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

December 16, 2024

ABSTRACT

The severe acute respiratory system coronavirus 2 (SARS-CoV-2) pandemic led to more than a 100 million infections and 1.2 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 and influenza 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 involvements in being recognized by the human immune system. Our contrastive learning framework is trained in an unsupervised manner, leveraging the Hamming distance from pairwise alignment of sequences when the contrastive loss is computed by the Siamese Neural Network. Finally, to test our framework, we perform experiments on two sets of sequences: one group belonging to a variant the model has not encountered in the training and validation phase (XEC), and the other group which the model has previously encountered (BA.2). We show that our model can acknowledge the sequences come from different groups (variants) as confirmed by a statistical Kolmogorov-Smirnov test. This shows that our proposed framework has properties suitable for identifying relationships among different SARS-CoV-2 sequences even in the absence of variant or lineage labels.

1 Introduction

The evolution of pandemics like COrona VIrus Disease (COVID-19) and influenza has 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 and influenza 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, enable 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 and influenza surface glycoprotein 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.

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` [29, 30]. 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 2020 to 2023. 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, our dataset had 134,510 records belonging to a total of 1495 different variants labeled by NCBI. These chronologically ordered records had sample collection dates from 2020 to 2023. For reproducibility, this was also split chronologically into 80% and 20% for our training and validation sets.

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` [29, 30]. 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). Similarly, influenza sequences were filtered based on country (United States), host (*Homo sapiens*), type of protein (surface glycoproteins), and subtype (H1N1, H3N2). The labels are obtained from the lineages.

Duplicate sequences were eliminated in both the SARS-CoV-2 and influenza datasets in the pre-processing step. After filtering and pre-processing the records for the comparison study, the SARS-CoV-2 and influenza datasets had 20,580 and 13,813 records respectively.

4.2 Contrastive Learning Framework

Contrastive learning models learn by comparing existing classes. Siamese Neural Networks (SNNs) [31, 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 [32]. 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 [33].

All code was implemented using Python on 1 GPU node and 4 CPUs with the HAL cluster [34].

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 and influenza 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.
- Dunn Index (DI): This metric measures the compactness of points within a cluster. The metric also measures how far apart points belonging to other clusters are. It is represented as $DI = min(a)/max(c)$, where a is the mean intra cluster distance and c is the mean inter-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.

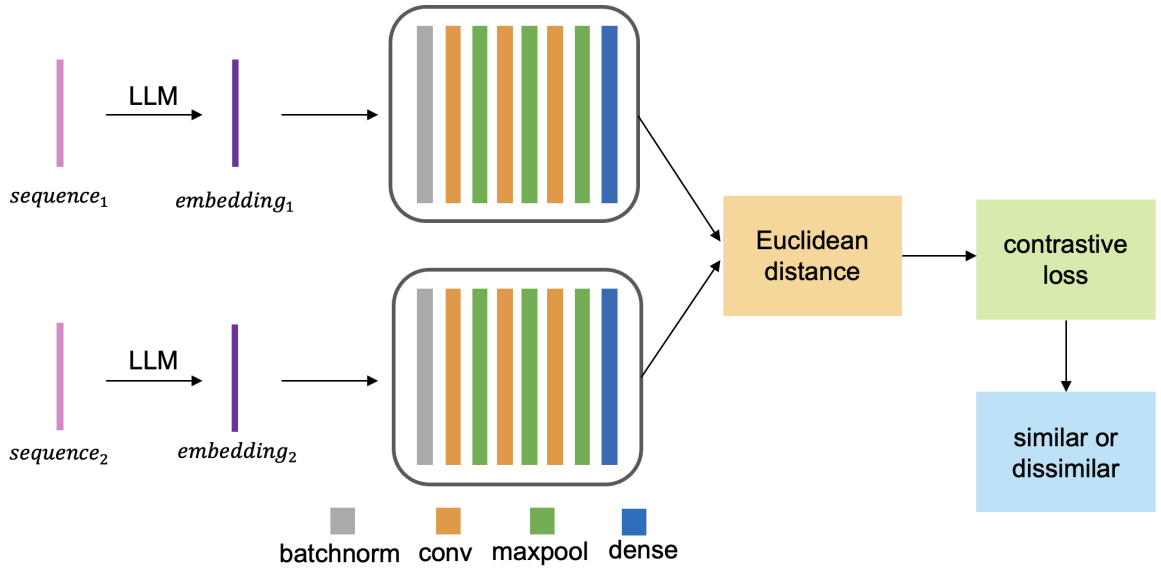

Figure 1: Architecture of contrastive learning framework (based on Basu et al. [28])

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 (February 2023) 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 [33], and a batch size of 512. There were five labels for SARS-CoV-2 and two labels for influenza 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 and 575 for influenza. 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 [35, 36] was only used to reduce the different embeddings to two dimensions for visualization purposes.

The results in Table 1 and Table 2 show that:

1. Protein large language models perform better than protein language models on classifying five different SARS-CoV-2 classes (lineages). The performance of all models is nearly at 1.0000 for all models in case of influenza. This is because embedding vectors belonging to the two influenza lineages have a broader separation between classes compared to the SARS-CoV-2 lineages. This makes classification an easier task on influenza compared to SARS-CoV-2.
2. When studying the performance of previously proposed protein large language models like GenSLM and ESM-2, GenSLM is slightly better at identifying SARS-CoV-2 lineages compared to ESM-2 while the performance is similar in case of influenza. Recent literature [37] also supports the claim that diffusion generative models which are present in GenSLM have the ability to learn distributions from high-dimensional bioinformatics data. For this reason, we choose to use GenSLM in the unsupervised contrastive learning framework.
3. Cluster metrics are included to show that embedding vectors do not cluster well based on metrics like SC and DI which are computed based on the SARS-CoV-2 and influenza lineages. This means that it is difficult to classify variants according to the embedding vectors as-is from protein language models, which motivates us to implement our unsupervised contrastive learning framework for this purpose.

Table 1: Comparison of protein LM embeddings on classification

Dataset	Model	val_acc	val_f1	val_loss
SARS-CoV-2	ProtVec [15]	0.9446	0.9729	0.1309
	ProteinBERT [16]	0.9643	0.9803	0.2535
	GenSLM [21]	**0.9905**	**0.9953**	0.0515
	ESM-2 [24]	0.9645	0.9829	0.2451
Influenza	ProtVec [15]	**1.0000**	**1.0000**	7.5e-05
	ProteinBERT [16]	0.9996	0.9997	0.0023
	GenSLM [21]	0.9996	0.9997	0.0011
	ESM-2 [24]	**1.0000**	**1.0000**	2.1e-04

Table 2: Comparison of protein LM embeddings on clustering

Dataset	Model	SC	DI
SARS-CoV-2	ProtVec [15]	0.2856	0.027
	ProteinBERT [16]	0.1517	0.0086
	GenSLM [21]	0.2196	0.0137
	ESM-2 [24]	0.2561	0.0357
Influenza	ProtVec [15]	0.2290	0.1162
	ProteinBERT [16]	0.3112	0.0257
	GenSLM [21]	0.2508	0.1014
	ESM-2 [24]	0.2616	0.1517

5.2 Unsupervised contrastive learning framework

The chronologically ordered SARS-CoV-2 dataset is partitioned into train (80%) and test (20%) respectively. In this subsection, we discuss results using GenSLM [21] because it is the best performing model on SARS-CoV-2 data based on the comparative study.

Sampling Pairs In order to prepare the dataset, we first need to sample pairs of embeddings and assign labels indicating whether they are similar or dissimilar. The pairs of embeddings are generated based on the GenSLM embeddings. Next, the sampled pairs are labeled according to the Hamming distance between the original spike protein sequences. We obtained 53,804 labels using this process. This is constrained based on a threshold parameter hd. When

the Hamming distances between the original sequences are computed, the pairs of embeddings are labeled similar if the distance is below the hd threshold. Otherwise, the pairs are labeled different. If we use randomly generated labels in the training and validation datasets, the proposed computation gives a validation accuracy of 0.7373 and F-1 score of 0.0. Instead, to improve the validation accuracy and F-1 score, we used the Hamming distance in computing the labels.

Table 3 shows the fine-tuning of parameters on the validation dataset. We use different values of Hamming distances thresholds against different values of the contrastive loss function margin to compare the changes in the validation accuracy and F-1 scores. The first row of each set of parameters represents the accuracy, while the second row represents the F-1 score. We observe that a lower value of the Hamming distance threshold at $hd = 500$ along with a margin value of $m = 0.3$ produces the optimal validation accuracy and validation F-1 scores of 0.7436 and 0.4101 respectively.

We used 2-dimensional t-SNE plots [35, 36] to observe how the contrastive learning framework contributes to the separation of the classes. The t-SNE plots reveal local relationships in the genomic data, which is why we choose to apply this method, and this is also shown in prior research [38]. The 2-dimensional t-SNE plots have the axes $x1$ and $x2$, which are used to project the 512-dimensional embedding vector points onto a 2-dimensional space. Figure 2 and Figure 3 show the 2-dimensional t-SNE plots of GenSLM embeddings before and after the SNN respectively. The cluster metrics are reported based on k-means clustering following the SNN training. The same cluster labels from k-means are used for the embedding vectors prior to training to highlight the intra cluster relationships. The k-means labels show that the SC score improved from 0.0817 to 0.2903 using the unsupervised contrastive learning framework which shows more cohesive clusters were formed while the DI improved from 0.0055 to 0.0097.

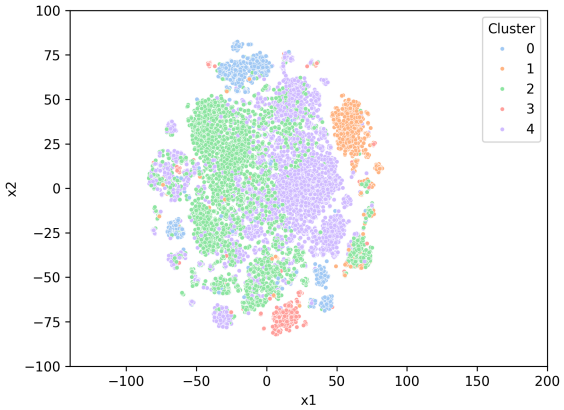

Figure 2: t-SNE embeddings of SARS-CoV-2 variants before learning

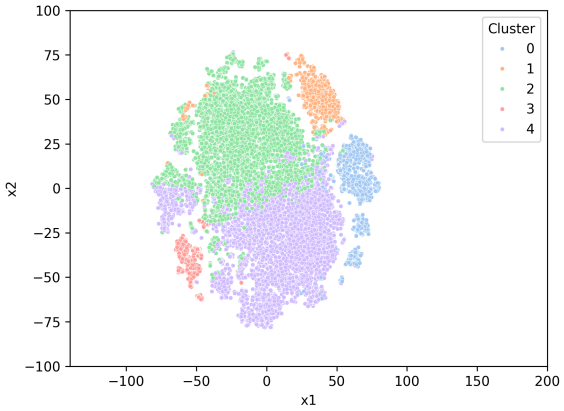

Figure 3: t-SNE embeddings of SARS-CoV-2 variants after learning

Table 3: Parameter tuning on validation data

Parameter	m=0.5	m=0.4	m=0.3	m=0.2
hd=500	0.7386	**0.7560**	**0.7436**	0.5610
	0.0033	**0.2279**	**0.4101**	0.0008
hd=600	0.3945	**0.7415**	**0.7235**	0.5785
	0.0305	**0.3037**	**0.3492**	0.0165
hd=700	0.7381	**0.7482**	0.5547	**0.7280**
	0.0000	**0.0935**	0.0002	**0.4589**

Finally, to test the model, we analyzed a group of 15 surface glycoprotein sequences sampled from January 2024 to November 2024 belonging to a novel variant BA.2 (we call this $Group_A$) which was not present in the training or validation dataset and another group of 15 surface glycoprotein sequences sampled from September 2024 to October 2024 belonging to variant XEC (we call this $Group_B$). All test sequences with labels were extracted from NCBI Virus randomly after filtering recently circulating SARS-CoV-2 variants in the United States from 2024. BA.2 sequences were previously present during the training and validation phases of the model, but because of the unsupervised framework, the model was trained without these labels. For testing purposes, we tokenized and obtained the GenSLM embedding of these test sequences in the same way as the training and validation datasets.

The similarity score obtained as a prediction from the trained SNN with the optimal parameters is a representation of training the model on self-labeled Hamming distances between sequences. The mean similarity score obtained from the trained SNN of one group was compared to the mean similarity score obtained from the other group using three combinations:

1. A vs. A ($Group_AA$): Comparing all group A sequences with all group A sequences
2. A vs. B ($Group_AB$): Comparing all group A sequences with all group B sequences
3. B vs. B ($Group_BB$): Comparing all group B sequences with all group B sequences

Table 4: SNN similarity scores on test data

AA_before	AB_before	BB_before	AA_after	AB_after	BB_after
0.9999	0.9997	1.0000	0.4621	0.4130	0.5259
0.9999	0.9997	1.0000	0.4597	0.3954	0.5259
0.9998	0.9999	1.0000	0.4248	0.4494	0.5259
0.9998	0.9999	1.0000	0.4641	0.4322	0.5259
0.9999	0.9997	1.0000	0.4647	0.3945	0.5259
0.9999	0.9998	1.0000	0.4698	0.4182	0.5259
0.9998	0.9999	1.0000	0.4365	0.4633	0.5259
0.9999	0.9997	1.0000	0.4653	0.3998	0.5144
0.9999	0.9997	1.0000	0.4580	0.3853	0.5229
0.9998	0.9999	1.0000	0.4396	0.4704	0.5259
0.9998	0.9999	1.0000	0.4465	0.4543	0.5259
0.9998	0.9998	1.0000	0.4516	0.4401	0.5259
0.9999	0.9997	1.0000	0.4578	0.4284	0.5259
0.9999	0.9997	1.0000	0.4680	0.4183	0.5259
0.9998	0.9997	1.0000	0.4609	0.3882	0.5259

Table 4 shows results on the test dataset. A statistical Kolmogorov-Smirnov test was used to determine if $Group_A$ and $Group_B$ similarity scores were significantly different with $p < 0.05$ considering the samples are drawn from a continuous distribution. To understand the effect of the SNN in evaluating differences in these groups, we performed one set of statistical tests on the embeddings generated directly by GenSLM prior to training our contrastive learning model. For the similarity score in this case, we take the cosine similarity between embedding vectors and report the average for each sequence. These results are shown in the AA_before, AB_before, and the BB_before columns. The other set of statistical tests is performed after passing the GenSLM embeddings through the trained SNN. We report the average similarity score obtained from the SNN for each sequence. These results are shown in the AA_after, AB_after, and the BB_after columns. The values after using the trained SNN show that there is more distinction in the comparisons compared to using only the embeddings for comparisons.

Table 5: Summary of statistical test results

sample	statistic	p-value
AA_before and BB_before	1.0000	1.29E-08
AA_before and AB_before	0.5333	0.0262
AB_before and BB_before	1.0000	1.29E-08
AA_after and BB_after	1.0000	1.29E-08
AA_after and AB_after	0.6000	7.65E-03
AB_after and BB_after	1.0000	1.29E-08

In order to determine whether the similarity scores are different, we compute p-values between each of the three pairs: AA_after and BB_after, AA_after and AB_after, and AB_after and BB_after. Our p-values of 1.29E-08, 7.65E-03, and 1.29E-08 respectively indicate that the samples are significantly different. A summary of these values is shown in Table 5. Furthermore, a lower p-value between AB_after and BB_after (1.29E-08) compared to the p-value between AA_after and AB_after (7.65E-03) statistically suggests that the samples in $Group_B$ are more distinct or novel from what the SNN has previously been trained on compared to those in $Group_A$. This is in consensus with the NCBI Virus labels assigned to these groups: BA.2 and XEC for $Group_A$ and $Group_B$ respectively. We observe that the summary statistics for the samples prior to training using the SNN are similar except in the case of AA_before and AB_before. An increase in the statistic and decrease in the p-value after training using the SNN for $Group_AA$ and $Group_AB$ also suggests that our contrastive learning model is able to improve performance on variants newly encountered during test time.

6 Conclusion

In this paper, we conducted a comparative study of previously proposed protein language models on SARS-CoV-2 and influenza surface glycoprotein 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. Following parameter tuning on validation data, we obtain validation accuracy and validation F-1 scores of 0.7436 and 0.4101 respectively. Furthermore, to test our framework, we perform experiments on two set of sequences: one group belonging to a variant the model has not encountered in the train and validation phase (XEC), and the other group which the model has previously encountered (BA.2). Using our framework, we show we can identify that the sequences come from different groups (variants) as confirmed by a statistical Kolmogorov-Smirnov test.

7 Code

Code for this paper is publicly available here: `https://github.com/sbasulittlefield/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 insights in discussing the methods used in this paper. 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. 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.

[30] 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.

[31] 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.

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

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

[34] 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.

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

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

[37] 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.

[38] 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.