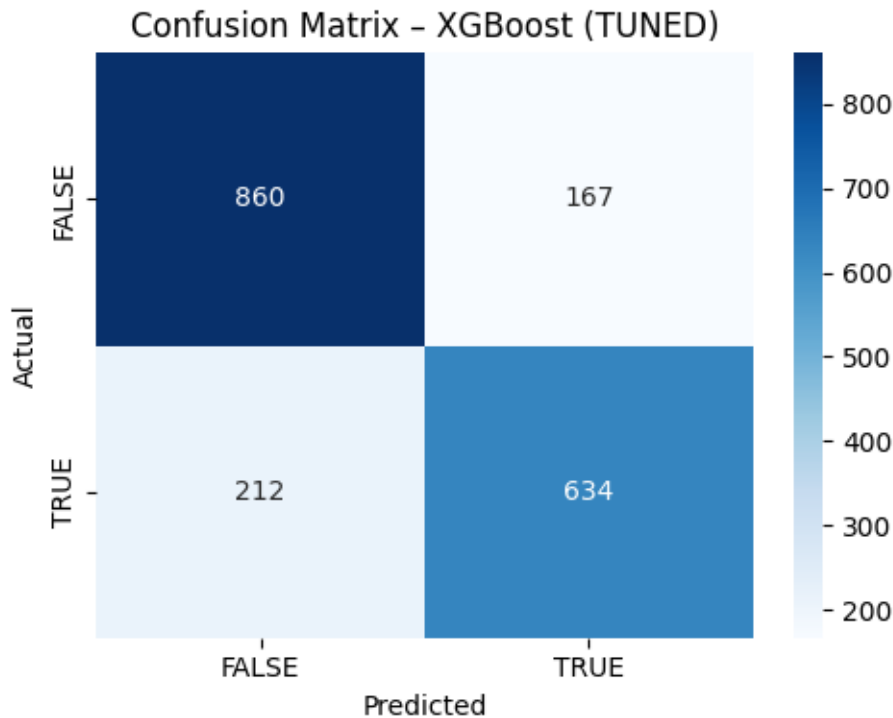


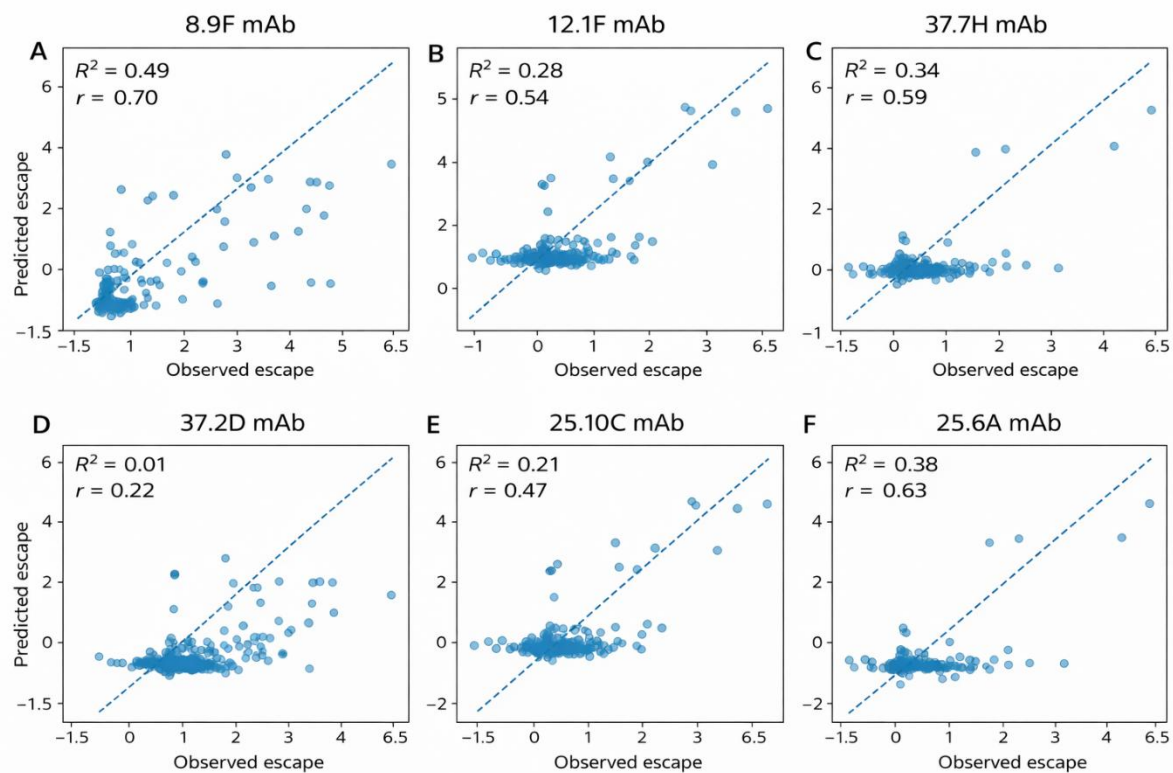
Figure 4. Confusion matrix for the tuned XGBoost classifier.

Confusion matrix summarizing classification outcomes for the tuned XGBoost model on the validation set. Counts indicate true negatives (860), false positives (167), false negatives (212), and true positives (634), demonstrating balanced sensitivity and specificity.

3.2 Regression Model Validation across Monoclonal Antibody Datasets

The scatter plots showed variability in the predictive power of six reference LASV monoclonal antibody data sets as shown in Figure 5 (Robinson *et al.*, 2016). The agreement between predicted and experimentally observed escape values varied greatly between the six antibodies indicating that model performance depends on natural distribution and variability of mutation escape profiles. Data points for 8.9F and 25.6A antibodies displayed very close correspondence between the predicted and observed values, relatively small scatter around the identity line, as well as relatively higher correlation of determination ($R^2=0.49$ and $R^2=0.38$, respectively) and Pearson correlation coefficient ($r=0.70$ and $r=0.63$). Antibodies; 37.7H, 12.1F and 25.10C showed moderate agreement between the predicted and experimentally observed values with a positive correlation ($r=0.47$ to $r=0.59$). However, the scatter around the identity line and relatively lower correlation of determination ($R^2=0.21$ to $R^2=0.34$) demonstrated lower performance over a wider range of observed escape values. Data for 37.2D antibody was different from those for other. LightGBM produced the lowest errors among all models, but both correlation coefficient and coefficient of determination were very low ($R^2=0.01$ and $r=0.22$, respectively). It can be clearly seen that predicted and observed escape values are not correlated with each other in this data set.

The performance of the models varied greatly between the antibodies, with CatBoost producing the best results for 4 out of 6 datasets, XGBoost for 12.1F dataset, and LightGBM for 37.2D. Despite the ability of machine learning methods to identify potential escape mutations, the performance of the models is dependent on natural distribution of mutation escape profiles. It seems that the models manage to maintain the rank of effects for these antibodies and can be used for screening the potential escape mutations for experimental verification.



336

337 **Figure 5. Observed vs predicted escape values across monoclonal antibody datasets.**

338 *Scatter plots show predicted escape values versus experimentally observed escape for held-out*
339 *validation sets across six monoclonal antibody datasets. The dashed line indicates the identity*
340 *line ($y = x$). Models capture the overall trend of escape but exhibit increased dispersion and*
341 *underestimation at higher escape values, reflecting data skew and limited representation of*
342 *extreme escape mutations.*

343 **Table 3. Antibody-specific regression performance of the best machine learning models for**
344 **LASV GPC escape prediction**

Antibody	Best Model	RMSE	MAE
8.9F mAb	CatBoost	0.4897	0.2716
12.1F mAb	XGBoost	0.5326	0.3092
37.7H mAb	CatBoost	0.5258	0.3202
37.2D mAb	LightGBM	0.1856	0.0823
25.10C mAb	CatBoost	0.4472	0.2615
25.6A mAb	CatBoost	0.3112	0.1581

345 **3.3 XGBoost Prediction of Mutation Effects Across Lineages**

346 Across representative sequences of the three lineages of Lassa virus (LI to LIII) as shown in
347 Figure 6, the prevalence of mutations which have been categorized as entry-competent and
348 poor-entry is almost equally distributed. The prevalence of entry-competent mutations has been
349 observed to be slightly more prevalent in all the lineages, being 52.93%, 52.86%, and 52.99%
350 for LI, LII, and LIII respectively. In contrast, poor-entry mutations were found to be about
351 47.0%-47.2% prevalent in the three lineages.

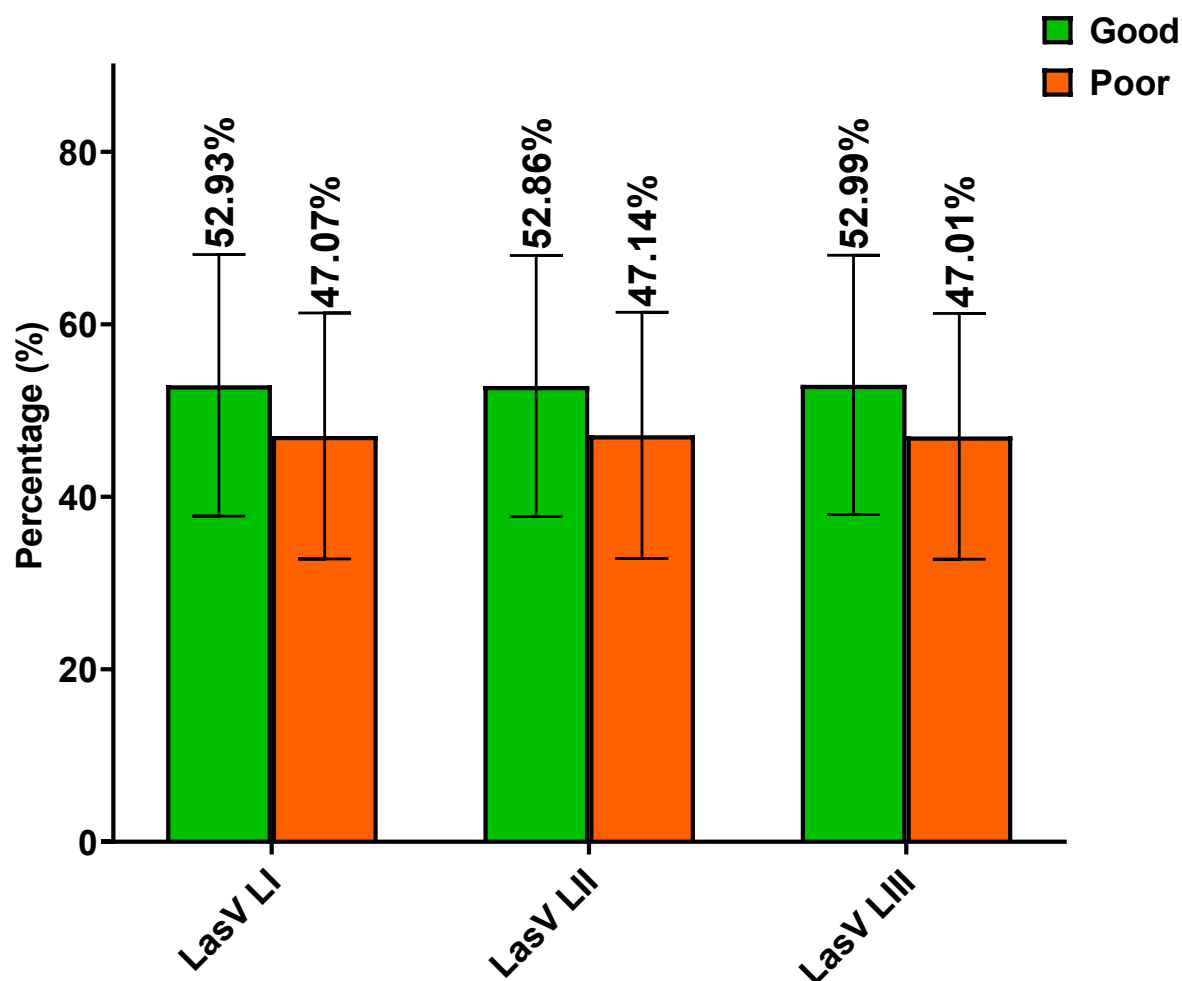


Figure 6. Viral entry classification and site conservation across Lassa virus lineages.
Percentage of entry-functional (Good) and poor-entry (Poor) mutations in Lineages I–III.

Among the different sets of mutations in the study, 3,923 mutations were found to be common between the three lineages and represent the highest number among all the other mutations assessed. On the other hand, the lineage-specific mutations have been observed to account for a relatively lower number compared to the conserved mutations with 265, 261 and 240 lineage-specific mutations being observed in Lineage III, Lineage II and Lineage I respectively. Pair wise analysis revealed that there are 182 mutations which are exclusive to Lineage I and Lineage II, 177 mutations which are exclusive to Lineage I and Lineage III and 157 mutations exclusive to Lineage II and Lineage III. It can therefore be noted that though each lineage possesses its own mutational pattern, most of the mutations in the glycoprotein occur in conserved regions.

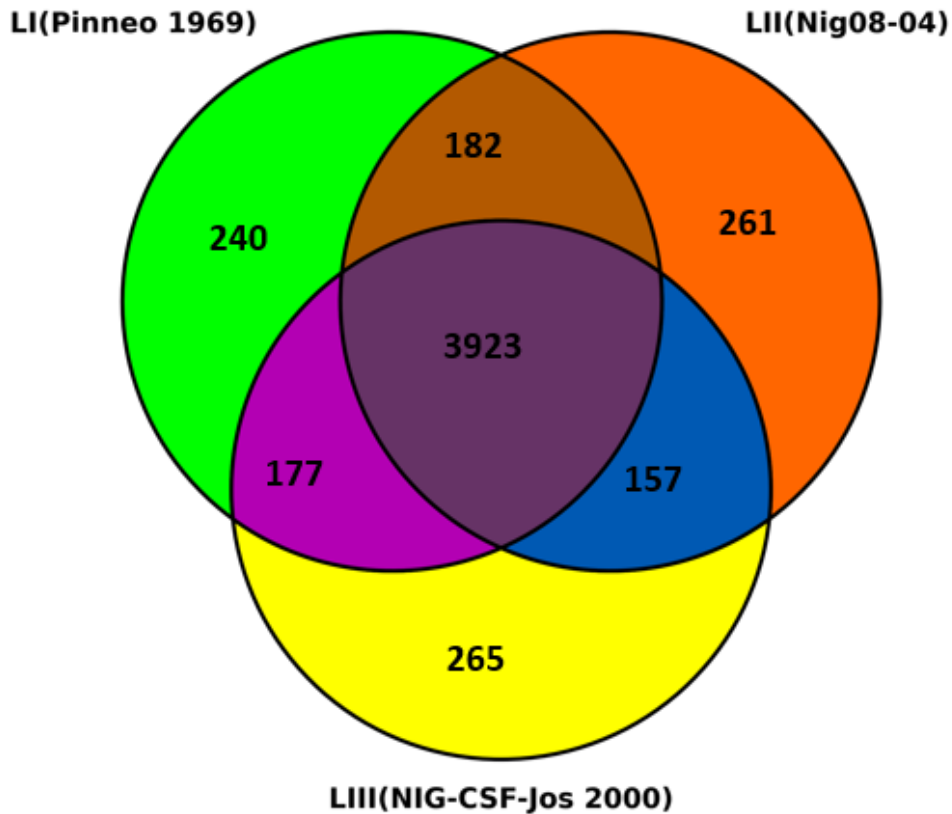


Figure 7. Overlap of mutations across three Lassa virus lineages.

Venn diagram indicating the distribution of mutations among the three lineages of Lassa virus: LI (Pinneo 1969), LII (Nig08-04), and LIII (NIG-CSF-Jos 2000). The number of unique mutations for each lineage is as follows: LI, 240; LII, 261; and LIII, 265. The number of mutations common between each paired of lineages are: LI–LII: 182, LI–LIII: 177, and LII–LIII: 157; mutations shared by all three lineages total 3923.

4. Immune Escape Profiles across Three Lassa Virus Lineages

In order to determine whether there were lineage-based and/or neutralizing antibody-based differences in escape mutation distribution, the escape mutations with values greater than 0.7 were selected from each of the representative LASV lineages (Table 4). These mutations were used for determining differences between the lineages and neutralizing antibodies in terms of their predicted escape. Amino acid mutations having predicted escape values of above 0.7 show significant variations in terms of Lassa virus lineages and antibodies. In the case of Lineage I, mutations having high escape values are mostly seen for antibodies 25.10C and 25.6A. The highest numbers of such mutations are found in the case of 25.10C antibody (19), while a smaller number is seen for 25.6A; in this lineage, no mutations of such a nature have been found for 12.1F and 8.9F antibodies. However, the same is seen to change in the case of lineage II, where high escape mutations have been found for three antibodies, and the highest number of mutations (18) have been seen in 8.9F antibody followed by 12.1F (5) and 25.10C (3), while none have been seen for 25.6A. The same trend continues even for lineage III, where high escape mutations have been seen for 12.1F (6) and 25.10C (3), while none have been seen for 25.6A and 8.9F antibodies. Thus, the above findings suggest that escape is both antibody and lineage dependent with the mutation for 25.10C being highly escapable in lineage I and 8.9F in lineage II, while 25.6A antibody shows very little escape potential.

Table 4. Distribution of unique mutations with predicted escape values greater than 0.7 across *Lassa virus* lineages I–III.

Lineage	25.6A	25.10C	12.1F	8.9F
LI	2	19	–	–
LII	–	3	5	18
LIII	–	3	6	–

The variation of escape values distribution in Figure 8 is clear among the antibodies since some of them are more vulnerable to escape than the others. 25.10C is the antibody with the highest escape values because there is a wide range of mutations which goes to higher numbers showing its high vulnerability to escape through several different ways of mutations. 8.9F, 12.1F and 25.6A have a moderate distribution of escape values because they have both low and intermediate values while 37.7H and 37.2D have consistently low values less than 0.7 set threshold. Geometric mean values of escape also indicate that 25.10C has the highest value of escape, followed by 12.1F, 25.6A and 8.9F.

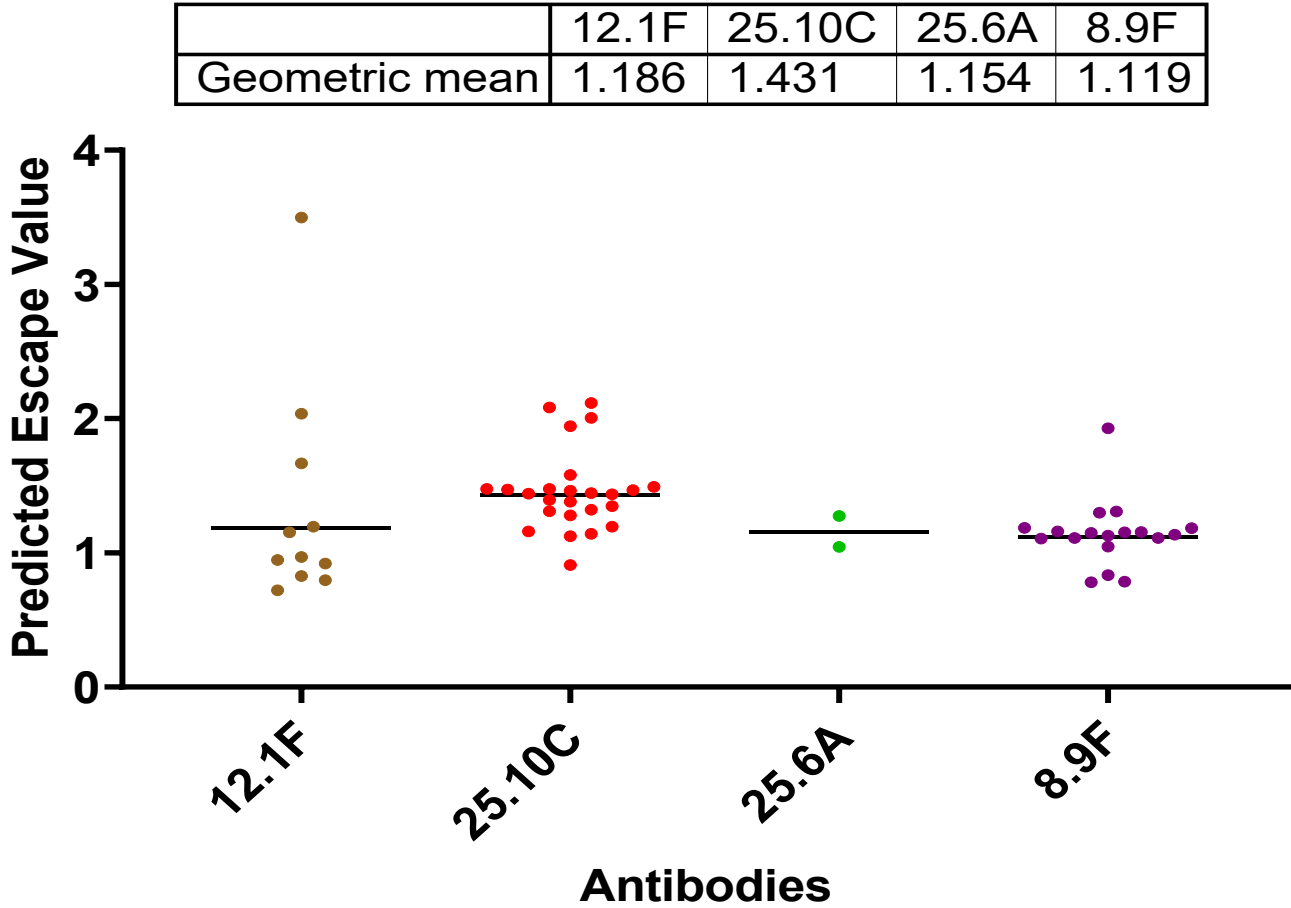


Figure 8. Escape values across antibodies targeting LASV GPC
Predicted escape values greater than 0.7 across *Lassa virus* lineages I–III.

5. Discussion

The LASV glycoprotein complex (GPC) is synthesized as a precursor protein that undergoes proteolytic cleavage by host-cell proteases, including the subtilisin-like proprotein convertase

SKI-1/S1P, to produce three subunits: the stable signal peptide (SSP), GP1, and GP2. The GP1 subunit mediates viral attachment through binding to α -dystroglycan on the host cell surface, whereas GP2 facilitates membrane fusion following viral endocytosis and exposure to the acidic environment of late endosomes, where subsequent interaction with lysosome-associated membrane protein 1 (LAMP1) occurs (Loureiro *et al.*, 2019; Pennington & Lee, 2022). The SSP plays an essential regulatory role by stabilizing the prefusion conformation of the GPC (Pennington & Lee, 2022). Given these critical roles in viral entry, mutations within the LASV GPC can influence both infectivity and susceptibility to antibody-mediated neutralization. Understanding and predicting the effects of these mutations on viral entry and immune responses are therefore essential for elucidating LASV evolution and for informing the development of effective vaccines and therapeutic antibodies, which remain limited for the management of Lassa fever. In this research, data on LASV GPC mutations were used alongside machine learning methods to explore the functional map of the LASV GPC. The correlation between amino acid changes, the ability of the virus to enter host cells, and evasion from antibodies as seen in Figure 6 & 8, clearly demonstrates the context-specific nature of mutations and provides further information about the relationship between structural restrictions and viral evolution under the pressure of the immune system.

The excellent classification capacity of tree-based ensemble models, especially XGBoost, in differentiating between poor-entry and entry-competent mutations is indicative of the need to detect non-linear dependencies in mutational datasets. Functionality of viral proteins is usually not determined by linear interactions of the type but involves intricate relationships between the residues, structural considerations, and epistatic effects. For that reason, ensemble learning is perfectly suitable for modelling such complexity and has proven superior to linear or kernel methods in predicting the fitness landscapes of viral proteins (Freschlin *et al.*, 2022). The minor boost in model accuracy after hyperparameter tuning is another indication that the experimental data contains substantial predictive information, which justifies the choice of XGBoost for mutational modelling in LASV. This is supported by existing literature on gradient boosting techniques like XGBoost, which have been shown to be able to effectively use patterns within the biological sequence data without needing to optimize parameters extensively, especially when there are informative features present (Chen & Guestrin, 2016; Lundberg *et al.*, 2020).

Moreover, the predicted distribution of entry phenotypes is an indicator of functional constraints affecting LASV glycoprotein. In the three different lineages of LASV (Figure 6), competent entry mutants slightly outscored incompetent entry mutants in number, contributing about 53% of the predicted amino acid substitutions, while incompetent entry mutants contributed about 47%. This consistent result across Lineages I, II, and III implies that the glycoprotein is located in a functional landscape in which many amino acid substitutions can be made without the ability to compromise the entry process, yet some substitutions are still deleterious. The constrained fitness landscape is due to the important role of the glycoprotein in the binding to the receptors, membrane fusion, and infectivity, thus maintaining enough stability but allowing evolution at the same time. A similar functional landscape has been observed in other enveloped RNA viruses in which only a fraction of mutations can be tolerated without fitness effects (Zapata & Salvato, 2013; Carr *et al.*, 2024).

The analysis at the lineage level also underscores these functional constraints. Even though each lineage included a specific mutation profile, most of the predicted glycoprotein mutations (3,923) were conserved among Lineages I–III, and only a small number of mutations were characteristic of particular lineages. The predominance of conserved mutations suggests that despite ongoing diversification, many portions of the LASV glycoprotein sequence are still evolutionarily constrained. The conserved sequences could include functional domains whose

alteration negatively affects glycoprotein folding, binding to receptors, or membrane fusion. On the other hand, lineage-specific mutations may be considered evidence of virus diversification in endemic areas with the preservation of the glycoprotein structure essential for its replication and propagation. These results support the findings from previous phylogenetic studies on the diversification of LASV in the presence of a highly conserved glycoprotein structure (Bowen *et al.*, 2000; Andersen *et al.*, 2015; Zapata & Salvato, 2013).

Despite significant constraints in the structure and function of the glycoprotein, mutations which do not affect the fitness of the virus can result in changes in susceptibility to antibody mediated neutralization (Carr *et al.*, 2024). This means that the study of the patterns of antibody escape offers an alternative perspective on the evolutionary paths of LASV that are open in response to immune selection.

In this case, the validation based on regressions has shown an important difference between the quality of predictions and their biological significance. While correlations between predicted and measured escape values are rather low, the model has demonstrated the ability to rank the order of mutations much better than predict their escape values quantitatively. The same trends have been reported in the context of deep mutational scanning experiments, where computational predictions are more often able to rank mutations of possible biological significance, compared to prediction of their quantitative effects (Reeb *et al.*, 2020). The higher variance of high escape mutations can be explained by their low occurrence and stochastic nature.

Substantial disparities were noted regarding the expected escape landscapes for the various antibodies tested. It is clear that the ability of antibodies to mutate is dependent mostly on epitope specificity, rather than being an inherent feature of LASV glycoprotein. The escape landscape was most expansive for antibody 25.10C and its geometric mean escape score was the largest. Therefore, there are many mutations that could potentially decrease the effectiveness of this particular antibody. On the other hand, 37.7H and 37.2D consistently had escape scores lower than the threshold and thus demonstrated restricted escape landscapes. Antibodies 12.1F, 25.6A, and 8.9F fall in between the two extremes, as most mutations result in low to moderate impact escape effects. It appears that different epitopes on LASV glycoproteins differ significantly regarding the tolerance to mutations (Carr *et al.*, 2024; Perrett *et al.*, 2022).

These structural distinctions are well aligned with the epitope structural classification proposed by Robinson *et al.* (2016) for LASV neutralizing antibodies belonging to the GPC-A, GPC-B, and GPC-C antibody competition groups, depending on the epitopes localized on the glycoprotein complex. 25.10C antibody is grouped with the GPC-A category of antibodies, while 37.2D and 37.7H are part of GPC-B antibodies, and 8.9F epitopes form the separate GPC-C class. As was revealed from structural analysis, GPC-A antibodies are highly susceptible to mutations in GP1 N-terminus and GP2 fusion loop, while GPC-B antibodies require maintaining of prefusion GP1-GP2 interface to neutralize the virus efficiently. In its turn, the GPC-C antibody neutralizes quaternary epitope composed of two glycoproteins and stabilized via different interactions rather than by a linear one (Hastie *et al.*, 2017; Robinson *et al.*, 2016; Perrett *et al.*, 2022). Thus, there is a potential explanation of different escape profiles of the antibodies examined in this study. Escape landscape prediction reveals higher tolerance of amino acids substitutions in GPC-A epitopes compared with more restricted profiles of 37.2D and 37.7H, which is related to the recognition of functional constraints in epitopes (Carr *et al.*, 2024; Aw *et al.*, 2025).

Additional evidence has been obtained through recent experimental studies of escape mapping. It has been established through deep mutational scanning that antibodies binding different sites of the LASV glycoprotein have unique escape maps where certain epitopes are prone to many escape mutations whereas other epitopes are very conserved as substitutions impair glycoprotein stability or mediate entry into host cells (Carr *et al.*, 2024). Similarly, Perrett *et al.* (2022) showed that many efficient antibodies neutralizing LASV use conformational or quaternary epitopes involving interaction between GP1 and GP2, positioning of glycans, and preservation of the prefusion trimer. These epitopes can be expected to allow fewer escape mutations than more flexible antigenic sites.

However, despite all of the antibodies used in Robinson *et al.* (2016)'s experiment being derived from individuals who had been infected with Lineage IV of LASV in Sierra Leone, there was broad cross-reactivity between the antibodies and glycoproteins from different LASV lineages, suggesting that these epitopes are conserved to some degree amongst the circulating strains. LASV has significant geographical genetic diversity, as Lineages I-III are predominantly found in Nigeria, and Lineage IV is found primarily in Sierra Leone, Guinea, and Liberia (Bowen *et al.*, 2000; Andersen *et al.*, 2015). The naturally occurring differences in amino acid sequences between these lineages are expected to play a role in how often escape mutations occur (Carr *et al.*, 2024; Zapata & Salvato, 2013). The lineage-specific nature of the predicted escape mutations is thus more reflective of genetic divergence constrained by the protein's functionality rather than the difference in source of the antibodies.

The identified patterns of virus escape from antibodies will significantly impact the design of antibody therapies against LASV. The antibodies that target structurally conserved and functionally constrained epitopes, especially those included in the GPC-B competitive group, should offer more resistance against virus escape as the number of mutations necessary for maintaining the protein's functionality while escaping antibodies seems to be limited (Carr *et al.*, 2024; Perrett *et al.*, 2022). Conversely, the antibodies that target more mutable epitopes, like 25.10C that is included in the GPC-A group, can be applied in combination with other antibodies that target different antigenic sites in order to minimise the chances for virus escape. This approach aligns with the current approach that focuses on creating broadly reactive antibody cocktails that target different antigenic sites at once (Robinson *et al.*, 2016; Perrett *et al.*, 2022; Carr *et al.*, 2024). Rational design of Arevirumab-3 is one such case wherein the cocktail is designed using GPC-C antibody 8.9F along with GPC-A antibody 12.1F and GPC-B antibody 37.2D for broad coverage by using antibodies which target different epitopes of the glycoprotein (Cross *et al.*, 2023).

6 Conclusion

This work explored how mutations in the LASV glycoprotein complex impact the processes of viral entry and antibody evasion, emphasizing the fact that viral evolution is driven by the need for balancing the ability to infect a cell with the adaptation to immune pressure from the host. It was demonstrated via the analysis of experimentally validated mutational data along with the use of machine learning approaches that the functional outcome of an amino acid substitution depends on its specific position, amino acid involved in the substitution and antibody type. The better accuracy of XGBoost compared to linear regression and single-tree model suggests that the connection between glycoprotein mutations and phenotypic changes is non-linear, and the same substitutions have different impacts depending on the place of their placement within the glycoprotein and substitutions at one place have different phenotypic consequences. It was also observed that LASV glycoprotein evolution is confined to a rigid yet

flexible space where there is an almost equal representation of entry competent mutations and non-entry mutations among lineages I-III along with a dominance of conserved over lineage-specific mutations, showing that viral evolution takes place on the conserved glycoprotein structure. In the case of antibodies, it was also observed that escape from antibodies was strictly epitope-dependent and antibodies 25.10C showed the most flexible escape landscape, followed by antibodies 37.2D and 37.7H showing the most restricted escape landscapes, signifying that escape-resistant antibodies target conserved and rigid epitopes. Though the models used for prediction were designed to use only three features mutation position, wild-type residue identity, and mutant residue identity, the models succeeded in capturing biologically relevant trends in relation to viral entry and escape from antibodies. Other features like glycosylation, residue conservation, solvent accessibility, physical and chemical properties, and epitope topology can also be included in future models to enhance performance as well as to gain more insight into virus fitness and immune evasion.

Together, these results highlight the importance of combining machine learning and experimental mutation studies to identify functionally important mutations, gain further insights into the evolutionary pressures responsible for the diversity of LASV glycoproteins, and develop rationally designed broadly neutralizing vaccines, antibody treatments, and genomic approaches for the detection of genetically diverse populations of Lassa virus.

Author Contributions

Oluremi I. Ajayi: Conceptualization, Data curation, Investigation, Formal analysis, Methodology, Data interpretation, Writing original draft.

Blessing O. Olorunfemi: Software, Formal analysis, Methodology, Visualization, Writing original draft, Writing–review & editing.

Jolly A. Adole: Investigation, Validation, Methodology, Writing review & editing.

Amoge C. Ogu: Data curation, Investigation, Validation, Writing review & editing.

Ifeoluwa O. Bejide: Writing review & editing.

Blossom Akinlawo: Writing review & editing.

Christian T. Happi: Conceptualization, Funding acquisition, Writing review & editing.

Chinedu A. Ugwu: Conceptualization, Supervision, Funding acquisition, Project administration, Methodology, Validation, Writing review & editing.

Declarations

Ethical approval

None of the authors involved in the writing of this manuscript has conducted any research involving human participants or animals.

Consent to participate

Not applicable.

Consent to publish

The final version of the manuscript has been reviewed and approved by all authors for publication.

Competing interests

The authors declare no competing interests.

Data availability

The datasets used and analysed during the current study are publicly available at https://github.com/dms-vep/LASV_Josiah_GP_DMS/tree/main/results/filtered_antibody_escape_CSVs and on the NCBI GenBank database with the corresponding accession numbers are NC_004297, KM822128, GU481068 and GU481072.

Funding source

This study was funded by the Bill and Melinda Gates Foundation through the GIISER Project ARIS3-1V-036304 (CTH and CAU). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

We wish to thank Tim Scott for helping with proofreading this manuscript to improve the clarity of the text.

REFERENCES

- Aloke, C., Obasi, N. A., Aja, P. M., Emelike, C. U., Egwu, C. O., Jeje, O., Edeogu, C. O., Onisuru, O. O., Orji, O. U., & Achilonu, I. (2023). Combating Lassa fever in west African sub-region: Progress, challenges, and future perspectives. *Viruses*, 15(1), 146. <https://doi.org/10.3390/v15010146>
- Andersen, K. G., Shapiro, B. J., Matranga, C. B., Sealfon, R., Lin, A. E., Moses, L. M., Folarin, O. A., Goba, A., Odi, I., Ehiane, P. E., Momoh, M., England, E. M., Winnicki, S., Branco, L. M., Gire, S. K., & Sabeti, P. C. (2015). Clinical sequencing uncovers origins and evolution of Lassa virus. *Cell*, 162(4), 738-750. <https://doi.org/10.1016/j.cell.2015.07.020>
- Asnicar, F., Thomas, A. M., Passerini, A., Waldron, L., & Segata, N. (2024). Machine learning for microbiologists. *Nature Reviews. Microbiology*, 22(4), 191–205. <https://doi.org/10.1038/s41579-023-00984-1>
- Aw DZH, Zhang DX, Vignuzzi M. Strategies and efforts in circumventing the emergence of antiviral resistance against conventional antivirals. *NPJ Antimicrob Resist*. 2025 Jun 9;3(1):54. doi: 10.1038/s44259-025-00125-z. PMID: 40490516; PMCID: PMC12149321.
- Ayulani, I. D., Yunawan, A. M., Prihutaminingsih, T., Sarwinda, D., Ardaneswari, G., & Handari, B. D. (2023). Tree-based ensemble methods and their applications for predicting students' academic performance. *International Journal on Advanced Science, Engineering and Information Technology*, 13(3), 919–927. <https://doi.org/10.18517/ijaseit.13.3.16880>
- Bowen, M. D., Rollin, P. E., Ksiazek, T. G., Hustad, H. L., Bausch, D. G., Demby, A. H., Bajani, M. D., Peters, C. J., & Nichol, S. T. (2000). Genetic diversity among Lassa virus strains. *Journal of Virology*, 74(15), 6992–7004.

625 <https://doi.org/10.1128/JVI.74.15.6992-7004.2000>

626 Bowyer, S., Allen, D. J., & Furnham, N. (2025). Unveiling the ghost: machine learning's
 627 impact on the landscape of virology. *The Journal of General Virology*, 106(1).
 628 <https://doi.org/10.1099/jgv.0.002067>

629 Burton, T. D., & Eyre, N. S. (2021). Applications of deep mutational scanning in
 630 virology. *Viruses*, 13(6), 1020. <https://doi.org/10.3390/v13061020>

631 Bushuiev, A., Bushuiev, R., Sedlar, J., Pluskal, T., Damborsky, J., Mazurenko, S., & Sivic, J.
 632 (2024). Revealing data leakage in protein interaction benchmarks. In *arXiv [cs.LG]*.
 633 <https://doi.org/10.48550/arXiv.2404.10457>

634 Carr, C. R., Crawford, K. H. D., Murphy, M., Galloway, J. G., Haddox, H. K., Matsen, F. A.,
 635 4th, Andersen, K. G., King, N. P., & Bloom, J. D. (2024). Deep mutational scanning reveals
 636 functional constraints and antibody-escape potential of Lassa virus glycoprotein
 637 complex. *Immunity*, 57(9), 2061–2076.e11. <https://doi.org/10.1016/j.immuni.2024.06.013>

638 Cheng, H.-Y., French, C. E., Salam, A. P., Dawson, S., McAleenan, A., McGuinness, L. A.,
 639 Savović, J., Horby, P. W., & Sterne, J. A. C. (2022). Lack of evidence for ribavirin treatment
 640 of Lassa fever in systematic review of published and unpublished Studies1. *Emerging*
 641 *Infectious Diseases*, 28(8), 1559–1568. <https://doi.org/10.3201/eid2808.211787>

642 Cheng, P., Mao, C., Tang, J., Yang, S., Cheng, Y., Wang, W., Gu, Q., Han, W., Chen, H., Li,
 643 S., Chen, Y., Zhou, J., Li, W., Pan, A., Zhao, S., Huang, X., Zhu, S., Zhang, J., Shu, W., &
 644 Wang, S. (2024). Zero-shot prediction of mutation effects with multimodal deep representation
 645 learning guides protein engineering. *Cell Research*, 34(9), 630–647.
 646 <https://doi.org/10.1038/s41422-024-00989-2>

647 Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of*
 648 *the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*,
 649 785–794.

650 Cross, R. W., Heinrich, M. L., Fenton, K. A., Borisevich, V., Agans, K. N., Prasad, A. N.,
 651 Woolsey, C., Deer, D. J., Dobias, N. S., Rowland, M. M., Lathigra, R., Borrega, R., Geisbert,
 652 J. B., Garry, R. F., Branco, L. M., & Geisbert, T. W. (2023). A human monoclonal antibody
 653 combination rescues nonhuman primates from advanced disease caused by the major lineages
 654 of Lassa virus. *Proceedings of the National Academy of Sciences of the United States of*
 655 *America*, 120(34), e2304876120. <https://doi.org/10.1073/pnas.2304876120>

656 Dadonaite, B., Brown, J., McMahon, T. E., Farrell, A. G., Figgins, M. D., Asarnow, D.,
 657 Stewart, C., Lee, J., Logue, J., Bedford, T., Murrell, B., Chu, H. Y., Veessler, D., & Bloom, J.
 658 D. (2024). Spike deep mutational scanning helps predict success of SARS-CoV-2
 659 clades. *Nature*, 631(8021), 617–626. <https://doi.org/10.1038/s41586-024-07636-1>

660 Elste, J., Saini, A., Mejia-Alvarez, R., Mejía, A., Millán-Pacheco, C., Swanson-Mungerson,
 661 M., & Tiwari, V. (2024). Significance of artificial intelligence in the study of virus-host cell
 662 interactions. *Biomolecules*, 14(8), 911. <https://doi.org/10.3390/biom14080911>

663 Ficenec, S. C., Percak, J., Arguello, S., Bays, A., Goba, A., Gbakie, M., Shaffer, J. G., Emmett,
 664 S. D., Schieffelin, J. S., & Bausch, D. (2020). Lassa fever induced hearing loss: The neglected
 665 disability of hemorrhagic fever. *International Journal of Infectious Diseases: IJID: Official*

666 *Publication of the International Society for Infectious Diseases*, 100, 82–87.
667 <https://doi.org/10.1016/j.ijid.2020.08.021>

668 Freschlin, C. R., Fahlberg, S. A., & Romero, P. A. (2022). Machine learning to navigate fitness
669 landscapes for protein engineering. *Current Opinion in Biotechnology*, 75(102713), 102713.
670 <https://doi.org/10.1016/j.copbio.2022.102713>

671 Hastie, K. M., Zandonatti, M. A., Kleinfelter, L. M., Heinrich, M. L., Rowland, M. M.,
672 Chandran, K., Branco, L. M., Robinson, J. E., Garry, R. F., & Saphire, E. O. (2017). Structural
673 basis for antibody-mediated neutralization of Lassa virus. *Science (New York,*
674 *N.Y.)*, 356(6341), 923–928. <https://doi.org/10.1126/science.aam7260>

675 Li, Y. (2023). Genetic basis underlying Lassa fever endemics in the Mano River region, West
676 Africa. *Virology*, 579, 128–136. <https://doi.org/10.1016/j.virol.2023.01.006>

677 Loureiro, M. E., D’Antuono, A., & López, N. (2019). Virus-Host interactions involved in Lassa
678 virus entry and genome replication. *Pathogens*, 8(1), E17.
679 <https://doi.org/10.3390/pathogens8010017>

680 Lundberg, S. M., Erion, G., Chen, H., *et al.* (2020). From local explanations to global
681 understanding with explainable AI for trees. *Nature Machine Intelligence*, 2, 56–67.

682 Michael, R., Kæstel-Hansen, J., Mørch Groth, P., Bartels, S., Salomon, J., Tian, P., Hatzakis,
683 N. S., & Boomsma, W. (2024). A systematic analysis of regression models for protein
684 engineering. *PLoS Computational Biology*, 20(5), e1012061.
685 <https://doi.org/10.1371/journal.pcbi.1012061>

686 Newby, M. L., Allen, J. D., & Crispin, M. (2024). Influence of glycosylation on the
687 immunogenicity and antigenicity of viral immunogens. *Biotechnology Advances*, 70(108283),
688 108283. <https://doi.org/10.1016/j.biotechadv.2023.108283>

689 Pennington, H. N., & Lee, J. (2022). Lassa virus glycoprotein complex review: insights into its
690 unique fusion machinery. *Bioscience Reports*, 42(2), BSR20211930.
691 <https://doi.org/10.1042/BSR20211930>

692 Perrett, H., Brouwer, P., Hurtado, J., Newby, M., Burger, J., Liu, L., Bouhuijs, J., Gibson, G.,
693 Messmer, T., Schieffelin, J., Antanasijevic, A., Boons, G.-J., Crispin, M., Sanders, R., Briney,
694 B., & Ward, A. (2022). Structural conservation of Lassa virus glycoproteins and recognition
695 by neutralizing antibodies. In *bioRxiv*. <https://doi.org/10.1101/2022.09.26.509601>

696 Reeb, J., Wirth, T., & Rost, B. (2020). Variant effect predictions capture some aspects of deep
697 mutational scanning experiments. *BMC Bioinformatics*, 21(1), 107.
698 <https://doi.org/10.1186/s12859-020-3439-4>

699 Robinson, J. E., Hastie, K. M., Cross, R. W., Yenni, R. E., Elliott, D. H., Rouelle, J. A.,
700 Kannadka, C. B., Smira, A. A., Garry, C. E., Bradley, B. T., Brannan, J. M., Olinger, G. G.,
701 Whitmer, S. L. M., Saphire, E. O., & Garry, R. F. (2016). Most neutralizing human monoclonal
702 antibodies target novel epitopes requiring both Lassa virus glycoprotein subunits. *Nature*
703 *Communications*, 7, 11544. <https://doi.org/10.1038/ncomms11544>

704 Shahin, O. R., Ibrahim, M. N., Alanazi, A., Alharithi, F. S., Alruwaili, Y., Alzahrani, A. A., &
705 El Azab, E. F. (2026). Predicting genetic evolution of viruses to identify suitable vaccines using

706 artificial intelligence. *Scientific Reports*, 16(1), 5512. [https://doi.org/10.1038/s41598-026-](https://doi.org/10.1038/s41598-026-35143-y)
707 35143-y

708 Uppala, P. K., Karanam, S. K., Kandra, N. V., & Edhi, S. (2025). Lassa fever: A comprehensive
709 review of virology, clinical management, and global health implications. *World Journal of*
710 *Virology*, 14(3), 108405. <https://doi.org/10.5501/wjv.v14.i3.108405>

711 Yang, Z., Sudjianto, A., Li, X., & Zhang, A. (2024). Inherently interpretable tree ensemble
712 learning. In *arXiv [stat.ML]*. <https://doi.org/10.48550/arXiv.2410.19098>

713 Zapata, J. C., & Salvato, M. S. (2013). Arenavirus variations due to host-specific
714 adaptation. *Viruses*, 5(1), 241–278. <https://doi.org/10.3390/v5010241>