AN UNSUPERVISED FRAMEWORK FOR COMPARING SARS-COV-2 PROTEIN SEQUENCES USING LLMS

Sayantani B. Littlefield
Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, IL 61801
basu9@illinois.edu

Roy H. Campbell
Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, IL 61801
rhc@illinois.edu

ABSTRACT

The severe acute respiratory system coronavirus 2 (SARS-CoV-2) pandemic led to 700 million infections and 7 million deaths worldwide. While studying these viruses, scientists developed a large amount of sequencing data that was made available to researchers. Large language models (LLMs) are pre-trained on large databases of proteins and prior work has shown its use in studying the structure and function of proteins. This paper proposes an unsupervised framework for characterizing SARS-CoV-2 sequences using large language models. First, we perform a comparison of several protein language models previously proposed by other authors. This step is used to determine how clustering and classification approaches perform on SARS-CoV-2 sequence embeddings. In this paper, we focus on surface glycoprotein sequences, also known as spike proteins in SARS-CoV-2 because scientists have previously studied their involvement in being recognized by the human immune system. Our contrastive learning framework is trained in an unsupervised manner, leveraging the Levenshtein distance from pairwise alignment of sequences when the contrastive loss is computed by the Siamese Neural Network. The final part of this paper focuses on a comparison with a previous approach on a test dataset containing data from the latter part of the pandemic. In the prediction of emerging variants, the proposed LLM-based approach shows an improvement of 0.1914 in terms of the adjusted rand index clustering compared to a previously proposed approach. This shows the potential of applying large-language models to this field.

1 Introduction

The evolution of pandemics like COrona VIrus Disease (COVID-19) have caused widespread infection and deaths worldwide [1]. Pandemics typically involve health officials introducing a set of mitigation measures while researchers develop ways to study vaccines and design drugs for the benefit of society [2]. In this process, there is a deluge of sequencing data from pandemics. This information is stored publicly by the National Institute of Heath — the National Center of Biotechnology Information (https://www.ncbi.nlm.nih.gov).

In this paper, we are interested in studying SARS-CoV-2 surface glycoprotein sequences because they are recognized by the human immune system [3, 4]. Protein language models are a set of models applicable to protein sequences which generate condensed representations in the form of embedding vectors. A comparison study on chronologically ordered SARS-CoV-2 sequences will help understand the performance of these embedding vectors relative to the variant labels [5].

Following our comparison study, we extend our previously proposed framework for supervised contrastive learning to unsupervised conditions. Contrastive learning is used in machine learning to identify differences among different samples. Such a comparison will eventually help with identification of clusters of variants [6]. We hypothesize that similarities in protein sequences that are reflected in their corresponding n-dimensional LLM sequence embeddings can be used to train a contrastive framework and observe similarities and differences in protein sequences.

The rest of this paper is structured as follows: Section 2 provides an overview of related work, Section 3 discusses the datasets used, Section 4 covers the methods used in our comparison study and unsupervised contrastive learning framework, Section 5 discusses the results obtained, and Section 6 concludes the paper.

2 Related Work

This section provides a review of existing literature in methods used for analysis of variant data. Computational software like Nextalign and Nextclade [7, 8, 9] applied to labeling SARS-CoV-2 variants rely on sequence alignment, which is computationally expensive. Approaches involving k-mers where k residues are used to form 'words' from a protein sequence to allow for an alignment-free approach. While prior work has also shown that k-mer based approaches perform better than alignment based approaches [10], they are subject to erroneous associations and also lead to high computational complexity in certain cases.

Several other machine learning approaches have been proposed on SARS-CoV-2 data, starting with classification of coronaviruses at various levels of taxa [11, 12]. Brierley and Fowler [13] classified intermediate COVID-19 hosts with random forests where the models were trained on SARS-CoV-2 spike proteins and gene sequences. The authors claimed deriving "evolutionary signatures" from genomic data may be possible, and this could indicate species origin [5].

Basu and Campbell [14] previously proposed a long short-term memory (LSTM) deep learning model that enabled the classification of different SARS-CoV-2 variants. However, the model was trained in 2021 on an early COVID-19 snapshot of twenty variants at the time. The structure of the proposed network required re-training each time newer SARS-CoV-2 variants added to the database.

ProtVec [15] and ProteinBERT [16] were language models in the field before the introduction of large-language models. ProtVec is inspired by word2vec [17]. The model was pre-trained on 546,790 SwissProt sequences. It uses a skip-gram neural network to train positive and negative 3-mers depending on the context. ProteinBERT [16] is an influential paper in this field modeled after BERT [18]. It consists of six blocks of transformers as well as fully connected and convolutional layers. The global attention layers exhibit linear complexity with respect to sequence length, which makes the model adaptable to sequences of different lengths. This model was pre-trained on 106M UniRef90 proteins. Large language models (LLMs) gained popularity in natural language applications with the introduction of transformers [19, 18] and generative pre-trained transformers (GPT) [20]. In the field of viral proteomics, protein LLMs have been used to study protein sequences, including their structure and function.

Protein LLMs have the advantage of being trained on larger context lengths compared to protein LMs. This enables the analysis of a large number of proteins with longer context length. In this paper, we discuss results using two long-context protein LLMs: Genome-Scale Language Model (GenSLM) and Evolutionary Scale Modeling (ESM-2). Genome-Scale Language Model (GenSLM) [21] uses attention to determine evolutionary dynamics. Evolutionary Scale Modeling (ESM) consist of a set of protein transformer models. The first of these is the ESM-1b model which was proposed by Rives et al. [22]. The objective was to apply generative models to protein sequences to determine protein structure and contact prediction using self-attention [23]. ESM-2 is the follow-up research model proposed by Lin et al. [24]. In this paper, we compare four previously published models ProtVec [15], ProteinBERT [16], GenSLM [21], and ESM-2 [24] in order to determine the classification and clustering performance on SARS-CoV-2 spike protein sequences.

Contrastive learning is a strategy used in machine learning where the model learns to distinguish among samples using comparisons. Siamese Neural Networks (SNNs) are a type of contrastive learning model originally used in computer vision [25]. SNNs operate by sharing weights and learn to minimize the distance between similar classes while separating the distance between different classes. In the field of viral genomics, SNN literature is limited. Madan et al. [26] used SNNs on spike protein embeddings to predict protein-protein interaction (PPI) networks for SARS-CoV-2. The authors applied ProteinBERT to generate the embeddings and showed how the properties of proteins can be learned from sequence data. This work is an extension of Tsukiyama et al. [27] where an (LSTM) model was applied to word2vec embeddings to study PPI interactions. From our previous work on using SNNs in supervised settings [28], we observed that SNNs have potential in differentiating various classes and choose to extend it to unsupervised settings in this paper. Our inspiration for comparison in an unsupervised manner also stems from recent similar work [29] on spike proteins that studied the detection of emerging variants in such settings.

In this paper, we first discuss a cursory comparison of various protein language models in a supervised manner, following which we discuss the unsupervised contrastive learning framework.

3 Datasets

SARS-CoV-2 surface glycoprotein sequences were downloaded from the National Center for Biotechnology Information (NCBI), an NIH sequenced data repository. The sequences were filtered based on completeness (complete), type of protein (surface glycoproteins), host (*Homo sapiens*) and country (United States).

3.1 Supervised comparison study

The PANGO lineage information was extracted from the same NCBI dataset, while the lineage classification (Omicron, Alpha, Beta, Delta, Gamma) was extracted from the `https://cov-lineages.org` [30, 31]. The cov-lineages dataset had information about the individual lineages and which variants belonged to each lineage. According to the website, B.1.1.529, BA.1, BA.2, BA.5, XBB, XBB.1, XBB.1.5, and XBB.1.6 belong to the Omicron lineage, B.1.1.7 belongs to the Alpha lineage, B.1.351 belongs to the Beta lineage, B.1.617.2 belongs to the Delta lineage, and P.1 belongs to the Gamma lineage. Following pre-processing, our dataset had 20,580 records. These chronologically ordered records had sample collection dates from January 2020 to February 2024. For reproducibility, it was split chronologically into 80% and 20% for our training and validation sets.

3.2 Unsupervised contrastive learning approach

The PANGO lineage information was extracted from the same NCBI dataset. Following pre-processing, there were 134,333 unique sequences divided into train (80%) and validation (20%). These were chronologically ordered records had sample collection dates from January 2020 to February 2024 with PANGO labels from NCBI. A separate test dataset of 40,143 sequences dated from February 2024 to February 2025 was applied to detecting emerging variants.

4 Methods

In this section, we introduce the methods used in the supervised comparison study and the unsupervised contrastive learning approach.

4.1 Comparison Study

Surface glycoprotein sequences were obtained from NCBI Virus (NIH), a public repository of sequencing data. SARS-CoV-2 sequences were filtered based on country (United States), host (*Homo sapiens*), type of protein (surface glycoproteins), and completeness (complete). Variant labels were extracted from the same NCBI Virus dataset and linked to the lineage classification from `https://cov-lineages.org` [30, 31]. We extracted five main lineages, namely Omicron (B.1.1.529, BA.1, BA.2, BA.5, XBB, XBB.1, XBB.1.5, and XBB.1.6), Alpha (B.1.1.7), Beta (B.1.351), Delta (B.1.617.2), and Gamma (P.1). The labels are obtained from the lineages.

4.2 Contrastive Learning Framework

Contrastive learning models learn by comparing existing classes. Siamese Neural Networks (SNNs) [32, 25] were originally used in computer vision to identify new classes at the inference step based on finding similarity among previously existing classes. We consider an SNN for the contrastive learning framework as previous literature has indicated that many bioinformatics tools currently struggle with the introduction of novel viral sequences [33]. Developing such a tool is useful in the case of viral pandemic data where new variants are on the rise, as it has the ability to handle newer sequences and variants after the pipeline is already trained on data from previously circulating variants. SNNs operate by sharing weights that minimize the distance between similar classes, while separating the distance between different classes. This changes the embeddings in the n-dimensional space while training the network, which we observe clarifies the clusters as indicated by cluster metrics.

Following our comparison study, we extend our previously proposed Siamese Neural Network (SNN) framework [28] for supervised contrastive learning to unsupervised conditions. The model architecture is shown in Figure 1, which shows the mirrored networks with shared weights. The objective function here is to minimize the contrastive loss. We retain the same architecture as empirically determined by Basu et al. [28] for the rest of the experiments. Each network contains three convolutional layers interleaved with maxpool layers. For training the network, a batch size of 256 and a learning rate of 0.0001 were used with Adam as the optimizer [34].

All code was implemented using Python on 1 GPU node and 4 CPUs with the HAL cluster [35].

5 Results and Discussion

5.1 Comparative analysis

In this section, we compare various protein language models like GenSLM [21] and ESM-2 [24] along with models like ProtVec [15] and ProteinBERT [16] on classifying and clustering pandemic viruses like that of SARS-CoV-2 into different lineages. The following metrics are used for this comparison:

- Silhouette Coefficient (SC): This metric measures similarity of points within a cluster and compares it to points belonging to different clusters. It ranges from -1 to +1. More positive values of SC indicate well-clustered data while lower values indicate poorly clustered data. It is represented as $SC = (b-a)/max(a, b)$, where a is the mean intra cluster distance and b is the mean nearest cluster distance.
- Accuracy: This metric is computed using the number of true positives and true negatives out of all samples.
- F-1 score: This is computed as the harmonic mean of precision and recall. This ranges from 0 to 1 where higher values are preferred.

The comparison study discussed in this paper is performed in a supervised setting and is evaluated using the PANGO labels in the NCBI dataset with their corresponding lineages. To enable a fair comparison, the variants were arranged chronologically from the earliest (2020) to the latest occurrence (2024) based on the date of collection of the saliva samples. Training was performed on the first 80% of the data using a multi-layered Dense network, Adam optimizer [34], and a batch size of 512. There were five labels for SARS-CoV-2 based on the corresponding lineages. Training used EarlyStopping and the validation loss was monitored with a patience parameter of 10. The comparison study shows the accuracy and F-1 on the test data (the remaining 20% of the data). The F-1 measure is used because it is the harmonic mean of the precision and the recall metrics. The metrics for four models (ProtVec [15], ProteinBERT [16], GenSLM [21], and ESM-2 [24]) are reported in Table 1.

The maximum lengths of the sequence for the attention-based GenSLM model (including the hierarchical diffusion model) were set to 1280 for SARS-CoV-2 sequences. These are the maximum sequence lengths observed on all sequences in the dataset after pre-processing. Computations were performed on the original embeddings, and t-SNE [36, 37] was only used to reduce the different embeddings to two dimensions for visualization purposes.

Table 1: Comparison of protein LM embeddings on classification

Dataset	Model	val acc	val f1
SARS-CoV-2	ProtVec [15]	0.9446	0.9729
	ProteinBERT [16]	0.9643	0.9803
	GenSLM [21]	**0.9905**	**0.9953**
	ESM-2 [24]	0.9645	0.9829

Table 2: Comparison of protein LM embeddings on clustering

Dataset	Model	SC
SARS-CoV-2	ProtVec [15]	0.2856
	ProteinBERT [16]	0.1517
	GenSLM [21]	0.2196
	ESM-2 [24]	0.2561

The results in Table 1 and Table 2 show that:

1. Large language models perform better than language models on classifying among five different classes of SARS-CoV-2. This is because large language models are trained with more parameters and on larger databases to recognize patterns among the different amino acids at the sequence level.
2. When critically analyzing the performance of the large language models GenSLM and ESM-2, GenSLM seems to perform slightly better compared to ESM-2. The performance of GenSLM can be attributed to the addition of the hierarchical diffusion model on top of the transformer framework because recent literature in bioinformatics has shows that diffusion generative models are capable of learning distributions from complex, high-dimensional samples [38].

3. The clusters of embedding vectors do not cluster well according to the SC standard metrics. This means such clusters cannot be used as-is for identifying emerging variants. This in turn leads us to form an approach to further separate the classes using the proposed contrastive learning framework (Siamese Neural Network). This approach makes the clusters more distinct to study emerging variants.

5.2 Unsupervised contrastive learning framework

This section details the method used in sampling the embedding pairs:

1. **Supervised case:** In the supervised framework, the PANGO labels are known along with the sequences and their embeddings at the time of training. The process of mining pairs and triplets of embeddings is as follows:
   Pairs: In this case, the sequence embeddings and the corresponding PANGO labels are passed to the model. The training sample is marked similar if the two variants are the same, else they are marked dissimilar. Here, a strict comparison at the variant level is considered—for example, embedding pairs from two variants BA.1 and BA.2 are labeled 'dissimilar' while embedding pairs from two variants B.1.1.7 and B.1.1.7 are labeled 'similar'.
   Triplets: In this case, three sequence embeddings and the corresponding PANGO labels are passed to the model at a time. The first embedding of the three is considered the 'anchor' while the remaining two are mined based on the constraint that one PANGO variant is similar to the anchor while the second PANGO variant is dissimilar to the anchor. For example, if the embedding corresponding to a sequence labeled 'B.1.1.7' is mined as the anchor, a corresponding positive or similar example would be an embedding corresponding to another sequence labeled 'B.1.1.7' while a negative or dissimilar example would be an embedding corresponding to a sequence labeled 'BA.2'.

2. **Unsupervised case:** The unsupervised framework is more complex since the PANGO labels are not provided at the time of training the model. Only sequences and their embeddings are available at the time of training. The novelty here is using the property of Levenshtein distance to compute the similarity between the sequences using that to self-label the dataset when the LLM embeddings are provided while training. The Levenshtein distance is defined as the number of steps it takes to transform one sequence into the other. The similarity metric is then computed by calculating the percentage of positions where the residues are identical across all positions, constituting a pairwise alignment of the sequences. This metric is used to determine whether two sequences are similar or not based on two thresholds: (i) positive threshold (pt): if the similarity metric is higher than pt, the sequences are similar, and (ii) negative threshold (nt): if the similarity metric is lower than nt, the sequences are dissimilar. The process of mining pairs and triplets of embeddings is as follows:
   Pairs: In this case, the sequence embeddings and the labels computed based on the thresholds are passed to the model for training. If the similarity between the sequences is greater than pt, then the pair is labeled similar. If the similarity between the sequences is less than nt, then the pair is labeled dissimilar.
   Triplets: In this case, three sequence embeddings and the labels computed based on the thresholds are passed to the model for training. If the similarity between the positive example and the anchor is greater than pt, then the two embeddings are passed to the model along with the similar label. If this constraint is not satisfied, the example is skipped. If the similarity between the negative example and the anchor is lesser than pt, then the two embeddings are passed to the model along with the dissimilar label. If this constraint is not satisfied, the example is skipped.

Following training of the SNN, in the test phase, the SNN performance metrics are generated for the test embedding pairs. The embeddings learned by the SNN are generated by passing the samples through one-half of the network to get the new condensed vector representations.

Results The chronologically ordered dataset is first split into 80% and 20% for the train and test data respectively.

Table 3 shows the complete parameter study conducted for this study varying positive threshold pt, negative threshold nt and the margin m. The silhouette coefficient (SC) for the validation set is reported at the end of each run. Based on the parameter study, the best results are observed on the unsupervised model using triplet loss, which gives a SC of 0.6757.

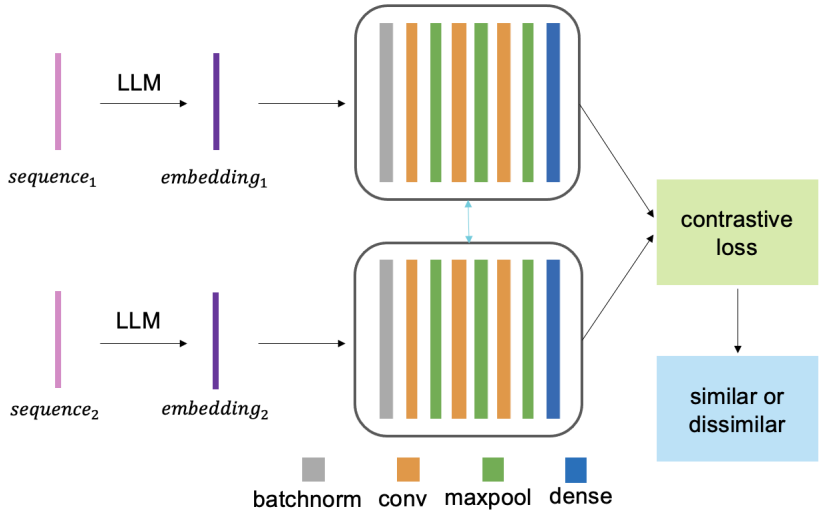

Figure 1: Architecture of contrastive learning framework with pairwise contrastive loss (based on Basu et al. [28])

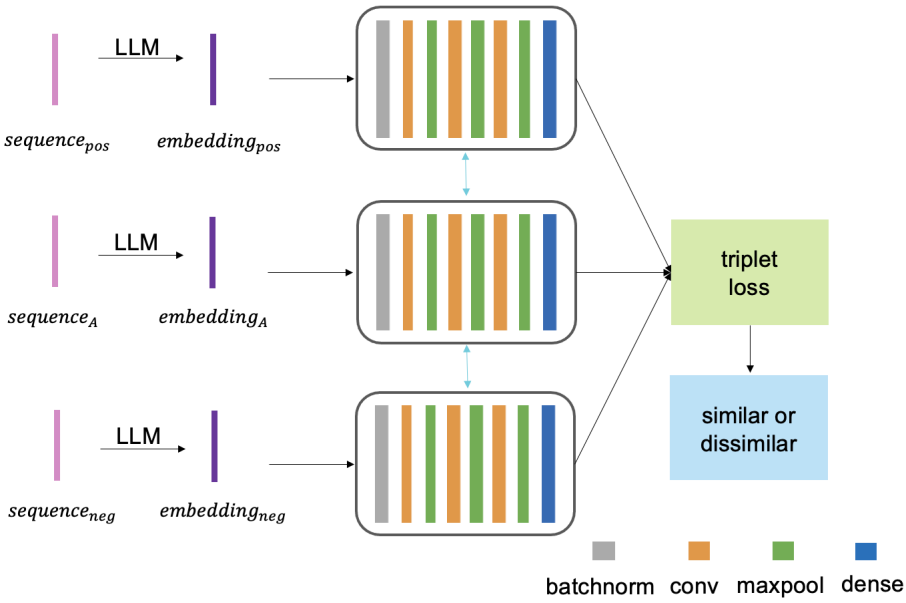

Figure 2: Architecture of contrastive learning framework with triplet loss (based on Basu et al. [28])

Table 3: Parameter study on training the proposed Siamese Neural Network

model	loss	pt	nt	m	SC
unsupervised	pairwise contrastive	0.5	0.5	0.2	0.5919
				0.3	0.3360
				0.4	0.3932
		0.6	0.4	0.2	0.3770
				0.3	0.6174
				0.4	0.4520
		0.7	0.3	0.2	0.4405
				0.3	0.5700
				0.4	0.4332
	triplet	0.5	0.5	0.2	0.5458
				0.3	0.4859
				0.4	0.4324
		0.6	0.4	0.2	0.3608
				0.3	0.4867
				0.4	0.3361
		0.7	0.3	0.2	**0.6757**
				0.3	0.6498
				0.4	0.4561
supervised	pairwise contrastive	0.5	0.5	0.2	0.2436
				0.3	0.2282
				0.4	0.2599
		0.6	0.4	0.2	0.2846
				0.3	0.2772
				0.4	0.2538
		0.7	0.3	0.2	0.2897
				0.3	0.2081
				0.4	0.2131
	triplet	0.5	0.5	0.2	0.2978
				0.3	0.2749
				0.4	0.3229
		0.6	0.4	0.2	0.3264
				0.3	0.3884
				0.4	0.3084
		0.7	0.3	0.2	0.3315
				0.3	0.3239
				0.4	0.3538

Using the best parameters from the parameter study ($pt = 0.7$, $nt = 0.3$, $m = 0.2$), the validation dataset embeddings were visualized before and after to observe how the contrastive learning framework contributes to better clustering of the classes. Figure 3 and Figure 4 show the t-SNE plots of GenSLM embeddings before and after learning. The SC improved from 0.4280 to 0.6757 on the validation set with the predicted clusters, which indicates tighter clustering of the embeddings following the SNN training. A 2-dimensional t-SNE plot [36, 37] was used to observe the different embeddings before and after the Siamese Neural Network was run. The reason for using t-SNE is that the local relationships and subtle structures in the data are revealed using t-SNE, which is why it is used on genomic data [39]. In all such figures, x1 and x2 represent the axes of the 2-dimensional plots for ease of observing the visualizations for these multi-dimensional embeddings.

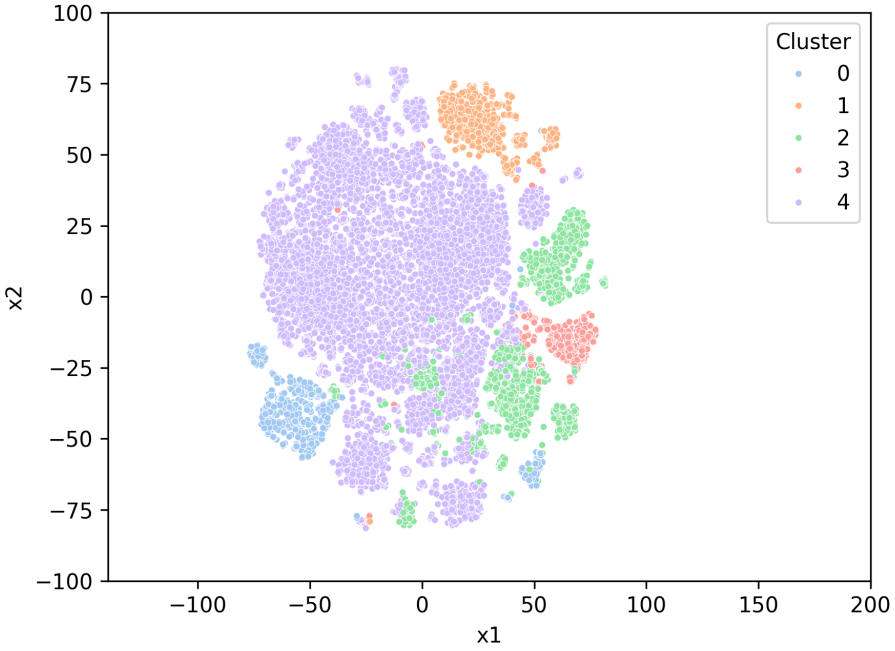

Figure 3: t-SNE embeddings of SARS-CoV-2 variants before learning

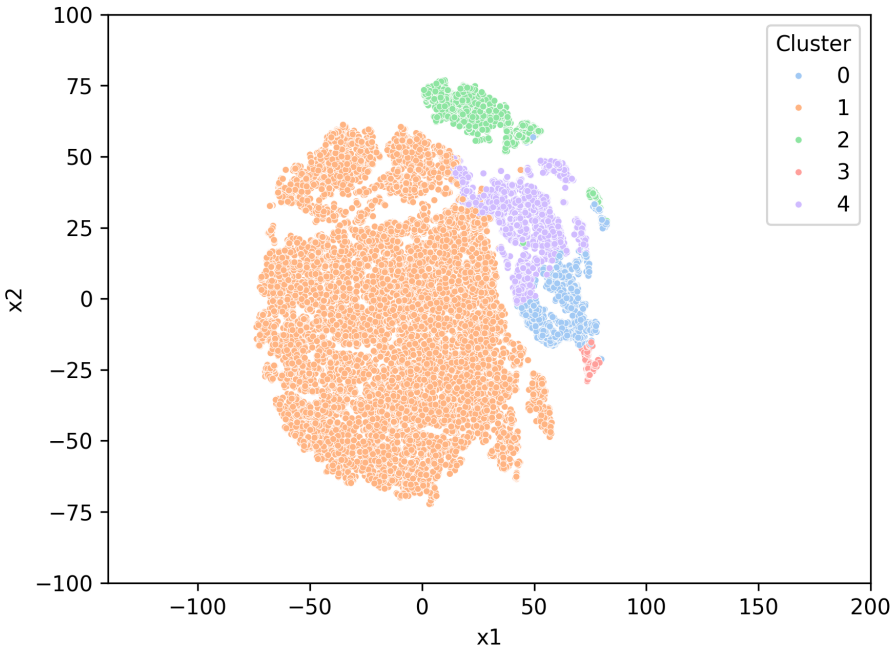

Figure 4: t-SNE embeddings of SARS-CoV-2 variants after learning

5.2.1 Application to detecting emerging variants

Results from clustering the dataset can be applied to the problem of detecting emerging variants of SARS-CoV-2. In order to prove that there is some benefit to using large-language models in this area, prior results from a previous model are replicated, and the proposed LLM-based approach is compared to it. The previously proposed framework by de Hoffer et al. [29] is an unsupervised framework that relies on the Levenshtein distances only as a similarity metric. In this approach, the authors start by computing the Levenshtein distances among all pairs of sequences. Following this, they generate a hierarchical clustering for each timestamp. This generates a dendrogram for each timestamp, and Ward's method is used to prune the tree and fix the number of clusters. The metric used by the authors is $rW = 100$ which is determined empirically, and this same metric is used in the proposed method for fair comparison. Finally, to ensure that the de Hoffer et al. [29] can be compared to the proposed approach, adjusted rand index (ARI) is used as a metric to compare the cluster alignment of their cluster predictions to the gold standard PANGO labels.

For the proposed framework, the large-language model embeddings and the contrastive learning framework are used from the previous section. In the first step, cosine distances are computed among all pairs of LLM embeddings predicted by the SNN with the optimal parameters as determined by the previous parameter study. The LLM used here is also GenSLM as it has shown the best performance among all the other LLMs. In the second step, a hierarchical clustering is generated for each timestamp and Ward's method is used with $rW = 100$ for fair comparison. In the last step, ARI is used to get a quantitative measure of how well the model predicts clusters compared to the gold standard PANGO labels.

Figure 5 and Figure 6 show boxplots of the two approaches over time. Based on the cluster alignment ARI metric, the proposed LLM-based method is better at reducing the distances within the same PANGO variants and increasing distances between different PANGO variants compared to the Levenshtein distance only approach by de Hoffer et al. [29]. The figures show that the proposed LLM-based approach is able to maximize the distances between sequence vectors belonging to different clusters while minimizing the distances between similar PANGO variants. The boxplots show that the LLM-based approach is able to consistently achieve this over all months in the test dataset compared to the previous approach by de Hoffer et al. [29]. Table 4 shows the results obtained using ARI on LD-only [29] and the proposed LLM-based approach. The gold-standard labels in this case considered for cluster alignment are the PANGO labels from the NCBI dataset. A significant improvement is observed for all months of the test dataset, with an average improvement of 0.1914 in cluster alignment using the LLM-based approach.

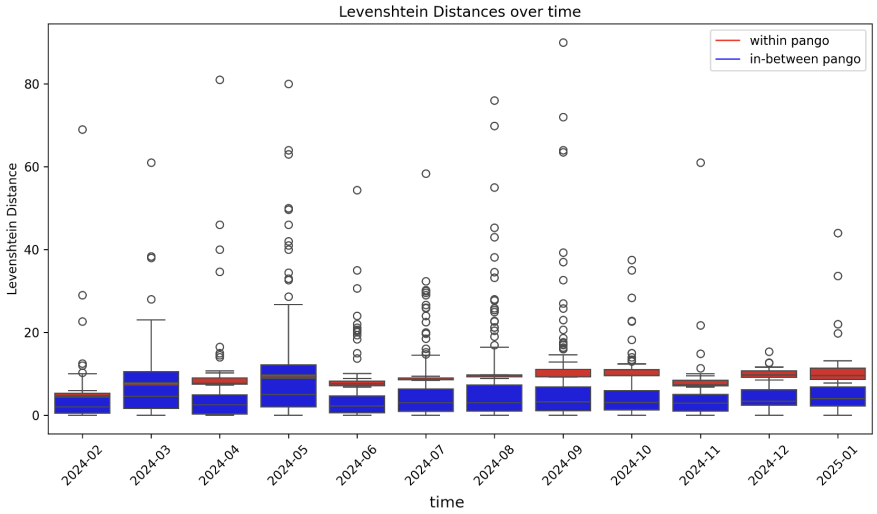

Figure 5: Levenshtein distance metrics over time based on de Hoffer et al. [29]

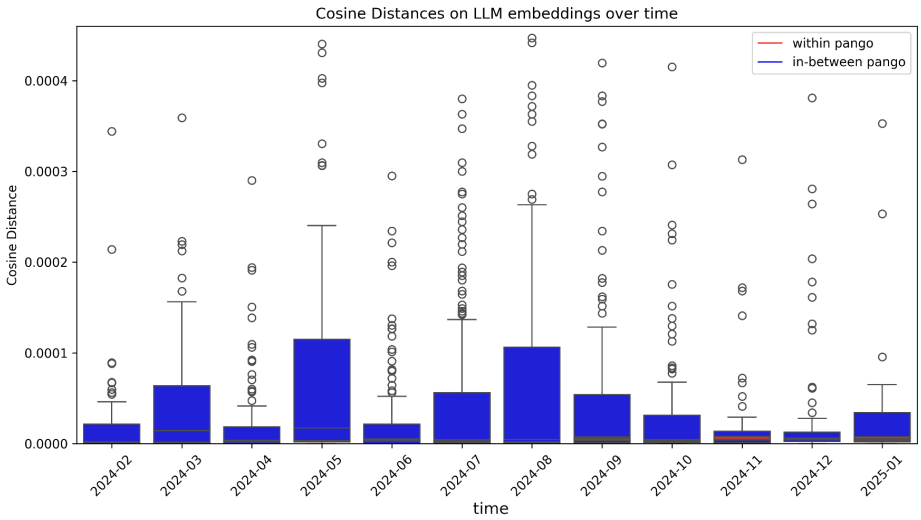

Figure 6: Proposed LLM-based approach metrics using contrastive learning

Table 4: Adjusted Rand Index (ARI) comparison of prior approach and proposed approach

Time	LD-only de Hoffer et al. [29]	LLM-based (proposed)
Feb-24	-0.0107	**0.1762**
Mar-24	0.0021	**0.1683**
Apr-24	-0.0011	**0.2406**
May-24	0.0000	**0.1908**
Jun-24	0.0424	**0.3032**
Jul-24	0.2042	**0.3350**
Aug-24	0.2524	**0.3946**
Sep-24	0.2652	**0.3905**
Oct-24	0.1297	**0.3667**
Nov-24	0.1614	**0.4250**
Dec-24	0.3220	**0.4106**
Jan-25	0.0958	**0.3584**

6 Conclusion

In this paper, we conducted a comparative study of previously proposed protein language models on SARS-CoV-2 surface glycoprotein (spike) sequences. We discuss how embeddings extracted from previously proposed models perform on classifying and clustering different variants. We then propose using a contrastive learning framework using a Siamese neural network in an unsupervised manner to compare different SARS-CoV-2 protein sequences. The benefit of using an unsupervised contrastive learning framework is that we do not rely on prior labeling methods or embedding vectors alone to compare different protein sequences. Using our framework, we show we can identify that the sequences come from different groups (variants) and that the proposed LLM-based approach shows an improvement of 0.1914 in terms of the adjusted rand index clustering on test data compared to a previously proposed approach.

7 Code

Code for this paper is publicly available here: `https://github.com/sblittlefield/covid-contrast`.

Acknowledgments

We would like to acknowledge Dr. Sergei Maslov from the Bioengineering department at the University of Illinois at Urbana-Champaign for his valuable feedback on our work. This work utilizes resources supported by the National

Science Foundation's Major Research Instrumentation program, grant #1725729, as well as the University of Illinois at Urbana-Champaign. We would also like to acknowledge all laboratories and authors that contributed sequence data used in this paper.

References

[1] B. Yu and T. Xiaonan, "Comparison of COVID-19 and influenza characteristics," *Journal of Zhejiang University. Science. B*, vol. 22, no. 2, p. 87, 2021.

[2] M. A. Johansson, T. M. Quandelacy, S. Kada, P. V. Prasad, M. Steele, J. T. Brooks, R. B. Slayton, M. Biggerstaff, and J. C. Butler, "SARS-CoV-2 transmission from people without COVID-19 symptoms," *JAMA network open*, vol. 4, no. 1, pp. e2 035 057–e2 035 057, 2021.

[3] C. B. Jackson, M. Farzan, B. Chen, and H. Choe, "Mechanisms of SARS-CoV-2 entry into cells," *Nature reviews Molecular cell biology*, vol. 23, no. 1, pp. 3–20, 2022.

[4] F. Krammer, "The human antibody response to influenza a virus infection and vaccination," *Nature Reviews Immunology*, vol. 19, no. 6, pp. 383–397, 2019.

[5] Z. N. Flamholz, S. J. Biller, and L. Kelly, "Large language models improve annotation of viral proteins," *Research square*, 2023.

[6] L.-Y. Wu, Y. Wijesekara, G. J. Piedade, N. Pappas, C. P. Brussaard, and B. E. Dutilh, "Benchmarking bioinformatic virus identification tools using real-world metagenomic data across biomes," *Genome Biology*, vol. 25, no. 1, p. 97, 2024.

[7] I. Aksamentov, C. Roemer, E. B. Hodcroft, and R. A. Neher, "Nextclade: clade assignment, mutation calling and quality control for viral genomes," *Journal of open source software*, vol. 6, no. 67, p. 3773, 2021.

[8] J. Huddleston, J. Hadfield, T. R. Sibley, J. Lee, K. Fay, M. Ilcisin, E. Harkins, T. Bedford, R. A. Neher, and E. B. Hodcroft, "Augur: a bioinformatics toolkit for phylogenetic analyses of human pathogens," *Journal of open source software*, vol. 6, no. 57, 2021.

[9] J. Hadfield, C. Megill, S. M. Bell, J. Huddleston, B. Potter, C. Callender, P. Sagulenko, T. Bedford, and R. A. Neher, "Nextstrain: real-time tracking of pathogen evolution," *Bioinformatics*, vol. 34, no. 23, pp. 4121–4123, 2018.

[10] S. Ali, B. Sahoo, A. Zelikovsky, P.-Y. Chen, and M. Patterson, "Benchmarking machine learning robustness in COVID-19 genome sequence classification," *Scientific Reports*, vol. 13, no. 1, p. 4154, 2023.

[11] S. Basu, "A study of the dynamics and genetics of COVID-19 through machine learning," Master's thesis, University of Illinois Urbana-Champaign, 2020.

[12] G. S. Randhawa, M. P. Soltysiak, H. El Roz, C. P. de Souza, K. A. Hill, and L. Kari, "Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study," *Plos one*, vol. 15, no. 4, p. e0232391, 2020.

[13] L. Brierley and A. Fowler, "Predicting the animal hosts of coronaviruses from compositional biases of spike protein and whole genome sequences through machine learning," *PLoS Pathogens*, vol. 17, no. 4, p. e1009149, 2021.

[14] S. Basu and R. H. Campbell, "Classifying COVID-19 variants based on genetic sequences using deep learning models," in *System Dependability and Analytics: Approaching System Dependability from Data, System and Analytics Perspectives*. Springer, 2022, pp. 347–360.

[15] E. Asgari and M. R. Mofrad, "Continuous distributed representation of biological sequences for deep proteomics and genomics," *PloS one*, vol. 10, no. 11, p. e0141287, 2015.

[16] N. Brandes, D. Ofer, Y. Peleg, N. Rappoport, and M. Linial, "ProteinBERT: a universal deep-learning model of protein sequence and function," *Bioinformatics*, vol. 38, no. 8, pp. 2102–2110, 2022.

[17] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in neural information processing systems*, vol. 26, 2013.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[20] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever *et al.*, "Improving language understanding by generative pre-training," *OpenAI*, 2018.

[21] M. Zvyagin, A. Brace, K. Hippe, Y. Deng, B. Zhang, C. O. Bohorquez, A. Clyde, B. Kale, D. Perez-Rivera, H. Ma *et al.*, "GenSLMs: Genome-scale language models reveal SARS-CoV-2 evolutionary dynamics," *The International Journal of High Performance Computing Applications*, vol. 37, no. 6, pp. 683–705, 2023.

[22] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma, and R. Fergus, "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences," *PNAS*, 2019. [Online]. Available: https://www.biorxiv.org/content/10.1101/622803v4

[23] R. M. Rao, J. Meier, T. Sercu, S. Ovchinnikov, and A. Rives, "Transformer protein language models are unsupervised structure learners," *bioRxiv*, 2020. [Online]. Available: https://www.biorxiv.org/content/10.1101/2020.12.15.422761v1

[24] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, A. dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido *et al.*, "Language models of protein sequences at the scale of evolution enable accurate structure prediction," *bioRxiv*, 2022.

[25] G. Koch, R. Zemel, R. Salakhutdinov *et al.*, "Siamese neural networks for one-shot image recognition," in *ICML deep learning workshop*, vol. 2, no. 1. Lille, 2015.

[26] S. Madan, V. Demina, M. Stapf, O. Ernst, and H. Fröhlich, "Accurate prediction of virus-host protein-protein interactions via a Siamese neural network using deep protein sequence embeddings," *Patterns*, vol. 3, no. 9, 2022.

[27] S. Tsukiyama, M. M. Hasan, S. Fujii, and H. Kurata, "LSTM-PHV: prediction of human-virus protein–protein interactions by LSTM with word2vec," *Briefings in bioinformatics*, vol. 22, no. 6, p. bbab228, 2021.

[28] S. Basu, R. H. Campbell, and K. Karahalios, "Detection of Novel COVID-19 Variants with Zero-Shot Learning," *HealthNLP'23, IEEE ICHI*, 2023.

[29] A. de Hoffer, S. Vatani, C. Cot, G. Cacciapaglia, M. L. Chiusano, A. Cimarelli, F. Conventi, A. Giannini, S. Hohenegger, and F. Sannino, "Variant-driven early warning via unsupervised machine learning analysis of spike protein mutations for covid-19," *Scientific Reports*, vol. 12, no. 1, p. 9275, 2022.

[30] A. Rambaut, E. C. Holmes, Á. O'Toole, V. Hill, J. T. McCrone, C. Ruis, L. du Plessis, and O. G. Pybus, "A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology," *Nature microbiology*, vol. 5, no. 11, pp. 1403–1407, 2020.

[31] Á. O'Toole, E. Scher, A. Underwood, B. Jackson, V. Hill, J. T. McCrone, R. Colquhoun, C. Ruis, K. Abu-Dahab, B. Taylor *et al.*, "Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool," *Virus evolution*, vol. 7, no. 2, p. veab064, 2021.

[32] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a "siamese" time delay neural network," *Advances in neural information processing systems*, vol. 6, 1993.

[33] K. Kieft and K. Anantharaman, "Virus genomics: what is being overlooked?" *Current opinion in virology*, vol. 53, p. 101200, 2022.

[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[35] V. Kindratenko, D. Mu, Y. Zhan, J. Maloney, S. H. Hashemi, B. Rabe, K. Xu, R. Campbell, J. Peng, and W. Gropp, "HAL: Computer system for scalable deep learning," in *Practice and experience in advanced research computing*, 2020, pp. 41–48.

[36] G. E. Hinton and S. Roweis, "Stochastic neighbor embedding," *Advances in neural information processing systems*, vol. 15, 2002.

[37] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of machine learning research*, vol. 9, no. 11, 2008.

[38] Z. Guo, J. Liu, Y. Wang, M. Chen, D. Wang, D. Xu, and J. Cheng, "Diffusion models in bioinformatics and computational biology," *Nature reviews bioengineering*, vol. 2, no. 2, pp. 136–154, 2024.

[39] W. Li, J. E. Cerise, Y. Yang, and H. Han, "Application of t-SNE to human genetic data," *Journal of bioinformatics and computational biology*, vol. 15, no. 04, p. 1750017, 2017.