Applications of Sliding Sampling to Biological Sequences

Octavio Gonzalez-Lugo
ogonzalezl@outlook.com

Abstract

Biological sequences contain information about the function and regulation of different components within a biological system. Accurate modeling will yield a deeper understanding of the system, as well as better experimental designs. Nonetheless, computational constraints of the different available methods continue to prevent its use on large scale. Sliding sub-sampling of biological sequences can generate either vector-based or graph-based sequence encodings. Each one of them can be used to train generative models for an unsupervised representation learning task. Offering a suitable tool to analyze fast-evolving biological systems such as SARS-Cov2. Analysis of variational autoencoders bottleneck representation shows a distinguishable temporal component. Changes in 4-mer composition in the region that codes for the structural SARS-Cov2 proteins. Non-symmetrical changes in 4-mer composition drive the viral temporal adaptation process. Furthermore, mean nucleotide composition and encodings of SARS-Cov2 appear to be constrained by day length. Development and refinement of sequence analysis methods will lead to a better understanding of viral adaptation and evolution.

Sequences

Biological sequences

Since the discovery of how information passes through generations, the importance of sequences in biology became obvious. From passing information through generations to the structural characteristics of biological components. The study of sequences leads to a better understanding of diseases and physiology. This information is coded in at least three main kinds of sequences DNA, RNA, and protein sequences.

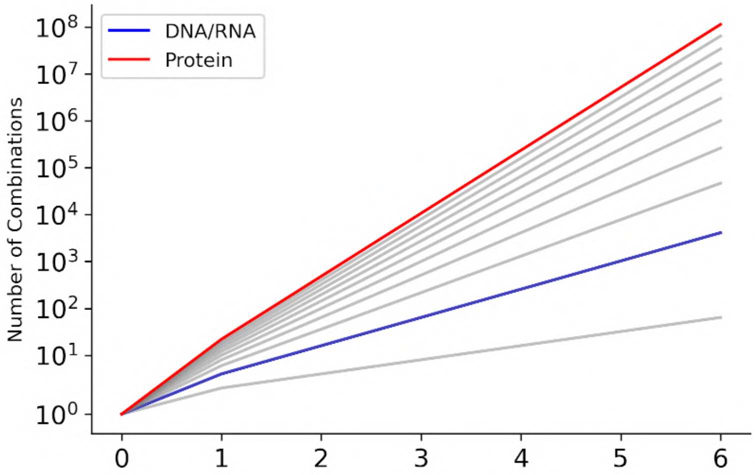

Figure 1. Alphabet Size and number of combinations.

DNA is a double-stranded polypeptide made of four different nucleosides, a combination of a monosaccharide and a nitrogenous base. While the RNA is a single strand polypeptide of four different nucleosides. DNA and RNA share more similarities between them. both share three out of the four nitrogenous bases, but a different monosaccharide. DNA and RNA contain encoded information to synthesize new proteins. A protein is another kind of polypeptide made by a combination of 21 different amino acids. The order and kind of amino acids in a protein are determined by DNA/RNA. DNA sequences are transcribed into RNA sequences and the resulting RNA sequence is translated into a protein. That roughly describes the central dogma of molecular biology [10].

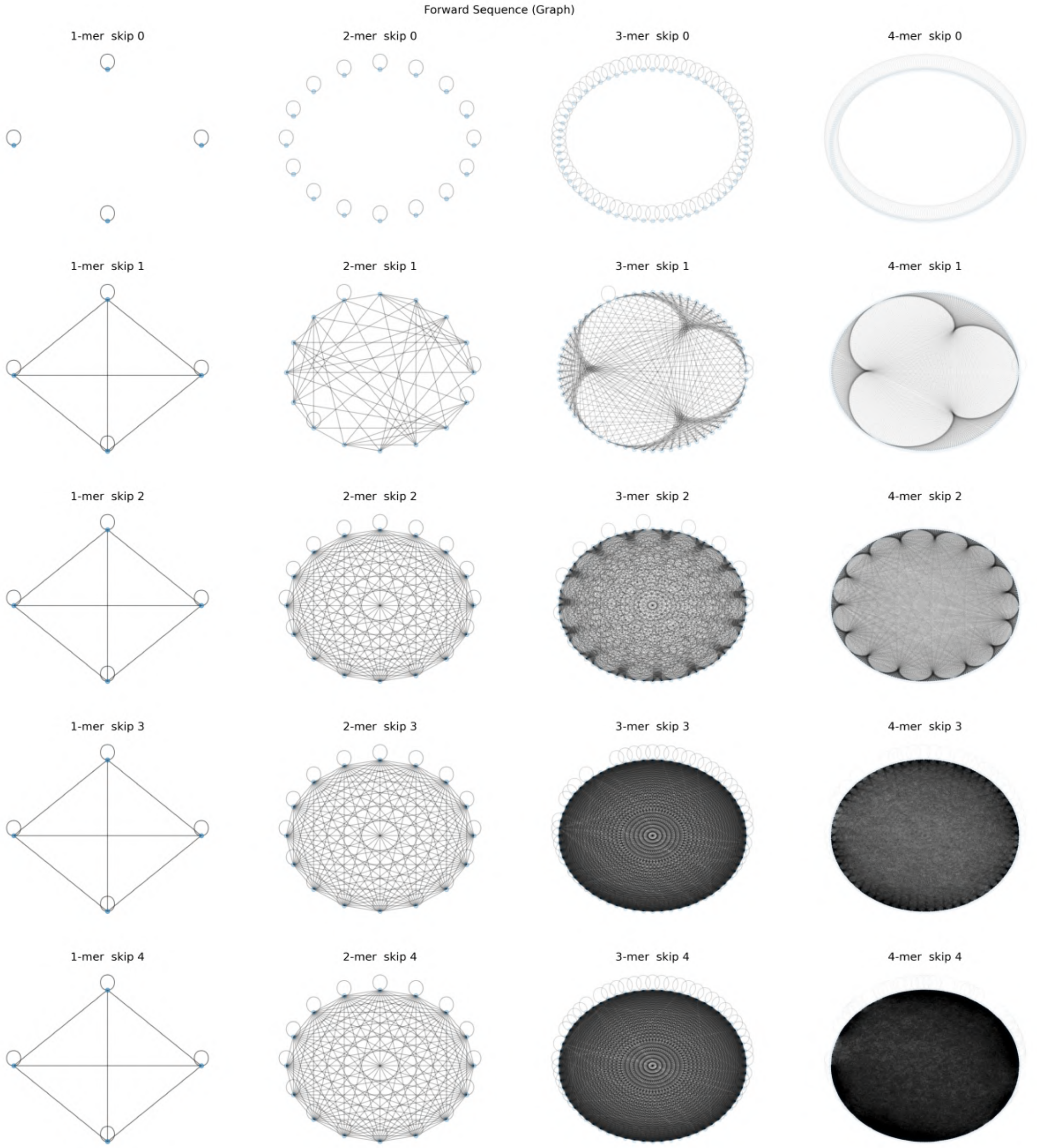

Figure 2. SARS-Cov2 genome graphs. Unique connectivity patterns show unique encoding capabilities for each graph.

DNA, RNA, and protein peptides are referred to as sequences and represented as a continuous string where each letter represents a single element in the peptide. Each letter in the string represents a nucleoside base or an amino acid following the same order as the peptide. This representation allows to compare sequences and measure their similarity. Similar sequences might share common properties and functions between them. [15]

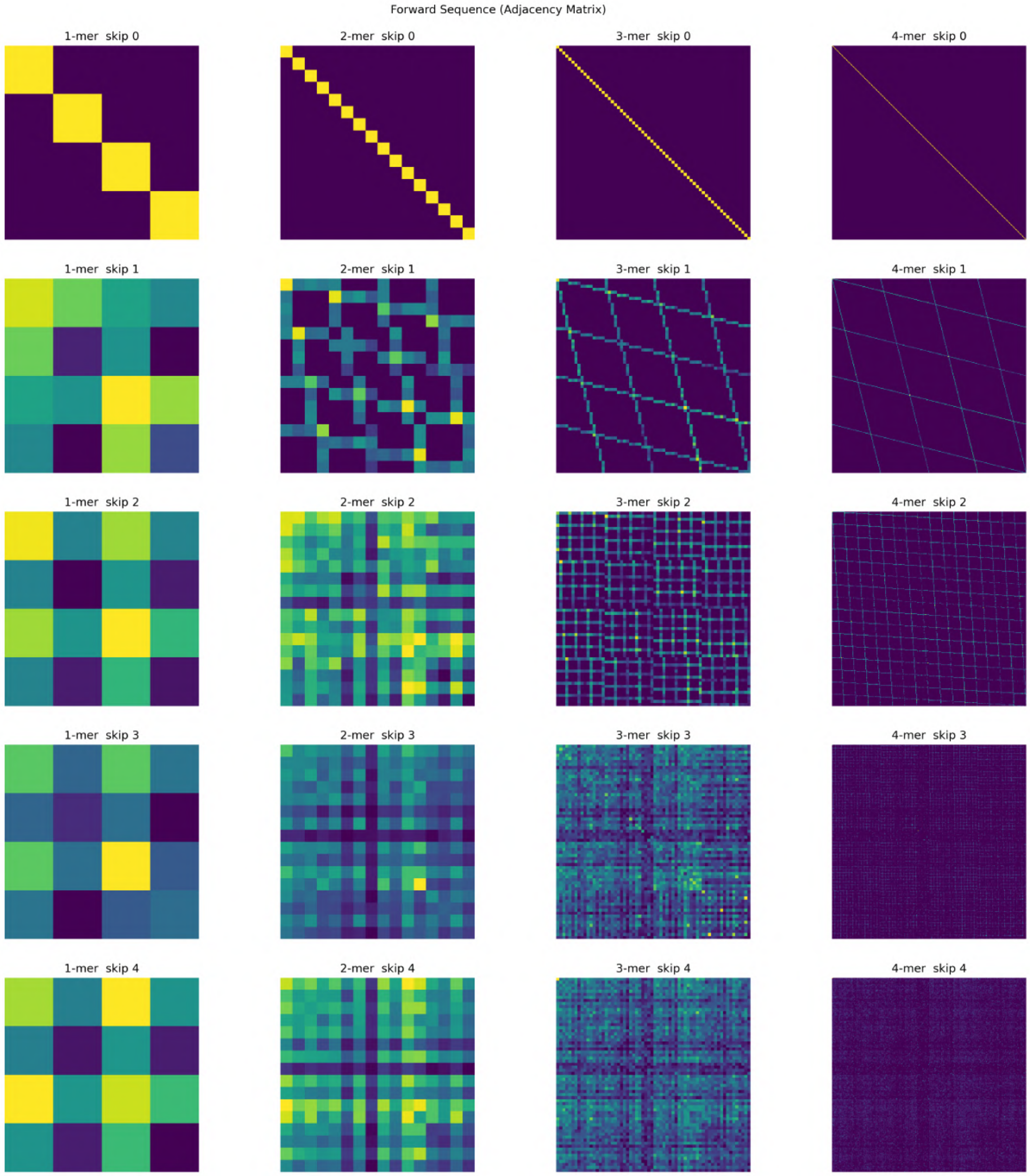

Figure 3. SARS-Cov2 genome adjacency matrices.Bidimensional representation of sequences shows small local patterns. Spatial closeness might provide enough information to find relations between long-range dependencies.

Comparison between two sequences is often made by aligning the sequences and minimizing the edit distance between them. Multiple alignments of sequences often result in shared sequence patterns. Patterns inside DNA/RNA sequences are referred to as consensus sequences and often represent regulatory elements. A regulatory element is a sequence fragment that regulates the expression of genes. [16] While patterns inside protein sequences or motifs represent among other things a binding surface of a protein. [18]

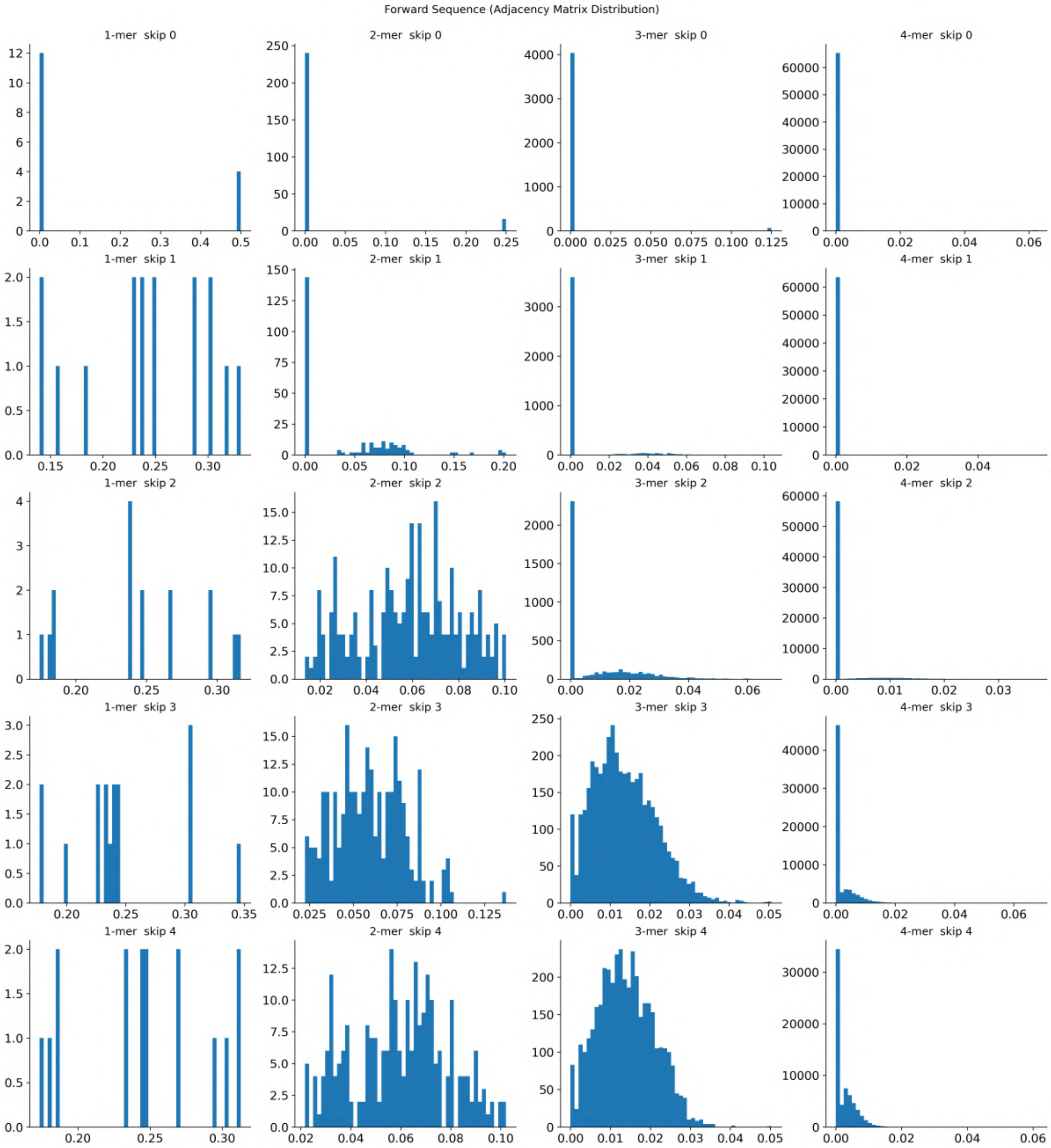

Figure 4. **SARS-Cov2 genome adjacency matrices value distribiution**.

Finding such patterns through sequence alignment can be computationally expensive as the number and length of the sequences to align increases. Although heuristics have been developed, multiple alignments of sequences continue to be a challenging task. [25] Thus the development of alignment-free methods provides an alternative to analyzing high numbers of sequences. There are roughly two categories of alignment-free methods, frequency-based and information theory based. [30] Both of those methods rely on the analysis of short fragments of sequences o k-mers, where k stands for the fragment size to be analyzed.

With the increased accuracy of machine learning models, several attempts to analyze and classify sequences have been made. Non deep learning methods often rely on the development of useful features that capture differences and similarities between the sequences. [13] While deep learning approaches are a mix of feature-based modeling and the use of successful neural network architectures adopted from NLP tasks. [11]

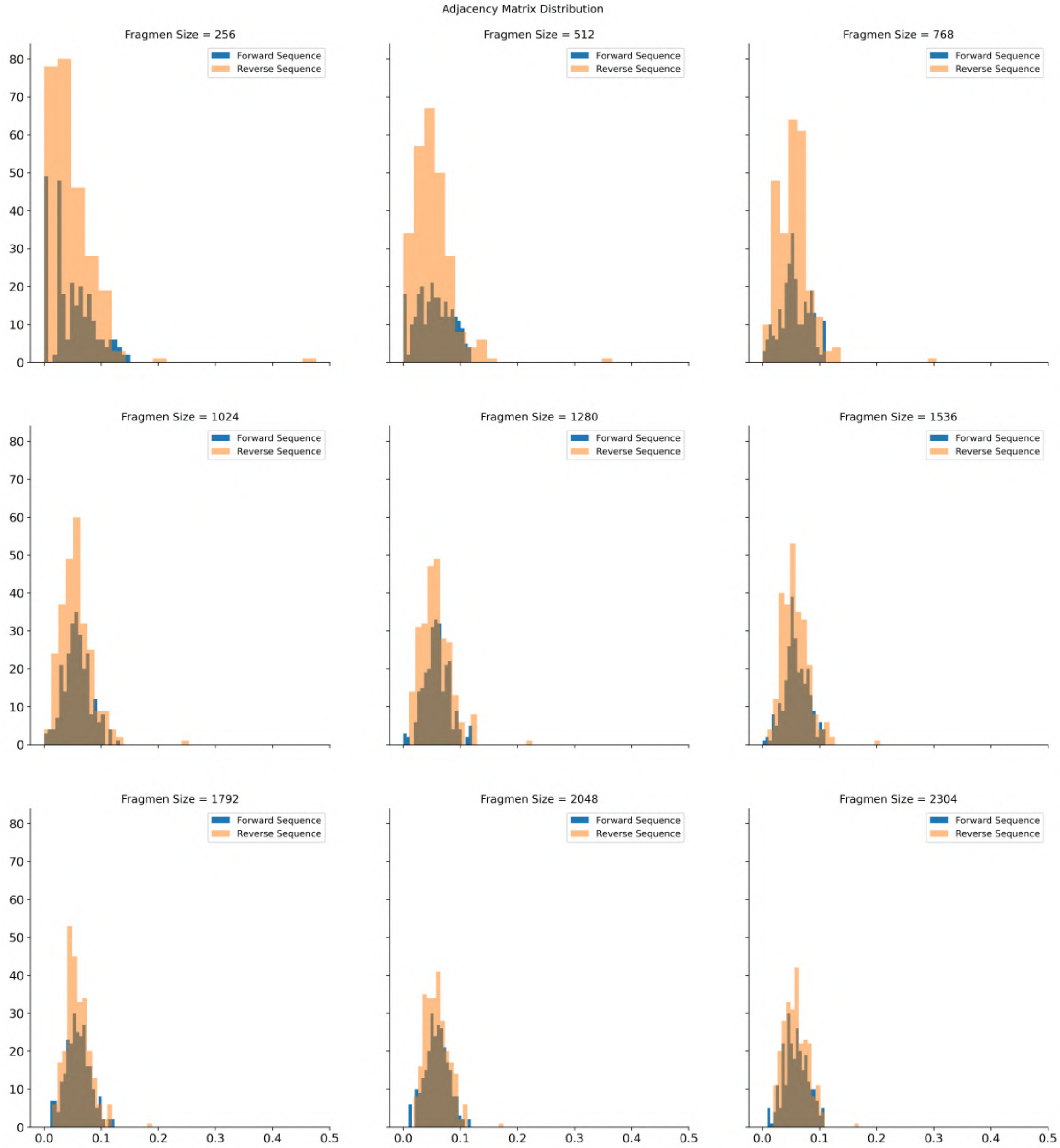

Figure 5. Adjacency matrices value distribuition. Forward and reversed sequence value distribution for k=s=2.

Sequences properties

A sequence can be described by three key properties. The different unique elements or alphabet, the connectivity between the different elements, and the direction in which the sequence has meaning.

For example, if the last paragraph is read backward, then the intended meaning of that paragraph is lost. The different letters, punctuation, and spaces will constitute the alphabet, and the connections between the different elements in the alphabet constitute the connectivity between the different elements in the sequence. This connectivity relation also contains encoded sequence size. If the sequence has the same meaning regardless of the direction in wight the sequence is read, such sequence is said to be palindromic.

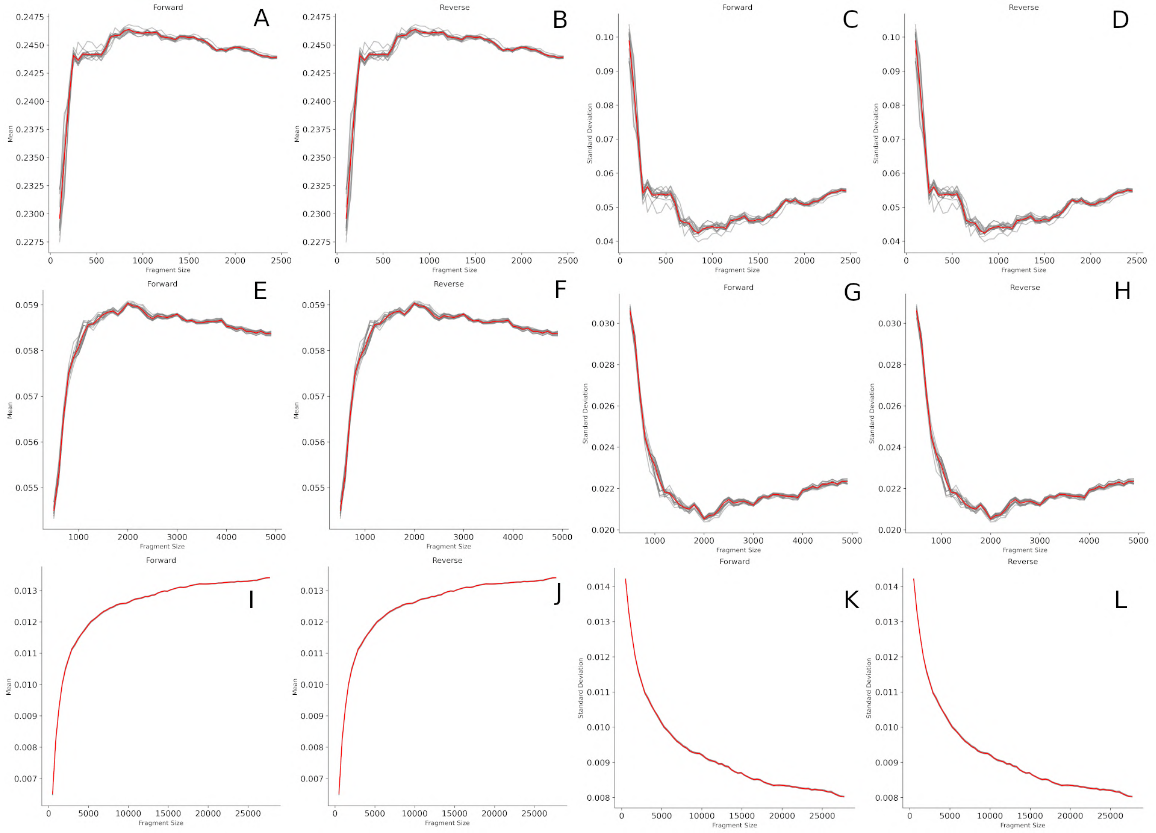

Figure 6. Mean and Standard deviation convergence. Mean and standard deviation are used as measures of information content. Sequence size at plateau is used to determine max fragment size. A,B,C,D convergence for k=s=1, E,F,G,H convergence for k=s=2, and I,J,K,L convergence for k=s=3

Thus a sequence can be defined as an ordered sequence of alphabet elements. And if the reverse element of a sequence is equal to the original sequence, such sequence is said to be palindromic.

Connectivity between the different elements in a sequence can be represented as a graph. Where each node stands for an element in the alphabet and each vertex is a connection between those two elements in the alphabet. An undirected graph will only capture the connectivity relations between the different elements in the alphabet. While a directed graph will retain both connectivity and direction.

The number of nodes in the graph will depend on the number of elements in the alphabet. For sequences with small alphabets, the number of nodes will be small. However, by constructing the graph with only the alphabet elements information could be lost due to the high compression of the sequence encoding. Sub-sampling the sequence by creating an auxiliary alphabet with the different combinations of original alphabet elements could lead to a sequence representation that accurately balances both compression and information.

In the case of DNA or RNA sequences the alphabet size is equal to four, and sub-sampling the sequence into two-character combinations leads to 16 different combinations. While for protein sequences the alphabet size is equal to 20 leading to 400 combinations. As the k-mer size increases the number of combinations increases in several orders in magnitude.(Figure 1) Due to the large size of the auxiliary alphabet, the following analysis will continue using only DNA/RNA sequences.

The auxiliary alphabet is then used to subsample the sequence, subsampling can be performed into two schemes, a non-overlapping scheme, and an overlapping scheme. In the non-overlapping scheme, the sequence is divided into non-overlapping k-size fragments, leading to L/k fragments where L is the sequence length. While in the overlapping scheme the sequence is divided into k size fragments but each fragment overlaps the next fragment by k-1 characters, leading to L-k fragments. The non-sliding scheme results in a lower number of fragments, but in the case of DNA/RNA sequence, such sampling could result in fragments with no meaning between them. While the sliding scheme adds the ability to check for every reading frame inside the DNA/RNA sequence.

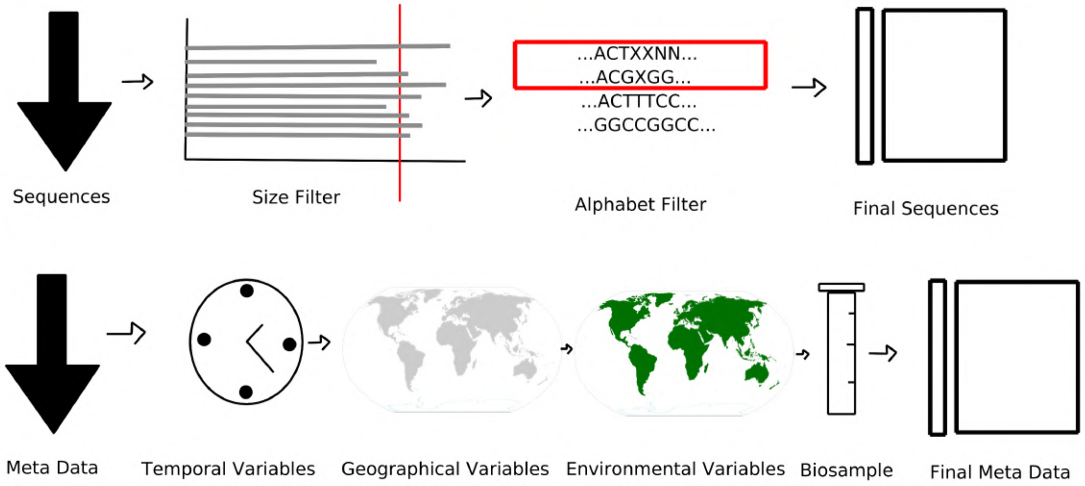

Figure 7. Sequence and metadata processing workflow.

A special case of sliding sampling of a sequence is the Debruijn graph. A graph is constructed from a sequence where each node shares k-1 characters. This property makes it useful for sequencing applications. As constructing the Debruijin graph and its Euler path is a method to ensemble a genome through its reads. [17] The overlap between fragments can be extended by adding a skip (s) during the construction of the de Bruijn graph. This extends the graph encoding scheme to a range from single element encoding to a gap between fragments.(Figure 2)

When there is no overlap the graph only encodes the information of a single element in the extended alphabet. As the fragment size increases the graph passes the Debruijn graph case and the resulting extended cases. Each graph shows unique connectivity patterns inside of them, a characteristic that can be visualized through its adjacency matrix.(Figure 3)

A honeycomb-like pattern is obtained in the Debruijin graph case, then as s increases the size becomes shorter merging into a unique pattern. However, increasing fragment size or k-mer size also increases the sparsity in the encoding.

The distribution of the normalized adjacency matrix confirms the sparsity of the values inside the matrix. With k=4 the distribution results skewed towards zero while increasing s removes such sparsity. (Figure 4) When s ¿ k there's little effect on the value distribution inside the adjacency matrix in either k=2 and k=3. Original sequence and reverse order sequence shows a different distribution in both k=2 and k=3 at s=k with different sequence size. (Figure 5) These changes in the distribution could lead to encoding a single sequence to a unique element.

Mean and standard distribution of original and reverse order sequence shows that given a sufficiently large sequence fragment both values reach a plateau. For k=1 and k=2 such behavior is observed, particularly from about 8 times the adjacency matrix size. (Figure 6) If that is the lower bound for sequence size complete SARS-Cov2 genome is not large enough to provide enough information for k=3 and beyond.

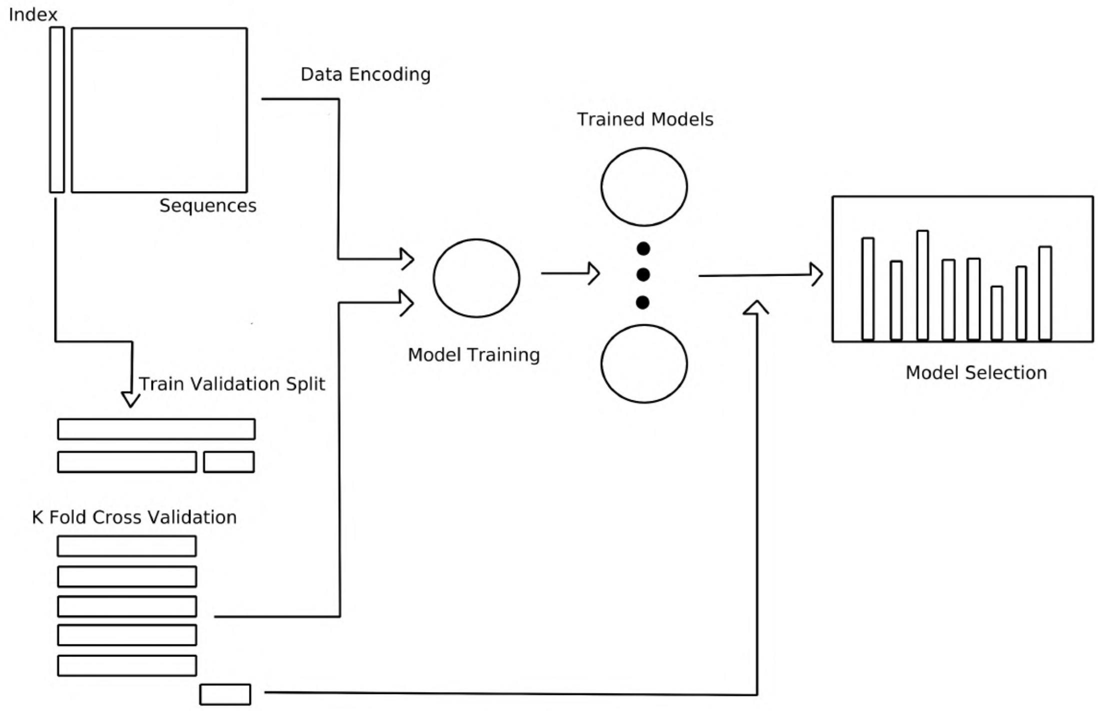

Figure 8. Data usage strategy. Valid sequences are split into two datasets, train and validation datasets. The train dataset is then used for five-fold cross-validation and the validation dataset is used to test for model performance.

Based on the information that can be encoded by the adjacency matrix representations the max k-mer size could be k=s=2 leading to encoding a four elements fragment. However, the sequence will need to be subdivided as the sequence encoding could end up overloaded with information. While k=s=3 might result in an encoding that does not fully capture the complexity of the sequence. But it might be able to capture better long-range dependencies or global structures within SARS-Cov2 sequences.

Applications to machine learning and the Covid-19 pandemic

Dataset Generation

Continuous genomic surveillance of the ongoing Covid-19 pandemic resulted in the generation of high amounts of genomic sequences. This can be exploited to understand the internal structure and patterns inside the SARS-Cov2 genome. Leading to different insights into the virus biology and evolution. Nonetheless, most of the machine learning applications have been developed for biomedical data such as chest x-rays and to predict cases throughout the pandemic. [29] [27] While sequence-based modeling is currently aimed at classification tasks and not understanding the inner characteristics of the genome or the sequence. [1]

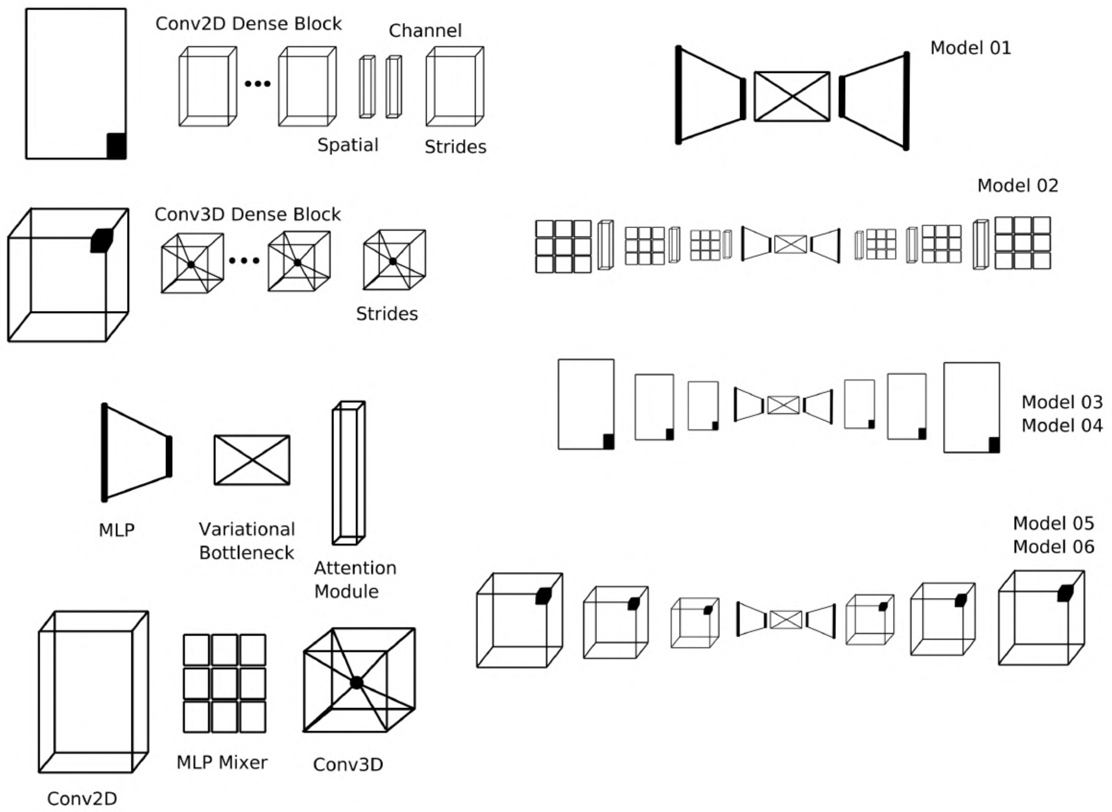

Figure 9. VAEs architectures. Different data structures are used to model the SARS-Cov2 sequences leading to different VAE architectures. Single MLP is used for a vector based datset. While MLP mixer is used for non pair size 2D data. Main convolutional blocks are generated by concatenating the different convolutional layers, either 2D or 3D. Downsampling is performed by transposed convolution.

Full genome sequences were downloaded from the NCBI SARS-Cov2 resources database with its corresponding metadata. Only two main filters were applied to select the sequences for further analysis. [9] In the first one, only complete genomes with no gaps or unresolved residues were selected for the analysis. And only sequences isolated from 110 longitude to the US east coast, were used, as most of the sequences were isolated in that region. (Figure 7)

Available metadata was used to retrieve geographical location by latitude, longitude, altitude, and other weather variables. The bio-sample identifier was used to retrieve PCR ct results, sex, age, and ethnicity.

Full sequence data is transformed into four different datasets. The first one contains the stacked k-mer frequency for k up to five. Raw frequency is used to prevent any kind of data leakage.

The second and third datasets contain the stacked difference of adjacency matrices for k up to 3. For the second dataset, the difference between the adjacency matrix with s=1 and s=k is the final sequence encoding.

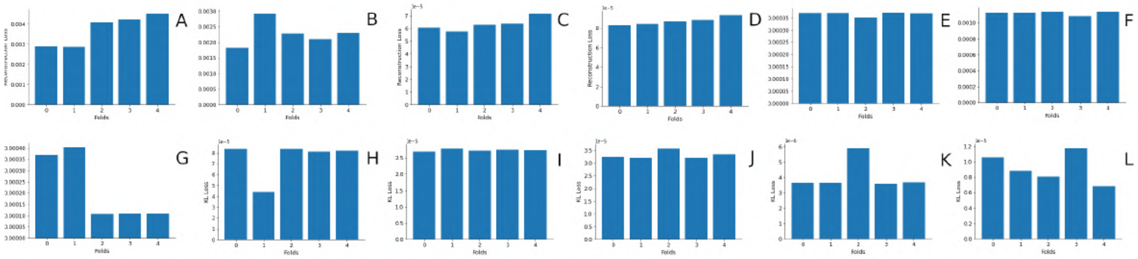

Figure 10. Model performance.. Reconstruction loss in the top row, while KL loss in the bottom row. Models show small variations within folds.

While for dataset three, the structure is like dataset 2, however both forward and reversed sequences are joined. The lower diagonal is equal to dataset two and the upper diagonal contains the corresponding representation of the reversed sequence. And the main diagonal contains the k-mer frequency.

For dataset four, the sequence is split into 16 fragments, cropping points will be calculated based on sequence length. Individual fragment size will change, however, a fixed size might push some differences towards the last fragment. For each fragment, the k=s=2 adjacency matrix is calculated with the forward and reverse fragments. This will result in data like a small video clip with two channels.

Dataset generation in a desktop machine with an AMD® Ryzen 5 2600 processor takes about a day and a half for datasets two and three. Frequency-based takes around 8hr, and for the final dataset, it takes around 5hrs.

Datasets can be found in CSV format for frequency-based modeling, where each row contains the information of a single sequence. While for adjacency matrix-based modeling, a single npy file contains the encoding of a single sequence. File name is equal to the accession id. Each file and dataset can be downloaded from the code repository.

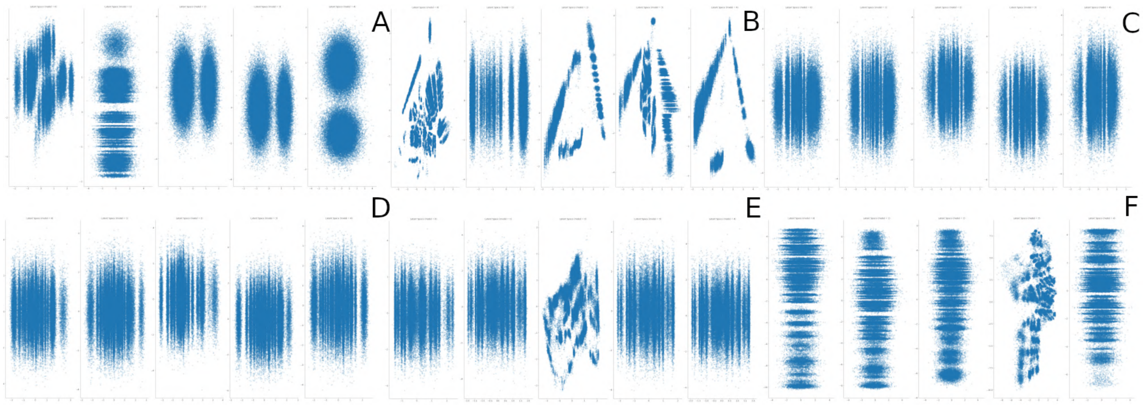

Figure 11. Bottleneck representations. Bottleneck representations show a particular inner structure. In most cases ordered towards a single axis.

Model Training

Graph or frequency-based encoding of sequences represents some data structures used to input sequences for analysis. Yet the high dimensionality of the data makes it difficult to understand those characteristics. Dimensionality reduction techniques offer a solution to the dimensionality problem by mapping high dimensional data into a lower-dimensional space. [21] A subset of dimensionality reduction is representation learning. Where the main goal is to find a suitable representation for a downstream task. One of the learning architectures used for representation learning is the variational autoencoder (VAE).

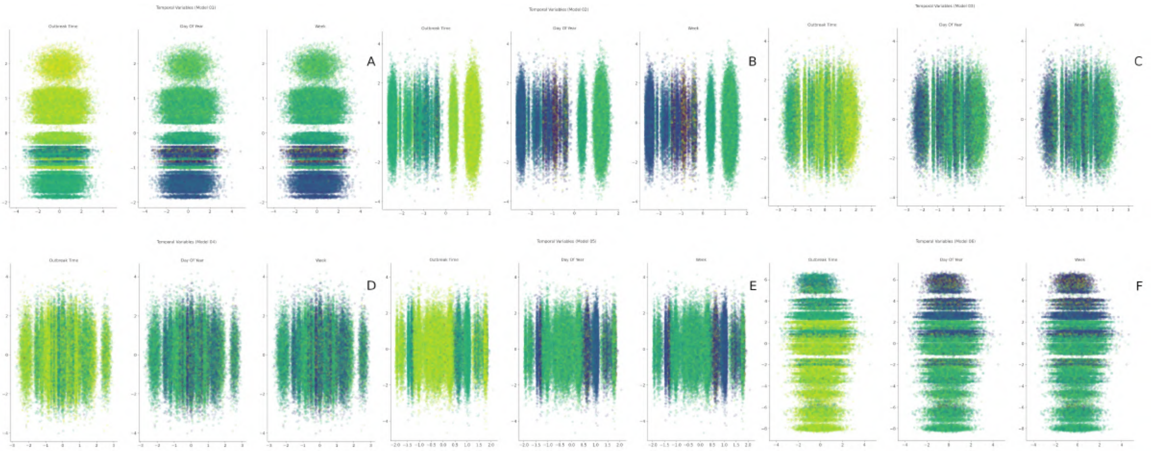

Figure 12. Temporal metadata.. Temporal ordering around a single axis in the different bottleneck representations. A,B,C,D,E,F,G correspond to models 1,2,3,4,5,6 respectively. The temporal component is the main learned component regardless of the learning architecture.

A VAE is a special kind of autoencoder, a neural network architecture that aims to predict the input data. An autoencoder has two main components, an encoder network, and a decoder network. The encoder network performs a downsampling operation, and the decoder an upsampling operation. In the middle of the autoencoder, there is a low dimensional representation of the data or bottleneck representation. [23] A VAE is a special kind of autoencoder, where the bottleneck representation behaves like a normal distribution. This allows create new samples with the decoder network by simply generating normally distributed random numbers. The decoder part of the autoencoder is also referred to as a generative model, as is capable to generate new samples. [14] VAEs offer a suitable tool to learn low dimensional representations of high dimensional data and have been used with success for biological sequences. [26]

Six different VAE models were used to model the different SARS-Cov2 datasets. Model 1 and model 2 use the same input data, but model 1 only uses the stacked frequencies until k=4. Model two uses all the dataset but rearranges the data into a 2d array, allowing the use of computer vision layer architectures such as MLP mixer [20]. Model 3 uses dataset two as input data, while model 4 uses dataset 3 as input data, and both models share the same neural network architecture. Models 5 and 6 also share the same architecture and input data with one difference. Dataset four is also used for model 6 but imported as integer values. This allows training the model with only the elements with higher frequency. Model size ranges from around 30k parameters to 300k parameters. (Figure 9) The size of the bottleneck representation is two, the same in all the different models. KL loss is scaled for all the models and cyclical annealed in all the models except the MLP model. [8]

For training and validation, sequence accession id is used as an identifier to split the data into different datasets. First, the data is split into a train validation data set. The validation dataset will be used to evaluate the model performance. While the Train dataset will be used for five-fold cross-validation.(Figure 8)

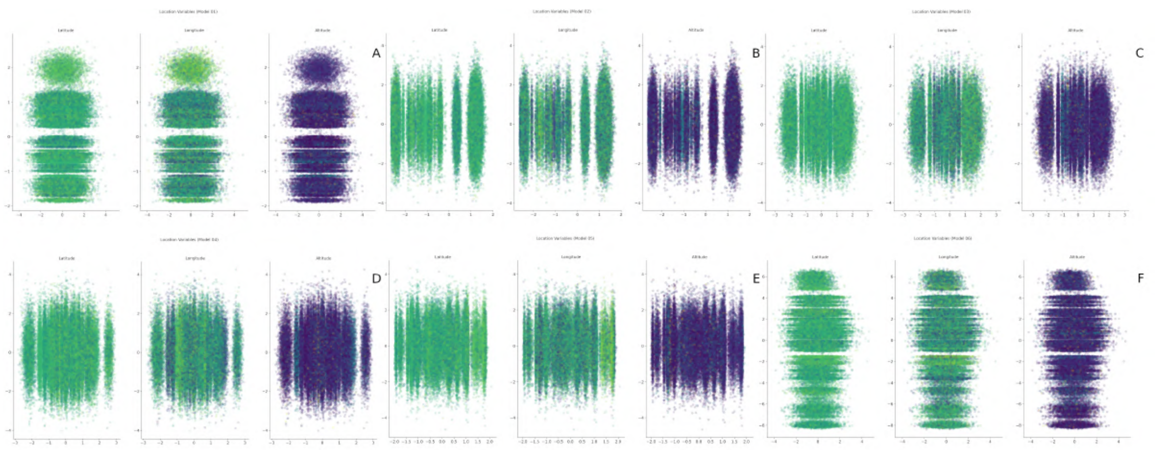

Figure 13. Geographical metadata.

Code is available in two formats, a Github repository that contains all the code used to process the data, create the datasets, train the models, and analyze de data. And also notebook examples hosted in kaggle with a few changes to fit the analysis within the computational constraints of the platform. Code and examples can be downloaded from the following link.

Training time for graph-based encodings takes around 13 hrs while vector-based representations take a couple of hours on a GeForce GTX 1660 Ti Nvidia graphics card. While Kaggle based examples take less than 12hrs to complete with GPU acceleration except for the vector-based model. Simple MLP does not need the use of GPU acceleration and can be completed within 9hrs. Thus full data generation and training on a desktop PC with AMD® Ryzen 5 2600 processor, a GeForce GTX 1660 Ti Nvidia graphics card, and 40GB of RAM ranges roughly between 12hr and 2 days.

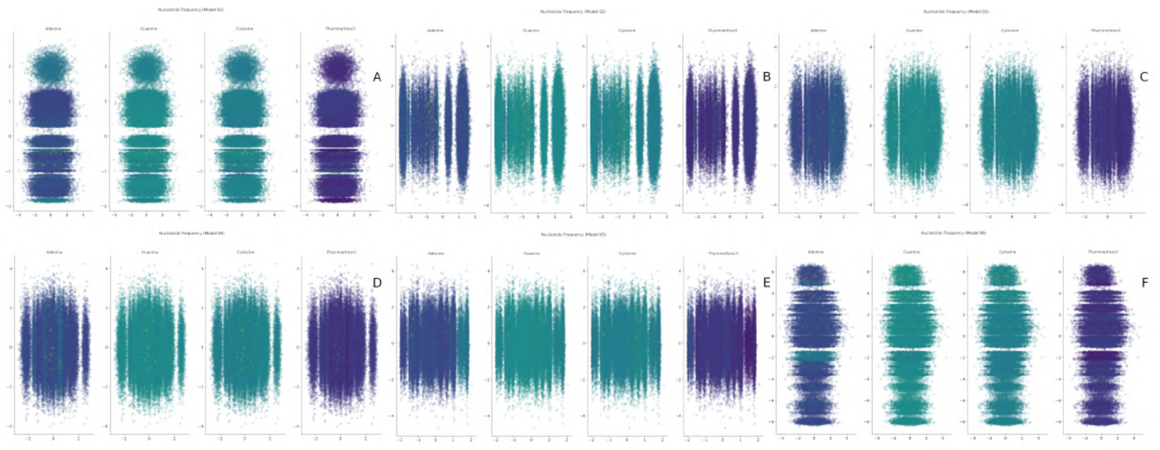

Figure 14. Nucleotide content.

Out of sample reconstruction loss within all the different models ranges from 0.001 to 0.0001. (Figure 10) Performance of small size models remains with small variations compared to larger models. Bottleneck representation shows some kind of structure, particularly ordered over one axis. While others show no clear structure. (Figure 11)

Latent dimension encoding

Sequence modeling of SARS-Cov2 sequences continues to be a topic of active research. However, most of the research is conducted on classification tasks. To predict viral strain in combination with different viral genomes, or to predict immune evasion. [2] [19] Thus understanding time-related changes and adaptations within the SARS-Cov2 sequence is the main underlying task. Luckily, one of the advantages of generative models is that each dimension in the bottleneck representation learns a particular pattern inside the data. However, the learned pattern will depend on the data and the model's capacity to capture such patterns.

Testing probably learned patterns within the bottleneck representation is done by a color encoding scheme. Numerical features within the metadata are used to add a single color to each data point, while categorical features are encoded to values in the range [0,1]. In the case of numerical features, similar colors will represent a closer relationship between the values. While similar colors within categorical features do not share any kind of similarity.

Time encoding defined in three different forms results in similar and well-established patterns for models 1, 2, 5, and 6. (Figure 12) Clear separations within different time-lapses show that one of the learned dimensions within the different models is some sort of temporal component within the sequences. Models 3 and 4 are also capable to make a kind of temporal separation, yet are not as informative as those obtained from the other models.

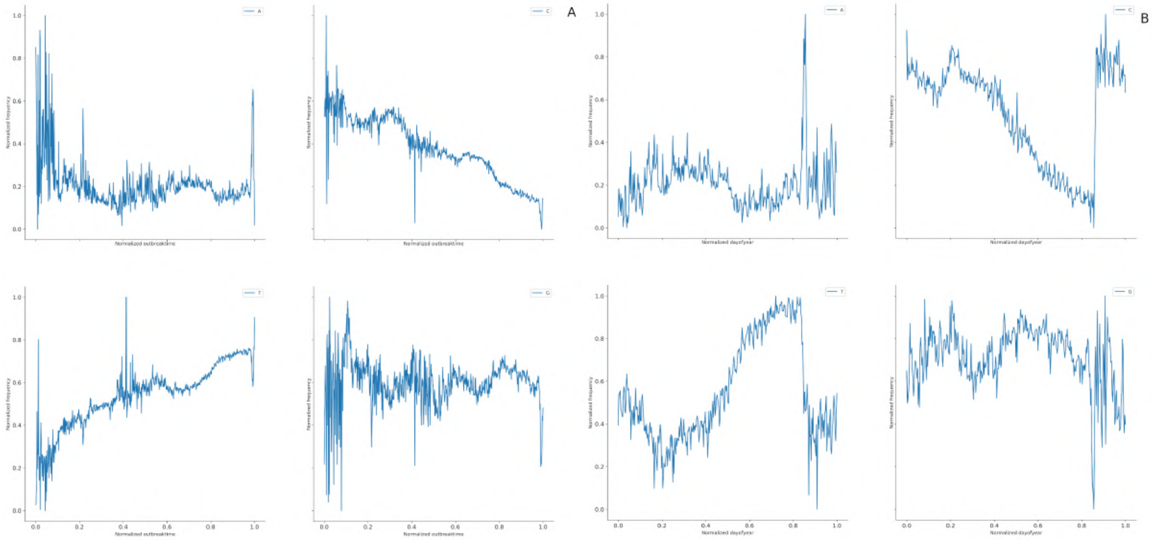

Figure 15. Nucleotide content through time. Panel A shows mean nucleotide content through the pandemic, an upward trend in the overall T/U content, and a downward in the overall C content. Panel B shows mean nucleotide content by day of the year, a small increase in adenine in the first half of the year.

Most models capture some sort of temporal information and this component might be cyclical. Models with a better disentanglement of temporal components show a continuous range of colors with little mixing in between them. Nevertheless, this can be due to sampling bias, the sampling tightly follows the number of cases in the pandemic. The high number of cases also results in a high number of sequenced samples. Subsampling the data and removing the temporal sampling bias results in similar learned representations able to encode a temporal component. Temporal subsampling to remove such bias can be found on kaggle.

Geographical locations do not seem to show any particular pattern within the learned representation. However small clusters of similar colors appear in some cases. (Figure 13)

Changes in the nucleotide composition might be an adaptation mechanism within the SARS-Cov2. The frequency of A and T/U shows a pattern with some resemblance to the one found in the temporal encoding. (Figure 14)

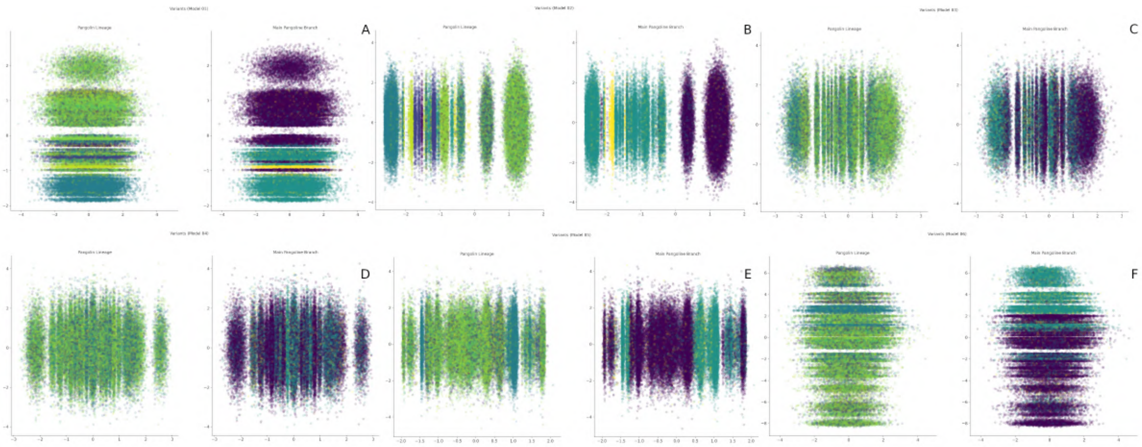

Figure 16. Lineage distribuition. Lineage distribution follows pandemic waves.

Remdesivir a nucleoside analog to adenine is one of the many treatments under evaluation for Covid-19, however, the effect appears to diminish over time. Changes in the adenine content might explain the contradictory efficacy of remdesivir. [12] First positive results were reported in clinical trials that took place from February through March 2020 and the main effect was a faster time to recovery. [24] Subsequent trials started in the middle of march and were unable to replicate such results, raising the concern if there's a place for remdesivir in Covid-19 treatment. [22] Adenine content shows a small increase in the first half of the year, yet the time range in which adenine remains high still matches the time in which the remdesivir trials took place. (Figure 15)

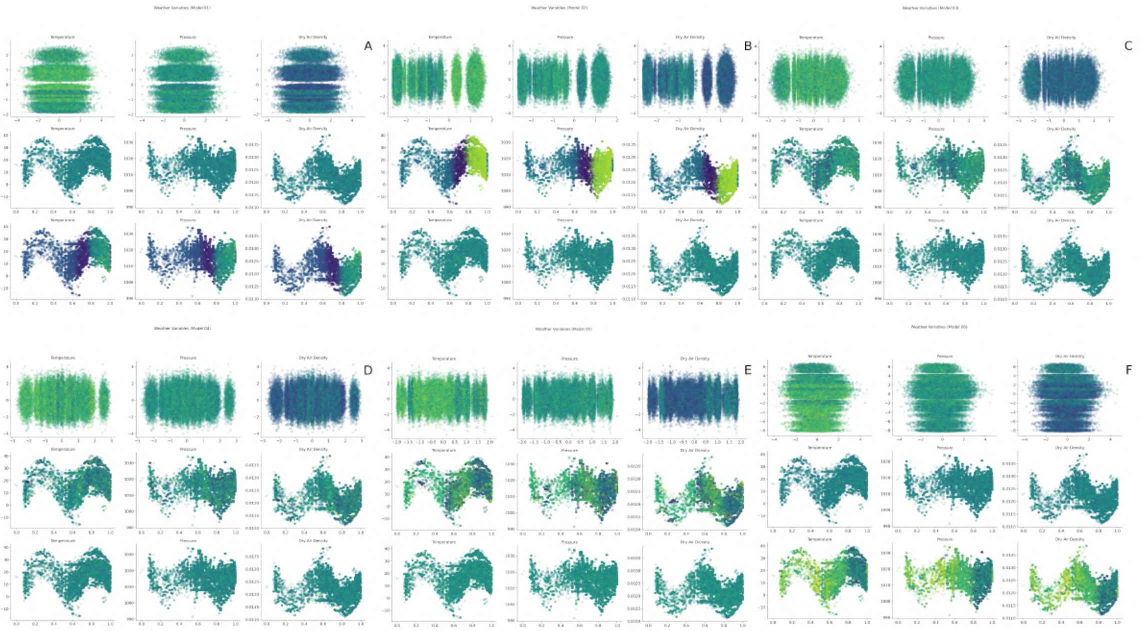

Figure 17. Weather data distribuition. Models with better separation between different time frames show a better agreement with weather variables.

Lineage distribution also follows clear patterns inside the learned representation. Particular lineages can be correctly classified in the models that capture a seasonal component. And the main branch of each lineage appears to be the one with greater accuracy. These models can offer a suitable option for new viral strain classification. (Figure 16)

Temporal adaptations might be a consequence of environmental changes. To test for possible environmental influences average temperature, pressure, and calculated dry air density corresponding to each sample are displayed by color encoding. Models that learn a temporal component have a better correspondence to each environmental variable. Nevertheless, as the seasonal nature of the environment is difficult to point out if the environment has some impact on the viral evolution. Or if environmental adaptations of the host drive such adaptations. (Figure 17)

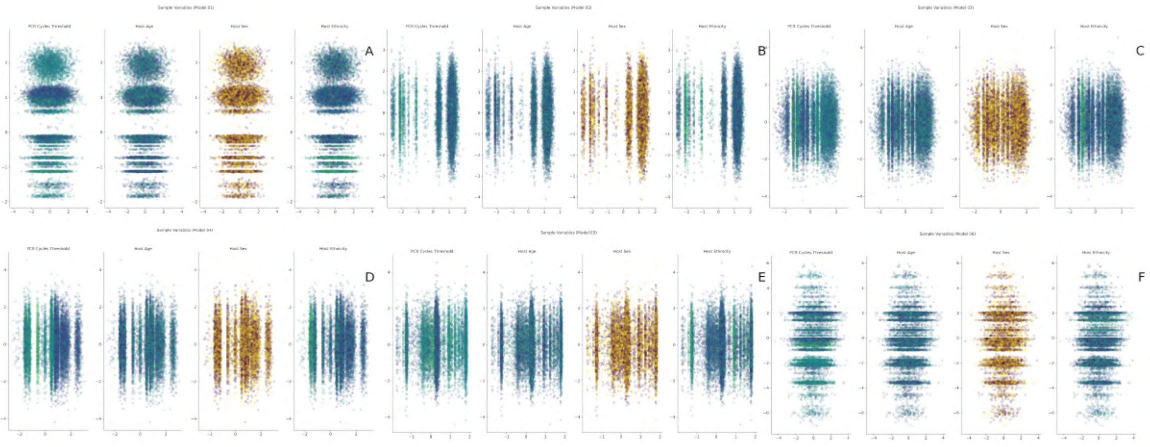

Figure 18. Biosample data.

SARS-Cov2 viral particles are assembled and synthesized inside the host making it the preferred environment of the virus. Host available data shows no particular pattern in age nor sex, and some level of association in the socially misunderstood environmental adaptation referred to as ethnicity. (Figure 18)

PCR ct values also show some degree of association, even in models that did not learn a clear temporal component. Although the final PCR ct value is an average of the different available values it shows that for a subset of samples the threshold is similar. This similarity might be due to individual host conditions.

Latent Space Walks

The output of the encoder results in the bottleneck representation and this and the associated metadata is used to analyze and understand the different patterns inside the bottleneck representation. Knowing the meaning of the different dimensions allows us to test for the impact of changes in that space. Thus changes in the different sequence encodings can be measured by making changes in the bottleneck representations and using those values as input for the decoder part of the model. A latent walk is a simple tool to analyze a continuous range of changes in the bottleneck representation and its effect on the expected output. Thus changes in the sequence encodings at fixed change values can be tested with relative ease. [4]

Latent space walk following the direction of the time-coded dimension of model 1 shows clear changes in the frequency of different 4-mers. Changes in the frequency of different 4-mers are the most noticeable changes. This also changes the frequency of the different remaining components that share some part of that particular 4-mer. (Figure 19)

The latent space walk of model 5 does not show any particular change in the different fragments inside the sequence. While its simplified version model 6, shows that most of the seasonal information is encoded in the structural region of the SARS-Cov2 genome. While the remaining region appears to be encoded without any noticeable change. (Figure 20)

This observation is in agreement with different reports of temporal adaptations, particularly over the spike protein. [6] However by analyzing the full genome the number of nucleotide combinations

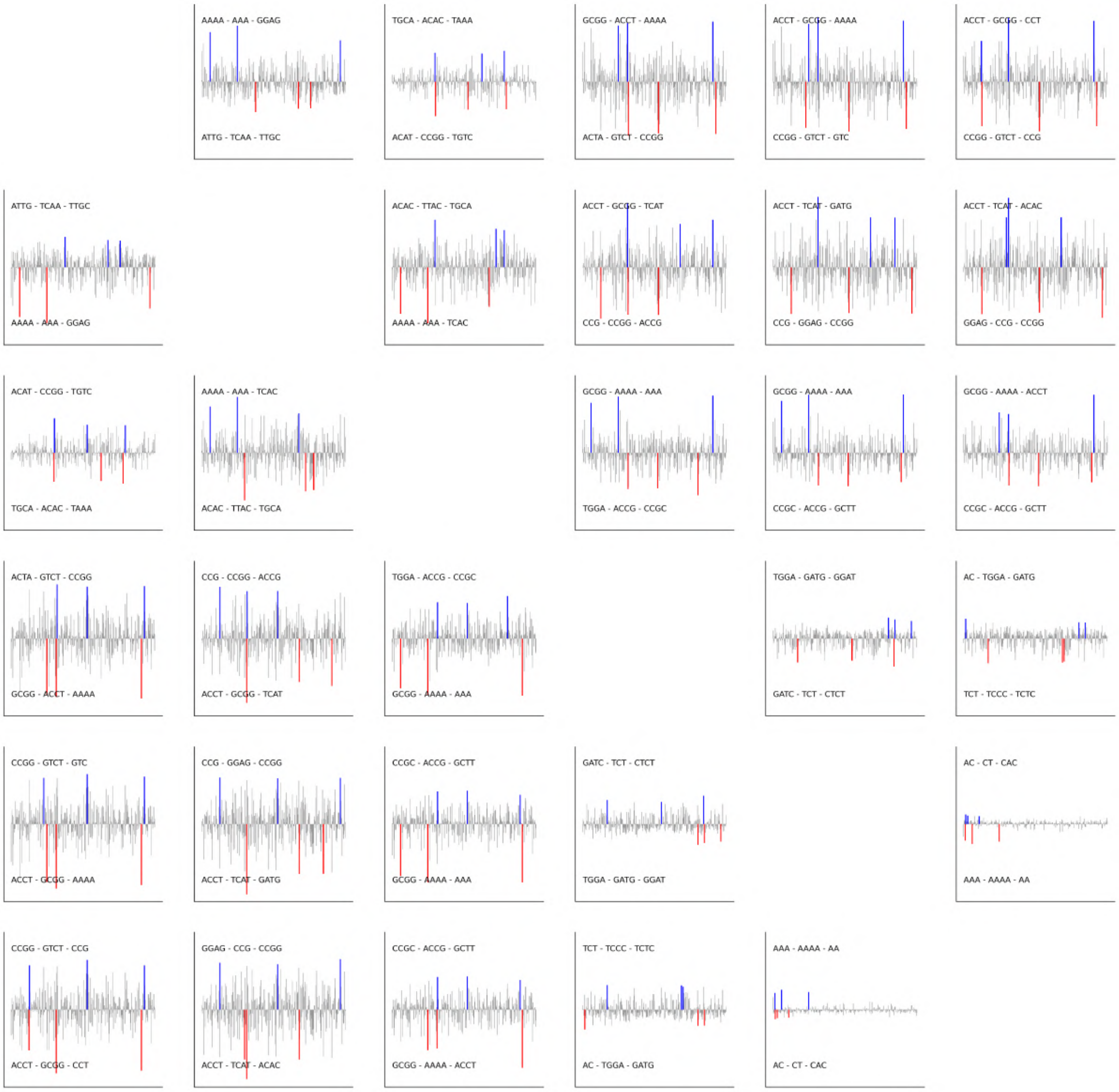

Figure 19. Latent space walk. Latent space walk from Model 1, only changes in the temporal encoding dimension show lack of symmetry and non-symmetrical changes in k-mer composition

that can influence the adaptation is expanded. And the effect of synonym codons inside the genome can also be taken into account. Both kinds of adaptations can be silent to protein base surveillance. [7]

Analyzing the top 4-mer found in either high or low frequency shows a series of 4-mers with different behaviors through time. However, those changes are not symmetrical, as different 4-mers changes throughout the year, thus such k-mers increase in frequency and then return to an equilibrium value. (Figure 21)

While 2-mer composition shows some particular combinations where the frequency appears to have a cyclic component. In some cases, the period is between half and one year. Changes in 2-mer composition can be used as a useful characteristic to design seasonal antiviral treatments. Particularly such combinations that show an upwards trend as the pandemic continues. (Figure 22)

Temporal adaptation followed by 4-mer composition is among the main patterns learned. However, the particular temporal scale that follows does not seem to be the linear yearly time scale.

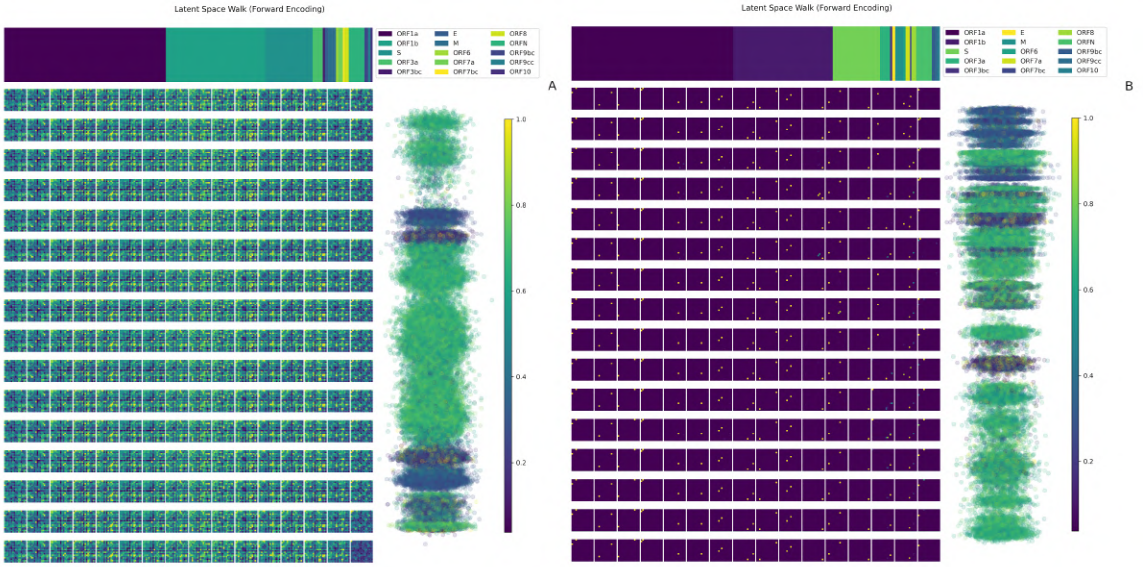

Figure 20. Latent space walk. Latent space walk from model 5 and model 6 shows changes in the structural components of the SARS-Cov2 genome. Also, non-symmetrical changes through the temporal scale.

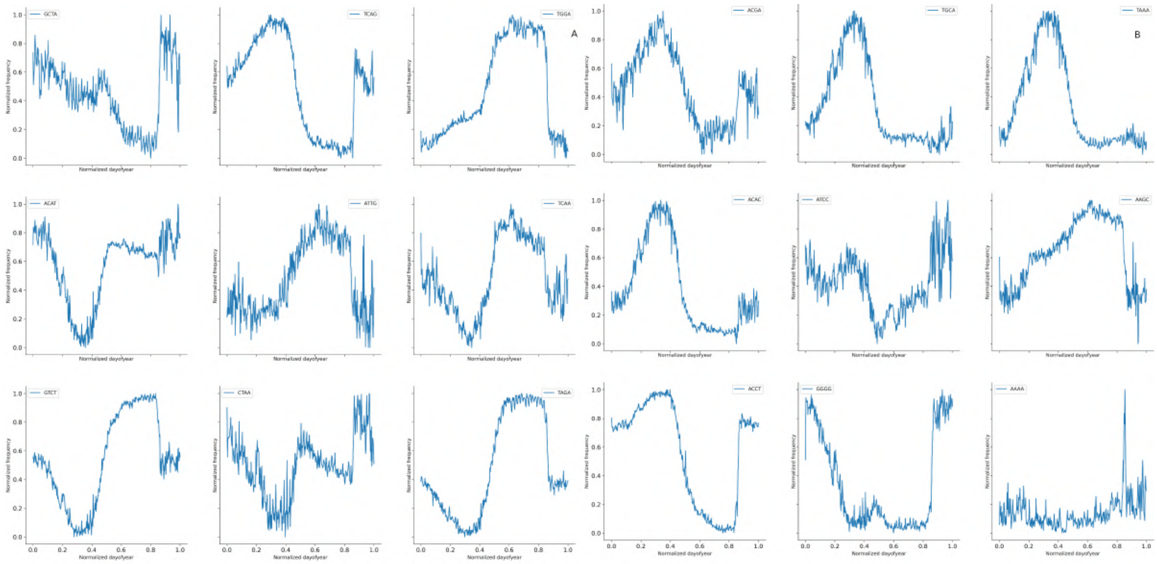

Figure 21. Top changing 4-mers. Mean 4-mer composition through the year of the different top changing 4-mers. Wave-like behavior in the different 4-mers.

A first approximation of the temporal path can be by adding the day length, this will in turn join similar geographical locations to a more accurate temporal representation. Mean daily single-nucleotide frequency shows a particular cycle for Cytosine and Timine/Uracil and a flat path for Adenine and Guanine. While mean single-nucleotide composition through the pandemic shows an upward spiral path for Timine/Uracil and a downward spiral path for Cytosine.

Similar results can be observed in the 2-mer case or the selected 4-mers. Also overlapping between different elements within the path might explain that some clusters in the learned representation contain sequences sampled at time frames with such differences. Going back to

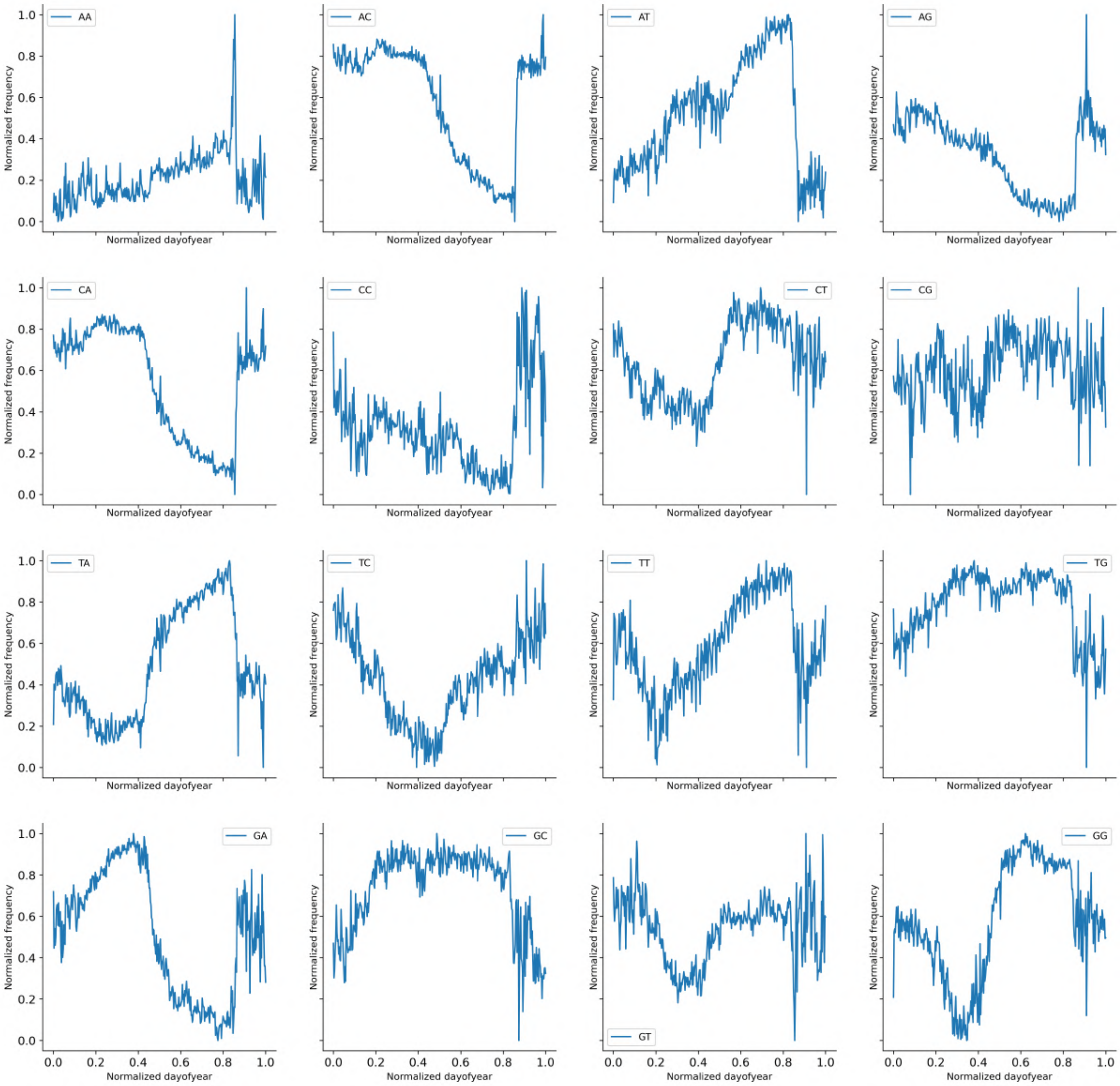

Figure 22. 2-mer change through time. Mean 2-mer composition through the year of the different possible 2-mers. Wave-like behavior with different shapes.

remdesivir, there's a particular spike in adenine composition, particularly around the last and first two months of the year. This narrows the window in which remdesivir might be clinically helpful both time and location-wise.

Although this might point out to a clear cyclical or seasonal component followed by SARS-Cov2 sampling bias remains high towards the second year of the pandemic. This could lead to more accurate statistics for the second year. Also, total pandemic paths show deviation from the cyclical path observed in yearly estimates.

This leads to at least two particular scenarios. SARS-Cov2 follows a seasonal yearly trend constrained by total day length spiraling towards a long-term evolutionary goal. And that particular evolutionary path is observed by the overall pandemic time estimate. Or SARS-Cov2 is trying to adapt to a specific seasonal path constrained by day length. It matches in some parts of the path but continues to try to adjust. (Figure 23)

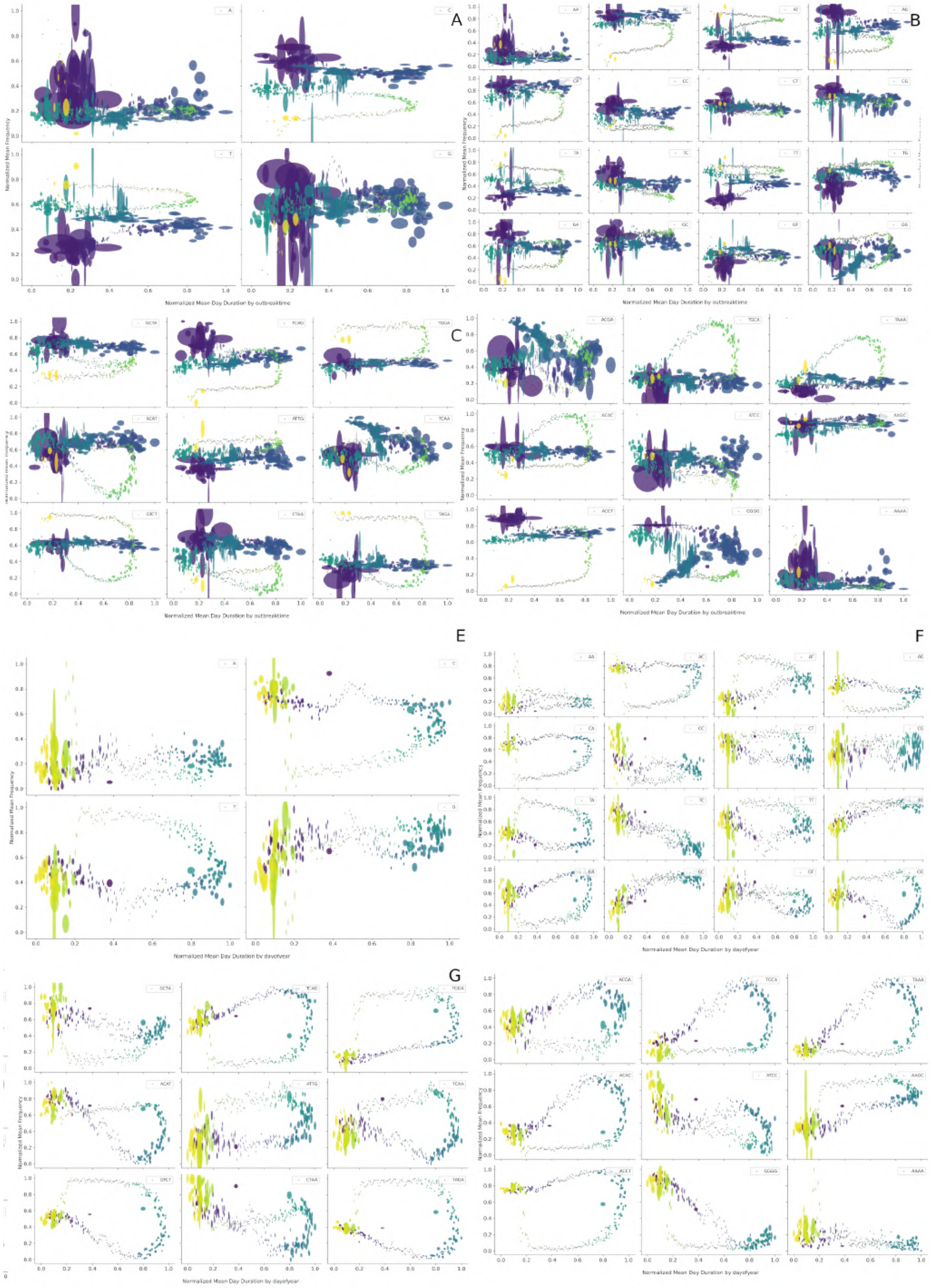

Figure 23. SARS-Cov2 composition path. Mean day duration vs mean composition through time. Shape size shows the mean standard error in day duration or composition, while color shows the linear selected time scale. For panels A, B, C, and D mean day duration and composition are calculated throughout the entire pandemic. While for the remaining panels mean values are calculated by the day of the year.

Transcripts screening

Adaptation might be the result of mimicking or following well-established temporal strategies of the host. As the nucleotide pool inside the cell is tightly regulated. And mechanisms of nucleotide starvation have been described as a mechanism of antiviral defense. [3] A possible way to predict the infectivity and adaptation checkpoints of SARS-Cov2 could be to map genes with similar nucleotide compositions of SARS-Cov2.

To test that hypothesis a series of filters are applied to the reference transcripts data. This will approximate the available mRNA composition and the available nucleotide pools. Nucleotide frequency is calculated for each mRNA from reference transcripts and compared to the mean SARS-Cov2 composition. Selected sequences are then compared to the 2D representation that contains the temporal patterns both simplified and full encoding resulting from models 5 and 6. From this selection, the sequences are then classified into two: coding sequences and non-coding sequences.

A total of 507 unique coding sequences were the result of the screening. Location and molecular function are retrieved from UniProt from each sequence. This results in 135 reviewed records from UniProt. Retrieved proteins can be clustered into nine different groups based on the available gene ontology information, and function information. Nevertheless, as each record is not complete and the resulting clusters might be a combination of lack of information and similarity. (Figure 24)

From the 135 selected records, 54 were involved in SARS-Cov2 infection in cell culture or patient samples. Nevertheless, there were no obvious functional associations or relations in the filtered proteins making it difficult to suggest mechanistic implications. However, some particular targets might provide some insights about SARS-Cov2 pathogenicity.

ILST6, the receptor of IL6 a cytokine involved in several immune functions, has been found up-regulated in COVID-19 patients. In combination with high levels of IL6 can induce a strong inflammatory response. If the nucleotide content of infected cells with high ILST6 expression matches SARS-Cov2 nucleotide needs it might also inadvertently favor the assembly of new SARS-Cov2 viral particles. [5]

Nsp1, a SARS-Cov2 protein, induced several factors involved in transcriptional regulation. Among them was CREBRF one of the genes selected by the previous filter. Although the link is weak, determining if SARS-Cov2 can induce or follow gene expression patterns with similar nucleotide composition is a hypothesis worth pursuing further. [28]

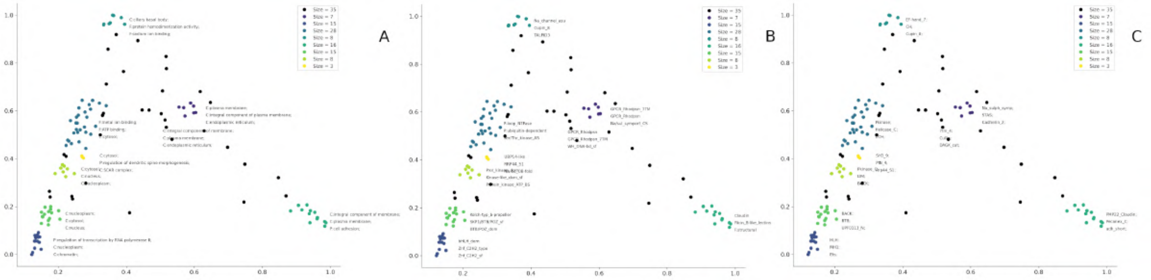

Figure 24. Function and location clustering. Function and location clustering. Location data is transformed to a 2D embedding and top occurring elements a are displayed. A top occurring GO terms, B top occurring InterPro terms, and C top occurring Pfam terms.

Conclusions

Fragment-based modeling of large sequences in the framework of this paper remains unable to reconstruct the original sequence. However, they offer a suitable option for sequence summarization at a low computational cost.

The current theoretical background for graph-based sequence encoding in this particular work remains insufficient. Further development will help to determine the correct use of sequence embeddings and improve other analysis tasks.

Changes in the 4-mer composition drive the temporal adaptation found within the SARS-Cov2 sequences. Nevertheless, even with a finite set of 4-mer driving these changes combinations of those could lead to a high number of possible SARS-Cov2 variants.

Design of new drugs and treatments targeting the relatively constant components of SARS-Cov2 offers a better long-term strategy for the current Covid-19 pandemic.

Uracil analogs might provide a better long-term solution as SARS-Cov2 composition follows an upward trend in uracil content. And the development of dinucleotide mimetics might provide a suitable seasonal treatment.

Although the temporal path is constrained by day length, it remains unknown the long-term evolutionary goal of SARS-Cov2. Extrapolations towards disease severity or transmissibility cannot be made with the currently available data.

Temporal effects in SARS-Cov2 research remain unknown, yet as experimental conditions are tightly regulated it can be assumed that to some degree results are skewed towards a particular time frame or environmental condition.

Application of similar analysis to existing datasets of viral sequences can lead to a better understanding of viral evolution and adaptation. As well as the design of better surveillance mechanisms for emerging pathogens.

Although there must be similarities between the different COVID-19 outbreaks, the previous analysis and its conclusions it's mostly applicable to the US.

Further refinement and improvement of the analysis pipeline will lead to new methods and tools for the fast characterization of emerging viruses.

Acknowledgments

I thank all the personnel involved in sample acquisition, processing, and labeling SARS-Cov2 samples. NCBI staff and policymakers who decided to make openly and publicly available SARS-Cov2 genomes. Without free and easy access to SARS-Cov2 sequences and metadata, this will not be possible.

Out of theme

Throughout the different tests made before arriving at the final form of this paper, there were many things left behind. Here I enlist some of the things that were tough and worked about, but did not make it to the final cut.

Although all the changes described above seem to be large in scale, all of them are small in the comparison with the overall sequence. Most of the normalizations are to present the data as clearly as possible and trying not to deal with hardcoded axis values and scales. Removing normalization from the code results in the same graphical patterns but different changing scales.

Models 3 and 4 have little temporal information and the clusters remain without a clear explanation of what they mean. Most of the metadata was iteratively added to find an explanation of what those clusters meant. Perhaps they encode a large shared pattern but I did not pursue the idea any further.

Encoding for model 4 was designed as an effort to try to add as much data as possible to the sequence representation. Tried to expand and find an explanation for what was encoded by model 2, but again without success.

For the adjacency matrix-based representations the data is scaled by its matrix norm and then min-max normalized. The main reason behind this was that data for models 3 and 4 resulted in the difference between two adjacency matrices. This first normalization allows reducing the effect of the magnitude of values in each matrix. Data form model 5 is also scaled by its matrix norm but for that particular model that step might not be necessary.

Model 6 was the result of mistakenly loading the data for model 5 in a different format. By loading it as integer values most of the values were flattened to zero except the max values within the matrix. That allowed for simple visualization of the viral temporal adaptations. A happy little accident if you will.

Following the results of models 5 and 6 attempts to design a fragment-based model were unsuccessful. Perhaps the data used for such models were inadequate. Or the sequence works as a whole rather than as a collection of small independent fragments.

The main viral replication model that I was thinking about was that the negative strand of the viral sequence anneals to small mRNA fragments resulted in the mRNA digestion by nsp15. This in turn allows the virus to capture seasonal changes through mRNA availability. Roughly SARS-Cov2 genome can be constructed by 2220 fragments of mRNA digested by nsp15, with a mean and median size of 13 and 9. However, this model led to a massive combinatorial problem. Even with those constraints continues to be difficult to assemble the full genome even if the replication model is correct.

Although there is no evidence a long-term goal for viral evolution might be to maximize the shortest common substring between the viral and host genome. Larger similar fragments lead to larger similar protein fragments. More host-like protein fragments can lead to a higher probability to develop autoimmunity.

Most of the analysis was an iterative trial and error process rather than a single iteration. If in some parts the information does not seem to follow the same train of thought is because they are not from the same train of thought.

References

1. S. Abubaker Bagabir, N. K. Ibrahim, H. Abubaker Bagabir, and R. Hashem Ateeq. Covid-19 and artificial intelligence: Genome sequencing, drug development and vaccine discovery. *Journal of Infection and Public Health*, 15(2):289–296, 2022.

2. S. Ali, B. Sahoo, N. Ullah, A. Zelikovskiy, M. Patterson, and I. Khan. A k-mer based approach for sars-cov-2 variant identification, 2021.

3. D. Ayinde, N. Casartelli, and O. Schwartz. Restricting hiv the samhd1 way: through nucleotide starvation. *Nature Reviews Microbiology*, 10(10):675–680, Oct 2012.

4. S. A. Billings. *Generation and Utilization of Latent Spaces for Prediction and Interpretation.* PhD thesis, University of Cambridge, 2018.

5. L. T. Bui, N. I. Winters, M.-I. Chung, C. Joseph, A. J. Gutierrez, A. C. Habermann, T. S. Adams, J. C. Schupp, S. Poli, L. M. Peter, C. J. Taylor, J. B. Blackburn, B. W. Richmond, A. G. Nicholson, D. Rassl, W. A. Wallace, I. O. Rosas, R. G. Jenkins, N. Kaminski, J. A. Kropski, N. E. Banovich, and the Human Cell Atlas Lung Biological Network. Chronic lung diseases are associated with gene expression programs favoring sars-cov-2 entry and severity. *bioRxiv*, 2021.

6. K. Caetano-Anollés, N. Hernandez, F. Mughal, T. Tomaszewski, and G. Caetano-Anollés. The seasonal behaviour of covid-19 and its galectin-like culprit of the viral spike. Methods in Microbiology. Academic Press, 2021.

7. F. Chen, P. Wu, S. Deng, H. Zhang, Y. Hou, Z. Hu, J. Zhang, X. Chen, and J.-R. Yang. Dissimilation of synonymous codon usage bias in virus–host coevolution due to translational selection. *Nature Ecology & Evolution*, 4(4):589–600, Apr 2020.

8. H. Fu, C. Li, X. Liu, J. Gao, A. Celikyilmaz, and L. Carin. Cyclical annealing schedule: A simple approach to mitigating kl vanishing, 2019.

9. E. L. Hatcher, S. A. Zhdanov, Y. Bao, O. Blinkova, E. P. Nawrocki, Y. Ostapchuck, A. A. Schäffer, and J. R. Brister. Virus Variation Resource – improved response to emergent viral outbreaks. *Nucleic Acids Research*, 45(D1):D482–D490, 11 2016.

10. A. M. Ille, H. Lamont, and M. B. Mathews. The central dogma revisited: Insights from protein synthesis, crispr, and beyond. *WIREs RNA*, n/a(n/a):e1718.

11. H. Iuchi, T. Matsutani, K. Yamada, N. Iwano, S. Sumi, S. Hosoda, S. Zhao, T. Fukunaga, and M. Hamada. Representation learning applications in biological sequence analysis. *Computational and Structural Biotechnology Journal*, 19:3198–3208, 2021.

12. Y. Jiang, D. Chen, D. Cai, Y. Yi, and S. Jiang. Effectiveness of remdesivir for the treatment of hospitalized covid-19 persons: A network meta-analysis. *Journal of Medical Virology*, 93(2):1171–1174, 2021.

13. U. Kamath, K. De Jong, and A. Shehu. Effective automated feature construction and selection for classification of biological sequences. *PLOS ONE*, 9(7):1–14, 07 2014.

14. D. P. Kingma and M. Welling. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019.

15. D. Lee, O. Redfern, and C. Orengo. Predicting protein function from sequence and structure. *Nature Reviews Molecular Cell Biology*, 8(12):995–1005, Dec 2007.

16. M. Levo and E. Segal. In pursuit of design principles of regulatory sequences. *Nature Reviews Genetics*, 15(7):453–468, Jul 2014.

17. Z. Li, Y. Chen, D. Mu, J. Yuan, Y. Shi, H. Zhang, J. Gan, N. Li, X. Hu, B. Liu, B. Yang, and W. Fan. Comparison of the two major classes of assembly algorithms: overlap–layout–consensus and de-bruijn-graph. *Briefings in Functional Genomics*, 11(1):25–37, 12 2011.

18. V. Neduva and R. B. Russell. Linear motifs: Evolutionary interaction switches. *FEBS Letters*, 579(15):3342–3345, 2005. Budapest Special Issue.

19. J. J. Park and S. Chen. Metaviromic identification of discriminative genomic features in sars-cov-2 using machine learning. *Patterns*, 3(2):100407, 2022.

20. I. O. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit, M. Lucic, and A. Dosovitskiy. Mlp-mixer: An all-mlp architecture for vision. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 24261–24272. Curran Associates, Inc., 2021.

21. M. Tschannen, O. Bachem, and M. Lucic. Recent advances in autoencoder-based representation learning, 2018.

22. A. Vitiello, F. Ferrara, and R. L. Porta. Remdesivir and covid-19 infection, therapeutic benefits or unnecessary risks? *Irish Journal of Medical Science (1971 -)*, 190(4):1637–1638, Nov 2021.

23. Y. Wang, H. Yao, and S. Zhao. Auto-encoder based dimensionality reduction. *Neurocomputing*, 184:232–242, 2016. RoLoD: Robust Local Descriptors for Computer Vision 2014.

24. Y. Wang, D. Zhang, G. Du, R. Du, J. Zhao, Y. Jin, S. Fu, L. Gao, Z. Cheng, Q. Lu, Y. Hu, G. Luo, K. Wang, Y. Lu, H. Li, S. Wang, S. Ruan, C. Yang, C. Mei, Y. Wang, D. Ding, F. Wu, X. Tang, X. Ye, Y. Ye, B. Liu, J. Yang, W. Yin, A. Wang, G. Fan, F. Zhou, Z. Liu, X. Gu, J. Xu, L. Shang, Y. Zhang, L. Cao, T. Guo, Y. Wan, H. Qin, Y. Jiang, T. Jaki, F. G. Hayden, P. W. Horby, B. Cao, and C. Wang. Remdesivir in adults with severe covid-19: a randomised, double-blind, placebo-controlled, multicentre trial. *The Lancet*, 395(10236):1569–1578, May 2020.

25. T. Warnow. *Revisiting Evaluation of Multiple Sequence Alignment Methods*, pages 299–317. Springer US, New York, NY, 2021.

26. R. Wei and A. Mahmood. Recent advances in variational autoencoders with representation learning for biomedical informatics: A survey. *IEEE Access*, 9:4939–4956, 2021.

27. L. Xu, R. Magar, and A. Barati Farimani. Forecasting covid-19 new cases using deep learning methods. *Computers in Biology and Medicine*, 144:105342, 2022.

28. S. Yuan, L. Peng, J. J. Park, Y. Hu, S. C. Devarkar, M. B. Dong, Q. Shen, S. Wu, S. Chen, I. B. Lomakin, and Y. Xiong. Nonstructural protein 1 of sars-cov-2 is a potent pathogenicity factor redirecting host protein synthesis machinery toward viral rna. *Molecular Cell*, 80(6):1055–1066.e6, 2020.

29. A. Zargari Khuzani, M. Heidari, and S. A. Shariati. Covid-classifier: an automated machine learning model to assist in the diagnosis of covid-19 infection in chest x-ray images. *Scientific Reports*, 11(1):9887, May 2021.

30. A. Zielezinski, S. Vinga, J. Almeida, and W. M. Karlowski. Alignment-free sequence comparison: benefits, applications, and tools. *Genome Biology*, 18(1):186, Oct 2017.