Supplementary Material for Generative AI in the Advancement of Viral Therapeutics for Predicting and Targeting Immune-Evasive SARS-CoV-2 Mutations

[1]Prem Singh Bist, [2,*]Hilal Tayara, and [1,3,*]Kil To Chong

Availability of Supplementary Tables: The detailed information for Tables S5 to S11 is available on GitHub at: data/Supplementary_Tables.pdf. Hyphens (-) within table cells indicate missing data that can be accessed through GitHub.

A. Extraction of embedded features using SARS-Escape Network

To characterize the encoded properties of both natural and generated spike protein sequences, we leveraged the feature learning capabilities of the SARS-Escape network [1]. We provided the network with an input file of generated spikes in CSV format (see dataset at Zenodo: data/Generated_spikes) [2]. The feature learning model, pre-trained on a vast dataset of SARS-CoV-2 spike proteins using a bi-directional LSTM architecture (see "Sars-escape network for escape prediction of SARS-COV-2") [1], extracted feature vectors for each sequence. These embedded features, having the shape of $(N, 512)$, where N is the number of spikes, were saved in NumPy [3] files for further analysis (see dataset at Zenodo: Generated_spikes/embed_features). Subsequently, these feature files were fed into a t-SNE algorithm for dimensionality reduction and visualized as a scatter plot using Matplotlib (Fig. S1).

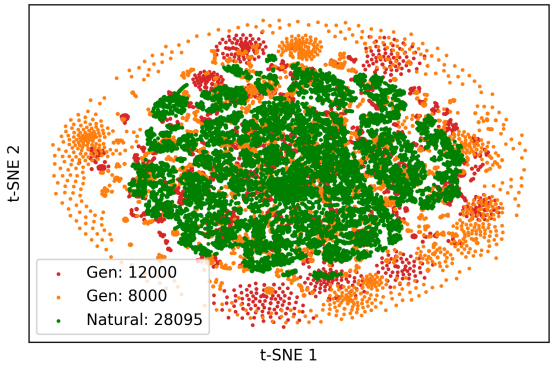

Fig. S1: **Exploration of Generated Spike Protein Landscape without filtering.** This scatter plot visualizes the embedded features of both natural (green) and unfiltered generated (red, orange) spike protein sequences extracted by the SARS-Escape network. The dispersed distribution of generated sequences compared to the tighter cluster of natural sequences suggests that not all generated spikes resemble the natural protein space. This highlights the need for filtration and validation to identify truly "escape" spikes with capabilities similar to natural ones.

B. Identity of Generated Spike Proteins across Training Steps

The scatter plot (Fig. S2) reveals the identity scores of generated spike protein sequences compared to both training (orange) and validation (blue) Blast databases across training steps. The regression line fit over sequence identity values of generated sequences across different training steps showed consistency, suggesting a stable generation. The X-axis represents training steps, indicating the progression of the learning process. The Y-axis represents sequence identity score, reflecting the similarity between generated sequences and natural sequences in the respective databases.

C. Multiple Sequence alignment of generated spikes

Understanding the resemblance between our generated spike protein sequences and their natural SARS-CoV-2 counterparts is crucial for evaluating their potential as escape mutants. To achieve this, we adopted the established blastp(4) to perform multiple sequence alignments. Our workflow involved constructing two essential data structures: a reference database containing natural SARS-CoV-2 spike protein sequences and

This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2020R1A2C2005612) and (No. 2022R1G1A1004613) and in part by the Korea Big Data Station (K-BDS) with computing resources including technical support.

Prem Singh Bist, Hilal Tayara, and Kil To Chong are with [1]Department of Electronics & Information Engineering, [2]School of International Engineering & Science, [3]Advances Electronics & Information Research Center, Jeonbuk National University, Jeonju 54896, Jeollabuk-do, South Korea (e-mail: premsing212@jbnu.ac.kr; Hilaltayara@jbnu.ac.kr; kitchong@jbnu.ac.kr).

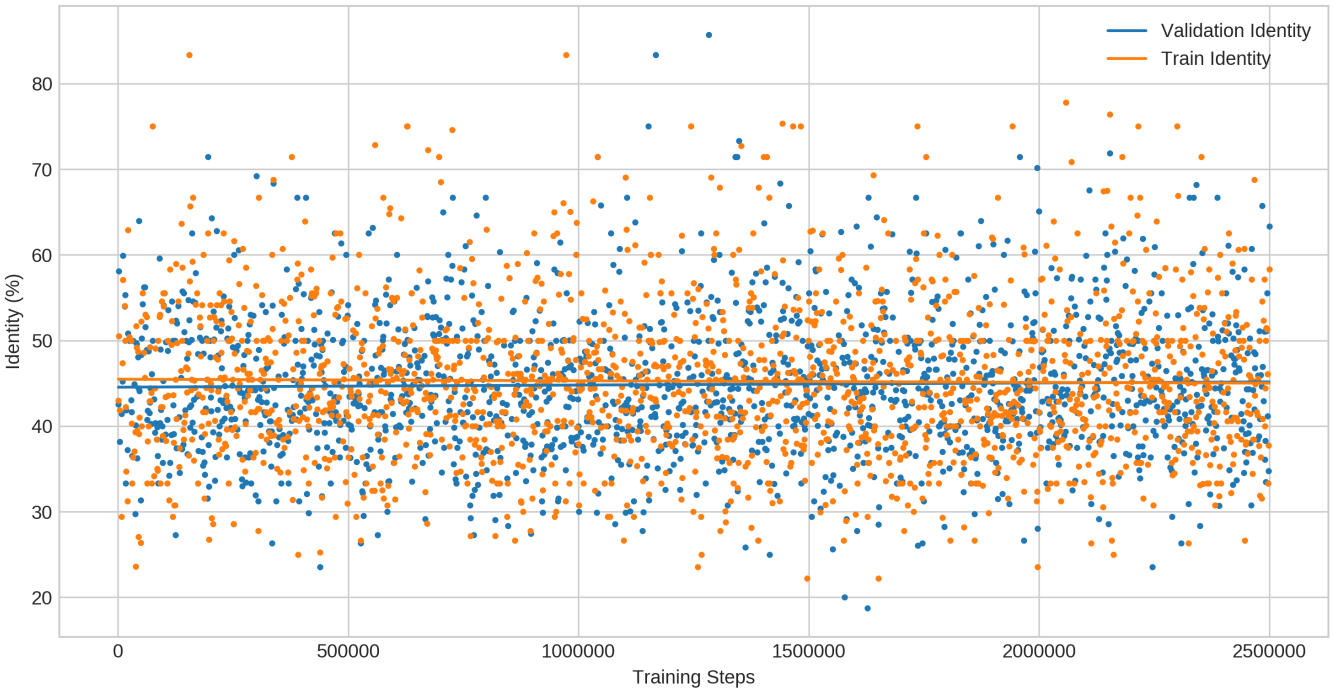

Fig. S2: **Exploration of Identity of Generated Spike Protein Sequences**

a query set comprised of generated spike protein segments, both formatted in Fasta files. By feeding these into blastp, we effectively launched a search for sequence similarities between the generated proteins (queries) and the natural spike protein database (subjects). Considering the 40-50% average identity between generated and natural spike sequences, we opted for the BLOSUM45 substitution matrix, known for its effectiveness with distantly related protein sequences. The resulting blastp alignments are readily available in the data/ directory for further analysis. The results can be executed using the following command:

```
blastp -db {SARS_BLAST_DB}
-max_target_seqs 1 -outfmt "10 qseqid
sseqid qstart qend sstart send nident
score evalue pident qseq sseq" -evalue
{Evalue} -matrix BLOSUM45 -query
{query.fasta}
```

D. Shannon Entropy Computation for Natural and Generated Spikes

To analyze sequence conservation, we aligned generated spike sequences with a natural SARS-CoV-2 spike protein database using Blastp [4]. We then calculated Shannon entropy for each aligned position of both sequences using the following formula:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_2(p(x_i)) \quad (1)$$

where $H(X)$ is the entropy, $p(x_i)$ is the probability of residue x_i in the spike, and the sum is taken over all unique residues in the sequence. The negative sign is used to ensure a positive entropy value.

E. Effect of Sample Size on Escape Sequence Generation and Escape Prediction Performance

To evaluate whether the generated sequences accurately exhibited escape behavior, we assessed the performance of the Sars-escape network after augmenting it with varying numbers of these sequences. The Sars-escape network is specifically designed to differentiate between escape and non-escape sequences. Therefore, if the generated sequences possess genuine escape potential, augmenting the network with them should enhance its escape prediction capability. Conversely, if these sequences lack escape capability, their inclusion in the training process would likely diminish the network's performance. Our generative model produced a substantial number of escape sequences, but it also generated a significant number of non-escape sequences. To isolate the true escape sequences, we employed a filtering process using the pretrained Sars-escape network. Sequences with prediction scores greater than 0.50 were classified as escape sequences, while those below the threshold were deemed non-escape sequences.

To validate the effectiveness of this filtering process in enhancing the reliability of escape sequences, we conducted a comparative analysis. We retrained the Sars-escape network [5] using two distinct sets of augmented sequences:

one containing only the filtered escape sequences, and the other incorporating both filtered and non-filtered sequences. Sequences fed into the pre-trained Sars-escape network [6] output a probability score. Sequences with a threshold score greater than 0.5 are considered filtered escape sequences. By comparing the performance of these two models, we aimed to determine whether filtering indeed improves the reliability and accuracy of escape sequence identification. The network's escape prediction capability was superior when augmented with filtered escape sequences compared to using non-filtered sequences.

Augmenting with 730 filtered sequences elevated the network's AUC scores to 0.767 from the original score of 0.70 on the Greaney dataset [7]. In contrast, augmenting with all 12,000 non-filtered sequences resulted in virtually no change in AUC scores(Table: S1). Augmenting with 871 filtered sequences yielded a slight improvement in AUC scores to 0.7197, whereas augmenting with all 8,000 non-filtered sequences led to a decline to 0.668 (Table: S2).

We assessed escape forecast capability by augmenting with combined samples from both the 8,000 and 12,000 generated sequences, resulting in an AUC score improvement to 0.746(Table: S3). Generating a larger number of samples (16,000) significantly diminished the network's prediction capability, with the AUC score dropping to 0.662 (Table: S4) compared to the original model score of 0.70 [5]. These results suggest that generating escape sequences through our model necessitates a delicate balance. An excessive number of sequences can introduce false positives, compromising prediction accuracy.

TABLE S1

Performance of Sars-Escape Network in Identifying Escape Mutations: From an initial set of 12,000 generated sequences, a filtering process with a threshold of 0.50 identified 730 sequences.

Metric	Unfiltered	Filtered
AUC	0.703	0.7672
Loss	0.7506	0.8003
Number of Augmented Sequences	12000	730

TABLE S2

Performance of Sars-Escape Network in Identifying Escape Mutations: From an initial set of 8000 generated sequences, a filtering process with a threshold of 0.50 identified 871 sequences.

Metric	Unfiltered	Filtered
AUC	0.6681	0.7197
Loss	0.8608	0.8003
Number of Augmented Sequences	8000	871

F. Decoding the Language of Spikes: Integer Encoding Scheme for SARS-CoV-2 Sequences

The SARS-COV-2 spike sequence input to the discriminator network was integer encoded with the amino acid symbols and their corresponding representative encoding values during the training. The full supplementary table data for Table S5 can be found on the Github repository.

TABLE S3

Performance of Sars-Escape Network in Identifying Escape Mutations: Combining augmented sequences from 12,000 and 8,000 samples resulted in 20,000 total sequences, of which 1,601 passed the filter ($Threshold > 0.50$).

Metric	Filtered
AUC	0.7465
Loss	0.9245
Number of Augmented Sequences	1601

TABLE S4

Performance of Sars-Escape Network in Identifying Escape Mutations: From an initial set of 16,000 generated sequences, a filter with a threshold greater than 0.50 identified 782 sequences for further analysis.

Metric	Filtered
AUC	0.6621
Loss	1.2351
Total Sequences	16000
Number of Augmented Sequences	782

TABLE S5: Amino Acid Encodings

Amino Acid	Description	Encoding value
A	Alanine	1
R	Arginine	2
N	Asparagine	3
-	-	-

G. Model Parameters

This section outlines the specific parameters of the discriminator and generator networks. Detailed Tables can be found in the GitHub repository. The key tables are listed below:

- **Table S6:** Key Architecture Parameters of Discriminator and Generator Networks.
- **Table S7:** Full Parameter Configuration of the Discriminator Network.
- **Table S8:** Full Parameter Configuration of the Generator Network.

H. Preparing Data for Rigorous Evaluation

To assess the performance of the SARS-Escape model [5] following its retraining with generated spike sequences, we employed three distinct datasets: the Validation dataset, the Greaney dataset [7], and the Baum dataset [8].

h1. Validation Dataset

The validation dataset served two crucial purposes. First, it enabled us to measure the similarity between generated

spike protein sequences and natural ones during our generative model's training phase. Second, it played a key role in assessing the performance of the SARS-Escape model [5] after retraining with the generated spikes. A detailed description of the data preparation process for both the training and validation dataset can be found in the section on Data Preparation for Feature Learning Network [5].

h2. Greaney Dataset

The Greaney dataset [7] provides a comprehensive resource for SARS-CoV-2 RBD mutations that evade antibody binding. It systematically lists mutations arising during viral growth under antibody pressure. Curated through deep mutational scanning, the dataset specifically assesses how all amino acid substitutions within the RBD impact binding to human monoclonal antibodies.

h3. Baum Dataset

The Baum dataset [8] comprises novel spike protein mutants that evade neutralization by specific antibodies. These mutants identified through in vitro experiments involving culturing a pseudo virus against four potent antibodies, exhibit resistance to one or more of these antibodies. A total of 19 escape mutants were identified in this way. All other 24,168 mutants in the dataset, generated by randomly mutating a single residue at specific positions in the wild-type SARS-CoV-2 Wuhan-Hu-1 sequence [9] were categorized as non-escape mutants.

I. Escape Strength Analysis of In Silico Spikes: Single-Residue Mutations

I1. Escape Strength of 464 highly probable escape mutants

Among the 88,920 analyzed in silico spike segments, 464 mutants exceeding an escape strength threshold of 0.95 are presented in the following table, ranked by decreasing escape strength. The mutated spike segments were constructed using the Wuhan-based sequence [9] as the template, and their scores were evaluated using the Sars-Escape network [1]. Detailed escape analysis output records for all mutated segments are available in supplementary table.

TABLE S9: Mutation and Escape Scores for Spike Segments

Id	Spike Segment	Mutation	Escape Score
1	VVLSDELLHAPATVCGPKKS	F515D	0.958565
2	VVLSCELLHAPATVCGPKKS	F515C	0.956172
.	-	-	-
.	-	-	-
463	NGVGYQPYRVVVLSFELLFA	H519F	0.952232
464	NGVGYQPYRVVVLSFELLHQ	A520Q	0.956128

I2. In Silico Escape Potential of Known Escape Mutants: Analysis and Comparison

We conducted a thorough review of the existing literatures [10], [11] to identify established escape mutants. Subsequently, we scrutinized our Insilco-generated escape score database to assess the presence of these known mutants and to evaluate how our model ranked their escape potential. In parallel, we explored potential escape mutants in the vicinity of those with elevated scores in our database. Comprehensive details of these analyses are elucidated in the subsequent tables. Previously confirmed escape mutants [10] identified among the top 464 high-confidence (score exceeding 0.95) mutants are labeled with green color. Full table records are available on supplementary Table (S10) on Github.

TABLE S10: **Insilico Mutated Spike Protein Analysis**

Insilico Mutated Spike	Mutant	Score	Established
EVRQIAPGQTGMIADYNYKL	K417M	0.959268	K417T/N/E
FNCYFPLNSYGFQPTNGVGY	Q493N	0.959997	Q493R/L
-	-	-	-

TABLE S11: **Scores of Known Escape Mutants**

The escape strength predicted for previously documented escape mutants from Source [10] was analyzed(Full data available on GitHub).

Known Escape Mutants	Score
E484Q	0.9716099
Q493H	0.9651634
V483A	0.9405139
N501Y	0.9400368
-	-

J. AUC Scores of Sars-CoV-2 Escape Model: Before and After Synthetic Data Augmentation

We evaluated AUC scores in two scenarios: after synthetic data augmentation and before synthetic data augmentation (Fig. S3).

K. Access to Resources

A. Data Availability

The dataset used in this research is publicly available at [https://doi.org/10.5281/zenodo.10628743].

B. Code Availability

The source code used in this research is publicly available on GitHub at [https://github.com/PremSinghBist/SarsGAN].

L. Declaration of Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

M. Funding

This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. 2020R1A2C2005612) and (No. 2022R1G1A1004613) and in part by the Korea Big Data Station (K-BDS) with computing resources including technical support.

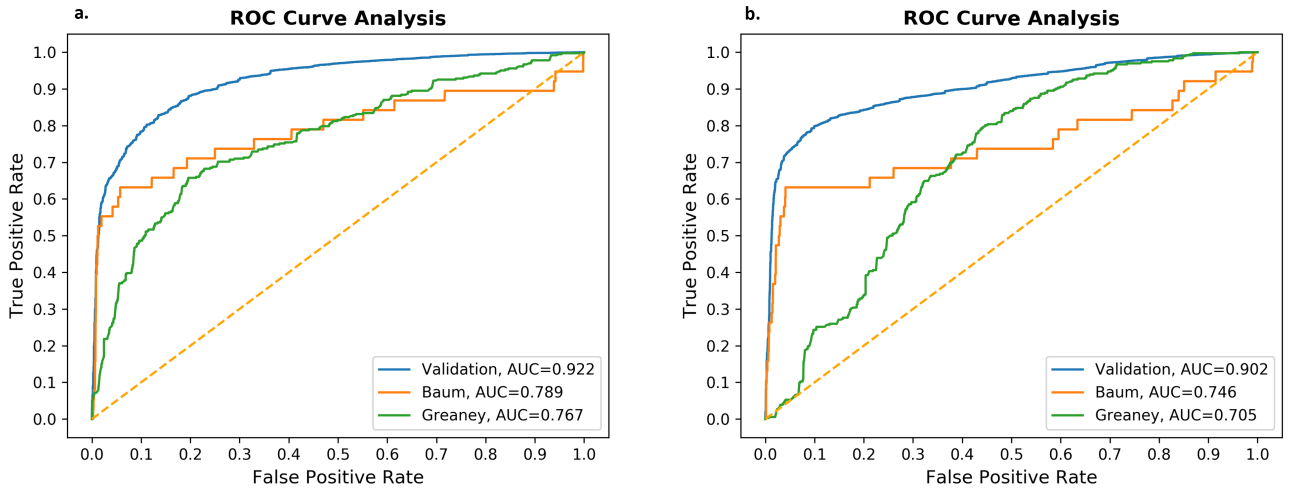

Fig. S3: **a.** AUC Scores After Synthetic Data Augmentation. **b.** AUC Scores Before Synthetic Data Augmentation

ACKNOWLEDGMENT

We extend our sincere gratitude to the authors from the Originating Laboratories who obtained the specimens and to the submission laboratories that generated and shared the genetic sequence data through the GISAID Initiative, upon which this research is founded.

REFERENCES

[1] P. Singh Bist, H. Tayara, and K. To Chong, "Sars-escape network for escape prediction of sars-cov-2," *Briefings in Bioinformatics*, vol. 24, no. 3, p. bbad140, 2023.

[2] P. Singh Bist, "Dataset For Generative Adversarial Network expands SARS-CoV-2 Spike Protein Diversity and Escape Prediction Potential," Feb. 2024. [Online]. Available: https://doi.org/10.5281/zenodo.10628743

[3] C. R. Harris, K. J. Millman, S. J. Van Der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith *et al.*, "Array programming with numpy," *Nature*, vol. 585, no. 7825, pp. 357–362, 2020.

[4] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, 1990.

[5] P. Singh Bist, H. Tayara, and K. To Chong, "Sars-escape network for escape prediction of SARS-COV-2," *Briefings in Bioinformatics*, vol. 24, no. 3, p. bbad140, 04 2023. [Online]. Available: https://doi.org/10.1093/bib/bbad140

[6] Bist Prem Singh, "Sars-CoV-2-Escape-Model," GitHub repository, 2023, retrieved from https://github.com/PremSinghBist/Sars-CoV-2-Escape-Model.

[7] A. J. Greaney, T. N. Starr, P. Gilchuk, S. J. Zost, E. Binshtein, A. N. Loes, S. K. Hilton, J. Huddleston, R. Eguia, K. H. Crawford *et al.*, "Complete mapping of mutations to the sars-cov-2 spike receptor-binding domain that escape antibody recognition," *Cell host & microbe*, vol. 29, no. 1, pp. 44–57, 2021.

[8] A. Baum, B. O. Fulton, E. Wloga, R. Copin, K. E. Pascal, V. Russo, S. Giordano, K. Lanza, N. Negron, M. Ni *et al.*, "Antibody cocktail to sars-cov-2 spike protein prevents rapid mutational escape seen with individual antibodies," *Science*, vol. 369, no. 6506, pp. 1014–1018, 2020.

[9] F. Wu, S. Zhao, B. Yu, Y.-M. Chen, W. Wang, Z.-G. Song, Y. Hu, Z.-W. Tao, J.-H. Tian, Y.-Y. Pei *et al.*, "A new coronavirus associated with human respiratory disease in china," *Nature*, vol. 579, no. 7798, pp. 265–269, 2020.

[10] K. Guruprasad, "Mutations in human sars-cov-2 spike proteins, potential drug binding and epitope sites for covid-19 therapeutics development," *Current research in structural biology*, vol. 4, pp. 41–50, 2022.

[11] R. Wang, J. Chen, K. Gao, and G.-W. Wei, "Vaccine-escape and fast-growing mutations in the united kingdom, the united states, singapore, spain, india, and other covid-19-devastated countries," *Genomics*, vol. 113, no. 4, pp. 2158–2170, 2021.

Prem Singh Bist earned his Bachelor of Computer Engineering from Pokhara University, Nepal, in 2010, and an MTech in Computer Science and Engineering from Dr. A.P.J. Abdul Kalam Technical University, India, in 2016. With over a decade of rich experience, he has climbed the professional ladder, holding positions such as Software Engineer, Senior Software Engineer, and Technical Architect at renowned software firms, notably including Impetus Infotech. Currently connected with Guru Technology in Nepal, Bist has cultivated extensive expertise in Enterprise Software Development, with a focus on Java, Big Data, and Cloud Computing technologies for more than 10 years. He is furthering his education by pursuing a PhD in Bioinformatics and Natural Language Processing at the Department of Electronics and Information Engineering at Jeonbuk National University, South Korea. His research is primarily dedicated to leveraging artificial intelligence in drug discovery, aiming to forge significant advancements in this frontier area.

Tayara Hilal received his PhD in Electronics and Information Engineering from Jeonbuk National University, South Korea. He is currently an assistant professor at the School of International engineering and science at Jeonbuk National University, South Korea. His research fields are Bioinformatics, Computational Biology,Deep Learning and Image Processing.

Kil To Chong received PhD from Texas A and M University in Mechanical Engineering. Currently, he is a Professor in the Electronics Engineering Division, president of Electronics and IT New Technologies Research Center in Jeonbuk National University, South Korea. His Interested fields are Bioinformatics, Computational Biology, Deep Learning, and Medical Image Processing.