Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Computers in Biology and Medicine 138 (2021) 104915

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Understanding mutation hotspots for the SARS-CoV-2 spike protein using Shannon Entropy and K-means clustering

Baishali Mullick [b,1], Rishikesh Magar [a,1], Aastha Jhunjhunwala [c], Amir Barati Farimani [a,*]

[a] Department of Mechanical Engineering, Carnegie Mellon University, Pittsburgh, PA, 15213, USA
[b] Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA, 15213, USA
[c] Department of Chemical Engineering, Carnegie Mellon University, Pittsburgh, PA, 15213, USA

ARTICLE INFO

Keywords:
SARS-CoV-2
Mutations
Clustering
Shannon entropy

ABSTRACT

The SARS-CoV-2 virus like many other viruses has transformed in a continual manner to give rise to new variants by means of mutations commonly through substitutions and indels. These mutations in some cases can give the virus a survival advantage making the mutants dangerous. In general, laboratory investigation must be carried to determine whether the new variants have any characteristics that can make them more lethal and contagious. Therefore, complex and time-consuming analyses are required in order to delve deeper into the exact impact of a particular mutation. The time required for these analyses makes it difficult to understand the variants of concern and thereby limiting the preventive action that can be taken against them spreading rapidly. In this analysis, we have deployed a statistical technique Shannon Entropy, to identify positions in the spike protein of SARS Cov-2 viral sequence which are most susceptible to mutations. Subsequently, we also use machine learning based clustering techniques to cluster known dangerous mutations based on similarities in properties. This work utilizes embeddings generated using language modeling, the ProtBERT model, to identify mutations of a similar nature and to pick out regions of interest based on proneness to change. Our entropy-based analysis successfully predicted the fifteen hotspot regions, among which we were able to validate ten known variants of interest, in six hotspot regions. As the situation of SARS-COV-2 virus rapidly evolves we believe that the remaining nine mutational hotspots may contain variants that can emerge in the future. We believe that this may be promising in helping the research community to devise therapeutics based on probable new mutation zones in the viral sequence and resemblance in properties of various mutations.

Contributions of the work:

1. The paper proposes a computational methodology to identify potential mutational hotspots in spike protein of SARS-CoV-2. The high throughput methodology can also identify some of the dangerous mutations emerging in the distant future
2. Understand and identify the similarities and patterns among the different type of mutations using clustering analysis. Such an analysis may possibly help biologists to better understand the relationships between SARS-CoV-2 mutations.

1. Introduction

The SARS-CoV-2 virus has rapidly evolved by continually mutating, affecting more than 180 million people across the globe. Ever since the genome sequence of SARS-CoV-2 became available, mutations at several sites in the genome have been identified raising concerns regarding enhanced transmissibility of the virus [1]. The mutating nature of the virus has inspired global efforts from research community to actively track and understand the emergence of variants of concern[2–4]. One of the first mutation that rapidly spread throughout the world, mutation D614G, was first reported in April 2020 [5]. This mutation has now been classified under several lineages and is found to be a factor in increased transmission of the virus [6–9]. The discovery of this mutation was followed by identification of a series of mutations in the virus belonging

* Corresponding author.
E-mail address: barati@cmu.edu (A. Barati Farimani).
[1] Equal Contributions.

https://doi.org/10.1016/j.compbiomed.2021.104915
Received 27 July 2021; Received in revised form 17 September 2021; Accepted 29 September 2021
Available online 5 October 2021

to the B. 1.1.7 lineage, which was first found in the Southeast of England [10]. The mutations, namely A222V, S477 N, N501, H69, N439K, Y453F,11S98F, D80Y, A626S, V1122L, have been noted as variants of interest in many studies [10–13] and are the focus of this work as well. These variants were selected because they were marked as the Variant Under Investigation SARS-CoV-2 VUI 202012/01 (Variant Under Investigation, the year 2020, month 12, variant 01) by different studies done in the United Kingdom [14]. The mutation, A222V belongs to the B.1.177 lineage and has been noted to have a dominating presence in European countries[15,16]. N439K and Y453F have been found to have a higher binding affinity to the hACE2 receptor and are noted to reduce the neutralizing potential of antibodies specific to SARS-CoV-2 [17–19]. N439K often co-occurs with 69–70 deletion in the spike protein, the effect of this combined double mutation is being investigated by researchers (COVID-19 Genomics UK consortium, 2021; [20]. N501Y is the causative factor in the increased infectiousness of the disease [21]. The numerous effects of such mutations on the increased transmissibility and lethality of SARS-CoV-2, make it imperative to study these mutations and understand their effects[22].

To tackle the COVID -19 pandemic, efforts from the researchers have involved exploring traditional paradigm of *in-vitro* experimentation and data analysis-based methodologies like machine learning. Data driven modelling techniques, with their ability to analyze large amounts of data, build a functional mapping between the input parameters and output. This paper explores the use of data-driven methodologies to understand the mutations in the SARS-CoV-2 spike proteins. To understand and identify the mutation hotspots we have examined the sequence entropy and its correlation with experimentally identified variants of concern.

Tomaszewski et al., defined mutational entropy as a measure of molecular heterogeneity of the SARS-CoV-2 proteome which is estimated from the positional variance in these sequences [7]. In our work, we measure the positional variance in the sequence of the SARS-CoV-2 spike proteins by calculating Shannon Entropy. In case of proteins, Shannon entropy is shown to have a strong correlation with protein structural entropy [23], and can provide insights into the compositional stability of the proteins. The Shannon entropy is also directly proportional to the inverse packing density of proteins [24], and the packing density is further related to increased mutagenesis. Moreover, higher local flexibility regions have an increased value of entropy and are prone to mutations [21]. Our study explores these relationships of Shannon entropy to estimate the mutational hotspots in the SARS-CoV-2 spike protein. Higher value of entropy at a position in the sequence is indicative of increased randomness at that site whereas low value of entropy at a certain site is indicative of an increased stability and decreased randomness at the said location.

Apart from identifying the hotspots of interest, we also analyze the similarity of these mutations by employing a k-means clustering algorithm. To generate the embedding for the clustering algorithm we leverage the protein sequence data by using language modeling approaches. Through transfer learning, some of the highly successful models in the Natural Language Processing (NLP) domain have been applied to protein sequence to generate meaningful representations that can be used in tasks like structure prediction [25]. We used the Prot-BERT language modeling to represent these spike protein sequences in the form of semantic rich embeddings [26]. The Prot-BERT model has been trained on 80 billion amino acids, representing wide variety of protein sequences. The embeddings generated via the Prot-BERT model can be used for different downstream tasks. In our work, we use embeddings to determine the similarities between mutations using unsupervised machine learning techniques. This analysis will help in understanding the relationships between the mutations and assist the research community to tackle the virus.

1.1. Related work

Machine learning models have been used in many ways to study and understand the different aspects of COVID-19 pandemic. These models have been previously used for forecasting the COVID-19 cases [27–29], propose the potential antibodies [30], understand the possible evolutions of the virus [31], understand the economic and social effects of social distancing [32,33], understand the efficiency of lockdowns [34], study the transmission and spread of the virus [35,36]. Data driven models have also been used to analyze the SARS-CoV-2 mutations. In their paper [37], use techniques topological like persistent homology to understand the SARS-CoV-2 mutations and uncover some underlying patterns. In another study [38], develop the Informative Subtype Markers (ISM) to visualize and analyze the spread of different mutated SARS-CoV-2 sequences.

2. Methods

2.1. Data

To understand the effect of the mutations we focus only on the spike protein of the virus sequence. We select the spike protein region because it is the major component of the SARS-CoV-2 virus that is responsible for eliciting host immune responses of neutralizing antibodies. It is the presence of this spike protein on the antigen that allows it to interact and penetrate the host cells. Therefore, more attention to spike protein has been given in the analysis of the mutations of the SARS-CoV-2 virus. To this end, we collect the spike protein data from the GISAID server to analyze the effect of the mutations on the spike protein on its transmissibility. We downloaded three hundred eleven thousand two hundred and fifty-six spike protein sequences from the GISAID server (https://www.gisaid.org/) on January 3, 2020 [11,39]. The comprehensive dataset had sequences related to the SARS-CoV-1 virus too, therefore the first stage of preprocessing involved the elimination of sequences that were not from 2020. This resulted in a dataset comprising three hundred ten thousand five hundred and ten sequences. Most of these sequences are comprised of 1273 amino acids, with maximum length being 1278 amino acids. To ensure uniformity in our calculation of the positional entropy, the ones with length less than 1278 were made up to length 1278 by appending the relevant number of 'X's to the end of the gene sequence for the entropy analysis. The original spike protein sequence found in Wuhan is referenced from Zhao et al. [1] and the mutations in all the collected sequences in the data are analyzed with respect to this sequence. There was a large presence of repeated spike protein sequences found in different countries, so we decided to curate the data further and create data with only the unique sequences as featurizing the same sequence twice using Prot-BERT would have been redundant. We found fifty-three thousand eight hundred and ninety-eight belonging to prime variants of interest that are unique sequences of the spike protein. Subsequently, this dataset was used to generate embedding via the ProtBERT Model. These embeddings were further used to carry out unsupervised machine learning analysis. To understand the spread of the data and visualize it, we generated the plot using t-SNE [40] shown in Fig. 1.

Further, we also analyze the geographical locations and the general distribution of the countries that were a part of the dataset we found that United Kingdom and Denmark contributed to over 50% of the mutation sequences in the dataset with 140458 mutation sequences from United Kingdom and 20346 from Denmark. These two countries have proactively studied the different mutations and made the data available for public use via the GISAID server. To analyze the mutation sequence data from other countries, a distribution of the dataset comprising of countries with more than 200 but less than 5000 mutation sequences is shown in Fig. 2.

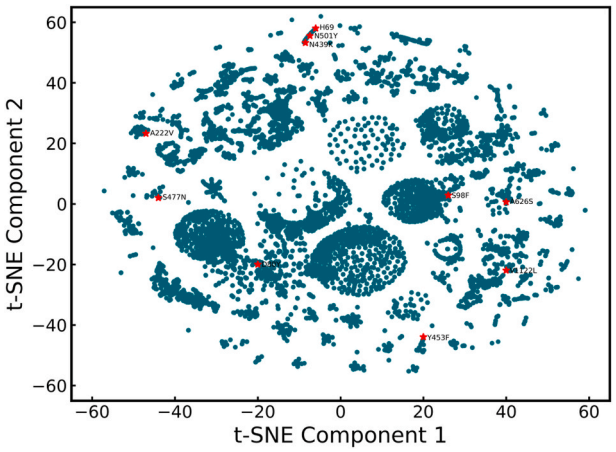

Fig. 1. t-SNE plot capturing the distribution of the data collected from the GISAID server. Some of the variants of concern like N439K, N501Y are clustered near each other. From the t-SNE, we can easily infer that the SARS-CoV-2 mutations have unique characteristics.

2.2. Positional entropy calculations

The positional entropy is a measure of the randomness at the given position in the sequence [41]. To calculate the positional entropy for our dataset we use Shannon Entropy formulation stated in Equation (1) [42]:

$$H(i) = -\sum_{k \in L} P_k(i) \times \log_2 P_k(i) \quad (1)$$

Where L is a list of all possible amino acids in all the sequences *and* $P_k(i)$ is the probability of finding the kth amino acid at that position.

We use equation (1) to find the positional entropy for all the positions in the SARS-CoV-2 spike protein sequence. Using the dataset obtained from the GISAID server, we first pre-process the data using Biopython[43] to extract the sequences from the FASTA file downloaded from the server. We found that the length of the spike protein sequence varied from 1270 to 1278, the distribution of the sequence lengths is shown in Fig. S1. We also observed that the positions that contain ambiguous sites or unidentified amino acid in the spike protein sequence have been denoted with character "X" in the dataset. These positions with character "X" are handled by a masking operation that calculates the entropy without considering them [38]. We proceed by calculating the positional entropy values using equation (1) and all the values for the positional entropy are stored in an array.

To identify the regions of high entropy that can possibly be associated with harmful mutations, we use a running mean (window length = 15, step size = 1), here the first positional index of the window gets assigned the value of the running mean. In the running mean calculation, we don't consider the first 60 and last 60 amino acids in the sequences because of the sequencing uncertainty. After calculating the running mean (window length = 15, step size = 1) for positional entropy, we stored it in another array. The array containing all the running means is then sorted and top 100 entropy values in the sequence are selected. Subsequently, we define the hotspots in the sequence as having $\geq$2 consecutive high entropy positions among the top-100 positional entropy values. For example: 210 and 211 both belong to the top 100 positional entropy values, and hence region 210–224 has been identified as a hotspot. To ensure both the positions (210 & 211) are included, we select the lowest index (210) as the start position of the hotspot and next 15 positions (included in the running mean) are considered as the hotspot (210–224). Additional details about the distribution of sequence lengths (Fig. S1) in the data and the starting positions of running mean windows for the top 100 positional entropy values are provided in the supplementary information (Table S1).

2.3. Prot-BERT model

The Prot-BERT trained on the UniRef100 dataset was used to generate sequence embeddings [26]. The Prot-BERT model has 30 layers, 16 attention heads, and embedding hidden size 1024. The Prot-BERT model was chosen because the embeddings generated have been used for different downstream tasks successfully increasing our

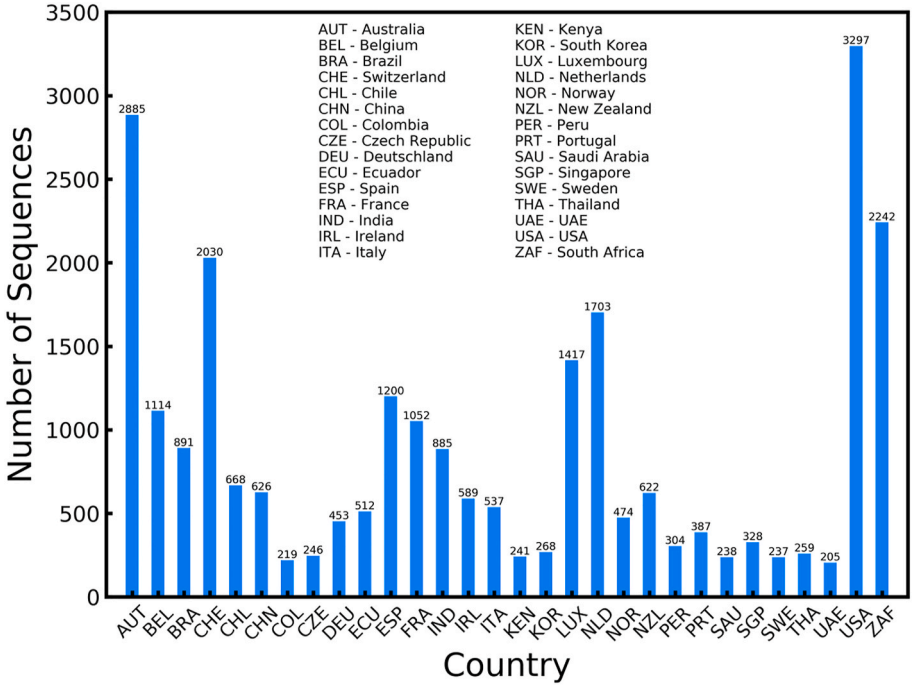

Fig. 2. Plot showing the distribution of the sequences in the data. Apart from United Kingdom and Denmark, the other countries actively tracking the variants of concern include USA, Australia, South Africa, and Switzerland.

confidence in using the same. We generate the embedding for the spike proteins of the mutated sequences using the pre-trained model on the hugging face api [44]. The Hugging face interface allows the users to easily use the pre-trained models on various Natural Language Processing (NLP) tasks. The curated data containing the unique sequences of spike protein were entered in the pre-trained Prot-BERT model and an embedding of size 1024 for every sequence. These embeddings are then used to study similarities and understand distributions between the mutations via K-Means clustering.

2.4. K means

Clustering is an unsupervised learning technique used to group a collection of unlabeled data sharing similarities. Each cluster comprises data sharing common traits which are distinct from members of other clusters, thereby resulting in clusters with high internal homogeneity and high external heterogeneity [45]. Clustering can be broadly classified into two categories, hierarchical and non-hierarchical clustering.

The k-means clustering technique used in this study is a non-hierarchical clustering approach. This technique involves defining the number of clusters 'k'. Each cluster is represented by a central location defined as the centroid, C_{kj} where k is the cluster number and j are the number of attributes. The algorithm allocates each data point to the nearest cluster by minimizing the distance from centroid. It starts off by randomly assigning centroids and thereafter continues as an iterative process to optimize the centroid locations depending on the points assigned to that cluster. This process continues until there is no further change in the centroid values or until the maximum number of iterations is reached [46].

Clustering is one of the most important data mining techniques to group unlabeled data based on common traits. In this work, we used K means clustering to group the different mutations based on similarities in properties. The embeddings generated using the ProtBert model were used as features for the clustering model.

To perform k-means clustering we use the scikit-learn library, that builds k-means model under the hood after entering the model parameters [47,48]. The number of clusters chosen for this task was 10, based on the number of different mutation types being 10 and also because we got the highest silhouette score of 0.7228 [49] when using 10 clusters. We also implemented the MST-kNN clustering technique but the algorithm did not perform very well, it had a very low silhouette score of −0.7638 and hence was not used for any further clustering analysis. We use the silhouette scores metric as it is a measure of how well an algorithm can differentiate between different clusters in the data. The score varies from −1 to +1 and high silhouette score indicates that the datapoints have been clustered appropriately, with similar datapoints clustered together and dissimilar datapoints clustered differently. Other parameters for k-means such as the maximum number of iterations was chosen to be 1000 and the total number of initializations was chosen as 50 after multiple trials with other values in order to stabilize the cluster formation.

3. Results

3.1. Positional entropy

The advantage of analyzing the entropy lies in the fact that sequential entropy is correlated to molecular motility is an important factor for the mutation [7,23,24]. Furthermore, studies have found a significant relationship between these high entropy hotspot regions of the viral sequence and enhanced virulence in the mutations associated with these regions, which have had a crucial role in the evolution of this disease. Hence, these sites are regions of interest in vaccine development and medicine formulation[38]. We calculated the positional entropy for all positions of the spike protein genomic sequence and have estimated the mutational hotspot regions in these viral sequences. Table 1 highlights some of these regions of interest we have identified which correspond to some of the most dominant mutations that have been noted in various countries. From this analysis, we have noted that the regions of interest have successfully captured the D614G mutation, which is one of the most dominant mutation and is found to enhance the replication of SARS-CoV-2 in the lung cells [50]. The regions of interest also captured the following mutations - A222V, N439K, Y453F, S477 N, N501, D614G and V1122L [12].

Apart from the above mutations, the following other mutations have also been correctly identified in our hotspots - E484K, T478K, and L452R. It has been shown that for the mutation, E484K along with the some mutations from B.1.1.7 lineage requires increased amounts of antibody serum to prevent infection [51] making it especially dangerous. Interestingly, our methodology is capable of capturing some of the potentially harmful mutations that may emerge in the future. For example: Our model that uses sequence data before 2020 identifies one of the hotspot regions from 439 to 453. A mutation of significance, L452R which was first identified by the California Dept of Public Health on 17th Jan 2021 [52] and was later found to be dominant mutation in the months of April and May 2021 worldwide. Similarly, another mutation E484K belonging to the B.1.25 family was recognized as variant of concern was recognized in South Africa in April 2021 [53]. This mutation lies in the region 473–487 which includes another mutation of significance S477 N [16,54]. This emergence of variants of concern from hotspot regions identified by our methodology demonstrates the accurate prediction of Shannon entropy based analysis.

To further illustrate the positional entropy hotspots, we have plotted the positional entropy for the entire sequence of the spike protein of SARS-CoV-2 in Fig. 3. Based on our analysis, we found nine other hotspot regions including 329–343, 386–400, 425–439, 530–544, 700–714, 763–777, 905–919, 955–968, 1172–1186. Based on validation analysis presented in Table 1 it is likely that the new mutation of concern may emerge in these hotspot regions.

To structurally understand the mutations further, we also identified the regions where the dangerous mutations belong in the structure of the spike protein. The analysis was based on study done by Huang et al., where they identify the different regions in the spike protein based on the positions in the sequence [55]. It must be noted that there are seven possible dangerous mutations in the receptor binding domain of the spike protein, these mutations are possibly more lethal because of their location on the binding interface. The locations of these mutations on the spike protein have been presented in Table 2.

We also validate the mutations in Table 1 by using– EV mutation[56] methodology that determines the favorability of a mutation by calculating the prediction epistatic score. The data for mutation effect using EV mutation for SARS-CoV-2 is available on the server created by Ref. [57], we used the data from this server to analyze the epistatic mutation effect predict for mutations presented in Table 1. The novel aspect of the EV mutation method is its ability to take into account epistasis by taking into consideration the interactions between all pairs of amino acids residues in the neighborhood to quantify the mutational effects. A higher value of the prediction score using EV mutation

Table 1
Hotspots found by analyzing the positional entropy. To determine a hotspot region a running mean (window length = 15, step size = 1) is calculated and top 100 value are selected. We found six such regions of interest in our analysis in which ten mutations of interest emerged.

Hotspots	Mutation
211–225	A222V
439–453	N439K, L452R, Y453F
473–487	S477 N, T478K, E484K
487–501	N501Y
602–616	D614G
1121–1135	V1122L

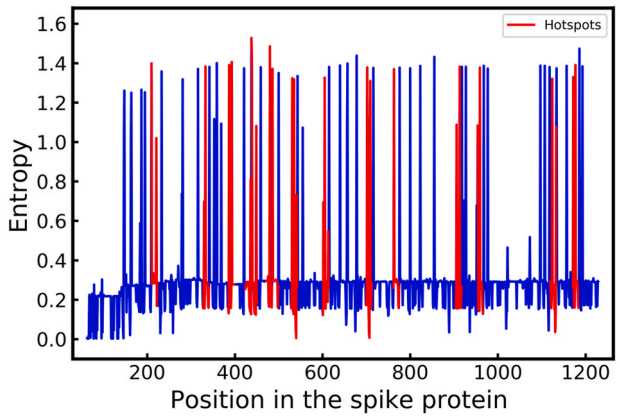

Fig. 3. Variation of entropy and the position in the spike protein. Hotspots with higher likelihood of mutagenesis and high entropy have been marked in red in the plot. The red regions(hotspots) have the maximum mean entropy over a window of length 15. The blue regions in the plot indicate the regions of relatively lower mean entropy over the window of length 15. According to positional entropy analysis the dangerous spike protein mutations are more likely to emerge from the hotspots (red regions).

Table 2
Location of the mutations in the spike protein of the SARS-CoV-2, we have 3 regions of the spike protein where mutations can be located.

Spike Protein Region	Mutation
N – Terminal domain	A222V
Receptor-Binding Domain	N439K, L452R, Y453F, S477 N, T478K, E484K, N501Y
Heptapeptide repeat sequence	V1122L

indicates a highly favorable mutation. The analysis using EV mutation has been presented in Table 3.

Among the ten different mutations in Table 1, Table 3 presents the EV mutation score for seven different mutations. The data for S477 N, E484K and N501Y is unavailable on the server (Nathan Rollins*, Kelly Brock*, Joshua Rollins* et al., 2020), and hence is not presented in Table 3. We observe that A222V and T478K are highly favorable mutations as they have the highest possible prediction epistatic score among all mutations for the wild-type residue (A for site 222 and T for site 478). The D614G mutations is also highly favorable, and mutations Y453F, V1122L and N439K may be considered as moderately favorable. On the other hand, the mutation L452R may not be as favorable based on prediction epistatic score. The EV mutation scores validate most

Table 3
Analysis of the SARS-CoV-2 mutations using EV mutation, the prediction epistatic score is an indicator of whether a mutation is fit or not fit. The higher score indicates that the mutation indicates that the mutation is a better fit. The third column indicates the rank among all the possible mutations at the site. The possible values for rank range from 1 to 19 as there are 20 amino acids and a single amino acids can mutate into 19 other amino acids. The rank depends on the EV mutation score, highest score will get rank-1 that indicates the mutation is highly favorable and lowest score gets rank-19 indicates that mutation is not favorable according to EV mutation calculations.

Mutation	Prediction epistatic score	Rank among all mutation possibilities
A222V	0.5465	1
N439K	−3.8605	10
L452R	−6.1483	15
Y453F	−6.5665	7
T478K	0.4154	1
D614G	−4.7144	2
V1122L	−6.9294	9

mutations identified in the hotspots from our methodology in Table 1, further indicating the calculating the positional entropy of the sequence can be a useful metric for identifying future mutation hotspots.

The positional entropy formulation developed in this work used the data from the year 2020 and yet was able to identify some of the mutations that emerge later in April and May 2021 such as E484K and L452R validating our methodology further. We believe that our method may potentially be used to identify the dangerous mutations in advance and aid in the fight against the pandemic.

3.2. *Clustering with K-means*

The clustering analysis was done on the embeddings generated from the Prot BERT model. The embeddings for all the sequences are a 2D array of shape (sequence length, 1024) where 1024 is the hidden dimension of the model. Subsequently, we applied mean pooling to the sequence length dimension of the embeddings and generate a vector of dimension 1024 for each sequence. This 1024-dimensional vector is used for k-means clustering analysis.

The cluster centers resulting from k-means clustering correspond to the different mutation types, thereby verifying our assumption that the different cluster types get grouped separately. We find that 7 out of 10 different mutations are identified as cluster centers with a few repeats. On analyzing the spike protein sequences that form the clusters and the sequence representative of the cluster center, we find that in most cases most of the sequences are identified to be of the same type as the cluster center whereas in most other cases the mutation type of the cluster center is amongst the top 3 mutation types present in the cluster, the other two types of possibly similar characteristics (Table 4). For example, from the plots (Fig. 4) show the clusters of S477 N and N439K have a majority of S477 N and N439K components. Furthermore, A222V has the second highest count in the cluster representing S477 N (Fig. 4) indicating similarities between them. D80Y is one of the majorities in the N439K cluster, thereby implying similarity in characteristics. In a study done by Ref. [58], it was found that A222V and S477 N are both stabilizing mutations thereby validating our findings that these two mutations may have some similar characteristics. This similarity analysis between the mutations is significant because when designing therapeutics that can counter new mutations understanding characteristics of mutations computationally can save a lot of experimental time and accelerate the therapeutic development process.

4. Conclusion

In this study, we developed a methodology to determine the hotspots for mutations in spike protein sequences of SARS-CoV-2. This study can enable us to know variants of interests beforehand so that therapeutics can be developed for them. We found fifteeen regions of interest in the sequence of the spike protein that may be the potential hotspots for novel mutations in SARS-CoV-2. Six of these hotspots contain ten mutations which have already been flagged as possibly more transmissible by the previous research. Interestingly, some of the new emerging variants from India and South Africa which have been marked dangerous in April 2021 and May 2021 were identified by our methodology even though we use the sequence data on the GISAID server before December

Table 4
Clusters where the top 3 dominant mutations in the cluster concur with the cluster center mutation. The top-3 dominant mutations are most likely to be similar in characteristics to the mutation in cluster.

Cluster Centers	Dominant Mutations in the Cluster
S477 N	**S477N**, A222V, S98F, N439K
N439K	**N439K**, D80Y, N501Y, H69-70
N501Y	D80Y, N439K, **N501Y**, H69
A222V	V1122L**, A222V**, N501Y, S477 N

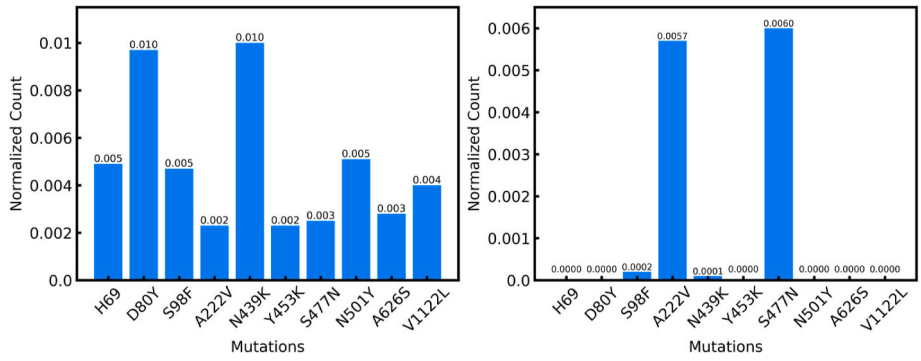

Fig. 4. A.) Clustering analysis for N439K mutation on the spike protein of SARS-CoV-2. After analyzing the cluster with cluster center as N439K we can conclude that the D80Y may have similar characteristics to that of N439K. B.) Clustering analysis for S477 N mutation on the spike protein of SARS-CoV-2. The majority of sequences in this cluster belong to the mutation S477 N and the next highest number is that of A222V suggesting similarity between them.

2020. Identifying hotspots beforehand may have implications in the development of therapeutics and be aware of the potential threats posed by the mutations in the virus. We also use the unsupervised learning-based clustering technique k-means to find the similarities between the variants of interests that have previously been found to be dangerous. The encode the protein sequences we use the Prot-BERT model and use features generated by it, for the k-means analysis. Clustering the mutation variants based on similarity reduces redundancy of time and resources, similar treatment techniques can be implemented for mutations that fall into the same cluster. One of the results of our analysis was the similarity between the S477 N and the A222V mutations, it implies that these mutations share common traits and occurrences and may be subjected to similar treatment strategies.

Declaration of competing interest

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

Acknowledgements

The authors would like to thank Prakarsh Yadav and Parisa Mollaei for their useful inputs and comments on the paper. This work is supported by the start-up fund provided by CMU Mechanical Engineering and support from Center for Machine Learning and Health (CMLH).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2021.104915.

References

[1] F. Wu, S. Zhao, B. Yu, Y.-M. Chen, W. Wang, Z.-G. Song, Y. Hu, Z.-W. Tao, J.-H. Tian, Y.-Y. Pei, M.-L. Yuan, Y.-L. Zhang, F.-H. Dai, Y. Liu, Q.-M. Wang, J.-J. Zheng, L. Xu, E.C. Holmes, Y.-Z. Zhang, A new coronavirus associated with human respiratory disease in China, Nature 579 (2020) 265–269, https://doi.org/10.1038/s41586-020-2008-3.

[2] I. Alam, A. Radovanovic, R. Incitti, A.A. Kamau, M. Alarawi, E.I. Azhar, T. Gojobori, CovMT: an interactive SARS-CoV-2 mutation tracker, with a focus on critical variants, Lancet Infect. Dis. 21 (2021) 602, https://doi.org/10.1016/S1473-3099(21)00078-5.

[3] A.T. Chen, K. Altschuler, S.H. Zhan, Y.A. Chan, B.E. Deverman, COVID-19 CG enables SARS-CoV-2 mutation and lineage tracking by locations and dates of interest, eLife 10 (2021), e63409, https://doi.org/10.7554/eLife.63409.

[4] Y. Xing, X. Li, X. Gao, Q. Dong, MicroGMT: a mutation tracker for SARS-CoV-2 and other microbial genome sequences, Front. Microbiol. 11 (2020) 1502, https://doi.org/10.3389/fmicb.2020.01502.

[5] B. Korber, W.M. Fischer, S. Gnanakaran, H. Yoon, J. Theiler, W. Abfalterer, N. Hengartner, E.E. Giorgi, T. Bhattacharya, B. Foley, K.M. Hastie, M.D. Parker, D. G. Partridge, C.M. Evans, T.M. Freeman, T.I. de Silva, C. McDanal, L.G. Perez, H. Tang, A. Moon-Walker, S.P. Whelan, C.C. LaBranche, E.O. Saphire, D. C. Montefiori, A. Angyal, R.L. Brown, L. Carrilero, L.R. Green, D.C. Groves, K. J. Johnson, A.J. Keeley, B.B. Lindsey, P.J. Parsons, M. Raza, S. Rowland-Jones, N. Smith, R.M. Tucker, D. Wang, M.D. Wyles, Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus, Cell 182 (2020) 812–827, https://doi.org/10.1016/j.cell.2020.06.043, e19.

[6] S. Laha, J. Chakraborty, S. Das, S.K. Manna, S. Biswas, R. Chatterjee, Characterizations of SARS-CoV-2 mutational profile, spike protein stability and viral transmission, Infect. Genet. Evol. 85 (2020) 104445, https://doi.org/10.1016/j.meegid.2020.104445.

[7] T. Tomaszewski, R.S. DeVries, M. Dong, G. Bhatia, M.D. Norsworthy, X. Zheng, G. Caetano-Anollés, New pathways of mutational change in SARS-CoV-2 proteomes involve regions of intrinsic disorder important for virus replication and release, Evol. Bioinf. Online 16 (2020), https://doi.org/10.1177/1176934320965149, 117693432096514.

[8] E. Volz, V. Hill, John T. McCrone, A. Price, D. Jorgensen, Á. O'Toole, J. Southgate, Robert Johnson, B. Jackson, F.F. Nascimento, S.M. Rey, S.M. Nicholls, R. M. Colquhoun, A. da Silva Filipe, J. Shepherd, D.J. Pascall, R. Shah, N. Jesudason, K. Li, R. Jarrett, N. Pacchiarini, M. Bull, L. Geidelberg, I. Siveroni, I. Goodfellow, N. J. Loman, O.G. Pybus, D.L. Robertson, E.C. Thomson, A. Rambaut, T.R. Connor, C. Koshy, E. Wise, Nick Cortes, J. Lynch, S. Kidd, M. Mori, D.J. Fairley, T. Curran, J. P. McKenna, H. Adams, C. Fraser, T. Golubchik, D. Bonsall, Catrin Moore, S. L. Caddy, F.A. Khokhar, M. Wantoch, N. Reynolds, B. Warne, J. Maksimovic, K. Spellman, K. McCluggage, M. John, R. Beer, S. Afifi, S. Morgan, A. Marchbank, A. Price, C. Kitchen, H. Gulliver, I. Merrick, J. Southgate, M. Guest, R. Munn, T. Workman, T.R. Connor, W. Fuller, C. Bresner, L.B. Snell, T. Charalampous, G. Nebbia, R. Batra, J. Edgeworth, S.C. Robson, A. Beckett, K.F. Loveson, D. M. Aanensen, A.P. Underwood, C.A. Yeats, K. Abudahab, B.E.W. Taylor, M. Menegazzo, G. Clark, W. Smith, M. Khakh, V.M. Fleming, M.M. Lister, H. C. Howson-Wells, Louise Berry, T. Boswell, A. Joseph, I. Willingham, P. Bird, T. Helmer, K. Fallon, C. Holmes, J. Tang, V. Raviprakash, S. Campbell, N. Sheriff, M.W. Loose, N. Holmes, Christopher Moore, M. Carlile, V. Wright, F. Sang, J. Debebe, F. Coll, A.W. Signell, G. Betancor, H.D. Wilson, T. Feltwell, C. J. Houldcroft, S. Eldirdiri, A. Kenyon, T. Davis, O. Pybus, L. du Plessis, A. Zarebski, J. Raghwani, M. Kraemer, S. Francois, S. Attwood, T. Vasylyeva, M.E. Torok, W. L. Hamilton, I.G. Goodfellow, G. Hall, A.S. Jahun, Y. Chaudhry, M. Hosmillo, M. L. Pinckert, I. Georgana, A. Yakovleva, L.W. Meredith, S. Moses, H. Lowe, F. Ryan, C.L. Fisher, A.R. Awan, J. Boyes, J. Breuer, K.A. Harris, J.R. Brown, D. Shah, L. Atkinson, J.C.D. Lee, A. Alcolea-Medina, N. Moore, Nicholas Cortes, R. Williams, M.R. Chapman, L.J. Levett, J. Heaney, D.L. Smith, M. Bashton, G.R. Young, J. Allan, J. Loh, P.A. Randell, A. Cox, P. Madona, A. Holmes, F. Bolt, J. Price, S. Mookerjee, A. Rowan, G.P. Taylor, M. Ragonnet-Cronin, F.F. Nascimento, D. Jorgensen, I. Siveroni, Rob Johnson, O. Boyd, L. Geidelberg, E.M. Volz, K. Brunker, K.L. Smollett, N.J. Loman, J. Quick, C. McMurray, J. Stockton, S. Nicholls, W. Rowe, R. Poplawski, R.T. Martinez-Nunez, J. Mason, T.I. Robinson, E. O'Toole, J. Watts, C. Breen, A. Cowell, C. Ludden, G. Sluga, N.W. Machin, S.S. Y. Ahmad, R.P. George, F. Halstead, V. Sivaprakasam, E.C. Thomson, J. G. Shepherd, P. Asamaphan, M.O. Niebel, K.K. Li, R.N. Shah, N.G. Jesudason, Y. A. Parr, L. Tong, A. Broos, D. Mair, J. Nichols, S.N. Carmichael, K. Nomikou, E. Aranday-Cortes, N. Johnson, I. Starinskij, A. da Silva Filipe, D.L. Robertson, R. J. Orton, J. Hughes, S. Vattipally, J.B. Singer, A.D. Hale, L.R. Macfarlane-Smith, K. L. Harper, Y. Taha, B.A.I. Payne, S. Burton-Fanning, S. Waugh, J. Collins, G. Eltringham, K.E. Templeton, M.P. McHugh, R. Dewar, E. Wastenge, S. Dervisevic, R. Stanley, R. Prakash, C. Stuart, N. Elumogo, D.K. Sethi, E. J. Meader, L.J. Coupland, W. Potter, C. Graham, E. Barton, D. Padgett, G. Scott, E. Swindells, J. Greenaway, A. Nelson, W.C. Yew, P.C. Resende Silva, M. Andersson, R. Shaw, T. Peto, A. Justice, D. Eyre, D. Crooke, S. Hoosdally, T. J. Sloan, N. Duckworth, S. Walsh, A.J. Chauhan, S. Glaysher, K. Bicknell, S. Wyllie, E. Butcher, S. Elliott, A. Lloyd, R. Impey, N. Levene, L. Monaghan, D.T. Bradley, E. Allara, C. Pearson, P. Muir, I.B. Vipond, R. Hopes, H.M. Pymont, S. Hutchings, M.D. Curran, S. Parmar, A. Lackenby, T. Mbisa, S. Platt, S. Miah, D. Bibby, C. Manso, J. Hubb, M. Chand, G. Dabrera, M. Ramsay, D. Bradshaw, A. Thornton, R. Myers, U. Schaefer, N. Groves, E. Gallagher, D. Lee, D. Williams, N. Ellaby, I. Harrison, H. Hartman, N. Manesis, V. Patel, C. Bishop, V. Chalker, H. Osman,

A. Bosworth, E. Robinson, M.T.G. Holden, S. Shaaban, A. Birchley, A. Adams, A. Davies, A. Gaskin, A. Plimmer, B. Gatica-Wilcox, C. McKerr, Catherine Moore, C. Williams, D. Heyburn, E. De Lacy, E. Hilvers, F. Downing, G. Shankar, H. Jones, H. Asad, J. Coombes, J. Watkins, J.M. Evans, L. Fina, L. Gifford, L. Gilbert, L. Graham, M. Perry, M. Morgan, M. Bull, M. Cronin, N. Pacchiarini, N. Craine, R. Jones, R. Howe, S. Corden, S. Rey, S. Kumziene-Summerhayes, S. Taylor, S. Cottrell, S. Jones, S. Edwards, J. O'Grady, A.J. Page, J. Wain, M.A. Webber, A. E. Mather, D.J. Baker, S. Rudder, M. Yasir, N.M. Thomson, A. Aydin, A.P. Tedim, G. L. Kay, A.J. Trotter, R.A.J. Gilroy, N.-F. Alikhan, L. de Oliveira Martins, T. Le-Viet, L. Meadows, A. Kolyva, M. Diaz, A. Bell, A.V. Gutierrez, I.G. Charles, E. M. Adriaenssens, R.A. Kingsley, A. Casey, D.A. Simpson, Z. Molnar, T. Thompson, E. Acheson, J.A.H. Masoli, B.A. Knight, A. Hattersley, S. Ellard, C. Auckland, T. W. Mahungu, D. Irish-Tavares, T. Haque, Y. Bourgeois, G.P. Scarlett, D. G. Partridge, M. Raza, C. Evans, K. Johnson, S. Liggett, P. Baker, S. Essex, R. A. Lyons, L.G. Caller, S. Castellano, R.J. Williams, M. Kristiansen, S. Roy, C. A. Williams, P.L. Dyal, H.J. Tutill, Y.N. Panchbhaya, L.M. Forrest, P. Niola, J. Findlay, T.T. Brooks, A. Gavriil, L. Mestek-Boukhibar, S. Weeks, S. Pandey, Lisa Berry, K. Jones, A. Richter, A. Beggs, C.P. Smith, G. Bucca, A.R. Hesketh, E. M. Harrison, S.J. Peacock, Sophie Palmer, C.M. Churcher, K.L. Bellis, S.T. Girgis, P. Naydenova, B. Blane, S. Sridhar, C. Ruis, S. Forrest, C. Cormie, H.K. Gill, J. Dias, E.E. Higginson, M. Maes, J. Young, L.M. Kermack, N.F. Hadjirin, D. Aggarwal, L. Griffith, T. Swingler, R.K. Davidson, A. Rambaut, T. Williams, C.E. Balcazar, M. D. Gallagher, Á. O'Toole, S. Rooke, B. Jackson, R. Colquhoun, J. Ashworth, V. Hill, J.T. McCrone, E. Scher, X. Yu, K.A. Williamson, T.D. Stanton, S.L. Michell, C. M. Bewshea, B. Temperton, M.L. Michelsen, J. Warwick-Dugdale, R. Manley, A. Farbos, J.W. Harrison, C.M. Sambles, D.J. Studholme, A.R. Jeffries, A.C. Darby, J.A. Hiscox, S. Paterson, M. Iturriza-Gomara, K.A. Jackson, A.O. Lucaci, E. E. Vamos, M. Hughes, L. Rainbow, R. Eccles, C. Nelson, M. Whitehead, L. Turtle, S. T. Haldenby, R. Gregory, M. Gemmell, D. Kwiatkowski, T.I. de Silva, N. Smith, A. Angyal, B.B. Lindsey, D.C. Groves, L.R. Green, D. Wang, T.M. Freeman, M. D. Parker, A.J. Keeley, P.J. Parsons, R.M. Tucker, R. Brown, M. Wyles, C. Constantinidou, M. Unnikrishnan, S. Ott, J.K.J. Cheng, H.E. Bridgewater, L. R. Frost, G. Taylor-Joyce, R. Stark, L. Baxter, M.T. Alam, P.E. Brown, P.C. McClure, J.G. Chappell, T. Tsoleridis, J. Ball, D. Gramatopoulos, D. Buck, J.A. Todd, A. Green, A. Trebes, G. MacIntyre-Cockett, M. de Cesare, C. Langford, A. Alderton, R. Amato, S. Goncalves, D.K. Jackson, I. Johnston, J. Sillitoe, Steve Palmer, M. Lawniczak, M. Berriman, J. Danesh, R. Livett, L. Shirley, B. Farr, M. Quail, S. Thurston, N. Park, E. Betteridge, D. Weldon, S. Goodwin, R. Nelson, C. Beaver, L. Letchford, D.A. Jackson, L. Foulser, L. McMinn, L. Prestwood, S. Kay, L. Kane, M. J. Dorman, I. Martincorena, C. Puethe, J.-P. Keatley, G. Tonkin-Hill, C. Smith, D. Jamrozy, M.A. Beale, M. Patel, C. Ariani, M. Spencer-Chapman, E. Drury, S. Lo, S. Rajatileka, C. Scott, K. James, S.K. Buddenborg, D.J. Berger, G. Patel, M. V. Garcia-Casado, T. Dibling, S. McGuigan, H.A. Rogers, A.D. Hunter, E. Souster, A. S. Neaverson, Evaluating the effects of SARS-CoV-2 spike mutation D614G on transmissibility and pathogenicity, Cell 184 (2021) 64–75, https://doi.org/10.1016/j.cell.2020.11.020, e11.

[9] L. Zhang, C.B. Jackson, H. Mou, A. Ojha, E.S. Rangarajan, T. Izard, M. Farzan, H. Choe, The D614G mutation in the SARS-CoV-2 spike protein reduces S1 shedding and increases infectivity (preprint), Microbiology (2020), https://doi.org/10.1101/2020.06.12.148726.

[10] E. Volz, S. Mishra, M. Chand, J.C. Barrett, R. Johnson, L. Geidelberg, W.R. Hinsley, D.J. Laydon, G. Dabrera, Á. O'Toole, R. Amato, M. Ragonnet-Cronin, I. Harrison, B. Jackson, C.V. Ariani, O. Boyd, N.J. Loman, J.T. McCrone, S. Gonçalves, D. Jorgensen, R. Myers, V. Hill, D.K. Jackson, K. Gaythorpe, N. Groves, J. Sillitoe, D.P. Kwiatkowski, , The Covid-19 Genomics UK (COG-UK) consortium, S. Flaxman, O. Ratmann, S. Bhatt, S. Hopkins, A. Gandy, A. Rambaut, N.M. Ferguson, Transmission of SARS-CoV-2 Lineage B.1.1.7 in England: insights from linking epidemiological and genetic data (preprint), Infectious Diseases (except HIV/AIDS) (2021), https://doi.org/10.1101/2020.12.30.20249034.

[11] S. Elbe, G. Buckland-Merrett, Data, disease and diplomacy: GISAID's innovative contribution to global health: data, Disease and Diplomacy, Glob. Chall. 1 (2017) 33–46, https://doi.org/10.1002/gch2.1018.

[12] J. Hadfield, C. Megill, S.M. Bell, J. Huddleston, B. Potter, C. Callender, P. Sagulenko, T. Bedford, R.A. Neher, Nextstrain: real-time tracking of pathogen evolution, Bioinformatics 34 (2018) 4121–4123, https://doi.org/10.1093/bioinformatics/bty407.

[13] T. Hayashi, N. Yaegashi, I. Konishi, Effect of RBD mutations in spike glycoprotein of SARS-CoV-2 on neutralizing IgG affinity (preprint). Infectious Diseases (except HIV/AIDS). https://doi.org/10.1101/2021.01.28.21250577, 2021.

[14] Galloway SE, Paul P, MacCannell DR, et al., n.d. Emergence of SARS-CoV-2 B.1.1.7 Lineage — United States, December 29, 2020–January 12, 2021, MMWR Morb Mortal Wkly Rep 2021.

[15] Covid-19 Genomics UK consortium, COG-UK Report on SARS-CoV-2 Spike Mutations of Interest in the UK 15th January 2021, 2021.

[16] E.B. Hodcroft, M. Zuber, S. Nadeau, T.G. Vaughan, K.H.D. Crawford, C.L. Althaus, M.L. Reichmuth, J.E. Bowen, A.C. Walls, D. Corti, J.D. Bloom, D. Veesler, D. Mateo, A. Hernando, I. Comas, F. González Candelas, SeqCOVID-SPAIN consortium, T. Stadler, R.A. Neher, Emergence and spread of a SARS-CoV-2 variant through Europe in the summer of 2020 (preprint), Epidemiology (2020), https://doi.org/10.1101/2020.10.25.20219063.

[17] R. Bayarri-Olmos, A. Rosbjerg, L.B. Johnsen, C. Helgstrand, T. Bak-Thomsen, P. Garred, M.-O. Skjoedt, The SARS-CoV-2 Y453F mink variant displays a pronounced increase in ACE-2 affinity but does not challenge antibody neutralization, J. Biol. Chem. 296 (2021) 100536, https://doi.org/10.1016/j.jbc.2021.100536.

[18] T.N. Starr, A.J. Greaney, S.K. Hilton, D. Ellis, K.H.D. Crawford, A.S. Dingens, M. J. Navarro, J.E. Bowen, M.A. Tortorici, A.C. Walls, N.P. King, D. Veesler, J. D. Bloom, Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding, Cell 182 (2020) 1295–1310, https://doi.org/10.1016/j.cell.2020.08.012, e20.

[19] E.C. Thomson, L.E. Rosen, J.G. Shepherd, R. Spreafico, A. da Silva Filipe, J. A. Wojcechowskyj, C. Davis, L. Piccoli, D.J. Pascall, J. Dillen, S. Lytras, N. Czudnochowski, R. Shah, M. Meury, N. Jesudason, A. De Marco, K. Li, J. Bassi, A. O'Toole, D. Pinto, R.M. Colquhoun, K. Culap, B. Jackson, F. Zatta, A. Rambaut, S. Jaconi, V.B. Sreenu, J. Nix, I. Zhang, R.F. Jarrett, W.G. Glass, M. Beltramello, K. Nomikou, M. Pizzuto, L. Tong, E. Cameroni, T.I. Croll, N. Johnson, J. Di Iulio, A. Wickenhagen, A. Ceschi, A.M. Harbison, D. Mair, P. Ferrari, K. Smollett, F. Sallusto, S. Carmichael, C. Garzoni, J. Nichols, M. Galli, J. Hughes, A. Riva, A. Ho, M. Schiuma, M.G. Semple, P.J.M. Openshaw, E. Fadda, J.K. Baillie, J. D. Chodera, S.J. Rihn, S.J. Lycett, H.W. Virgin, A. Telenti, D. Corti, D.L. Robertson, G. Snell, Circulating SARS-CoV-2 spike N439K variants maintain fitness while evading antibody-mediated immunity, Cell 184 (2021) 1171–1187, https://doi.org/10.1016/j.cell.2021.01.037, e20.

[20] B. Meng, S.A. Kemp, G. Papa, R. Datir, I.A.T.M. Ferreira, S. Marelli, W.T. Harvey, S. Lytras, A. Mohamed, G. Gallo, N. Thakur, D.A. Collier, P. Mlcochova, L. M. Duncan, A.M. Carabelli, J.C. Kenyon, A.M. Lever, A. De Marco, C. Saliba, K. Culap, E. Cameroni, N.J. Matheson, L. Piccoli, D. Corti, L.C. James, D. L. Robertson, D. Bailey, R.K. Gupta, Recurrent emergence of SARS-CoV-2 spike deletion H69/V70 and its role in the variant of concern lineage B.1.1.7, Cell Rep. (2021) 109292, https://doi.org/10.1016/j.celrep.2021.109292.

[21] H. Tegally, E. Wilkinson, M. Giovanetti, A. Iranzadeh, V. Fonseca, J. Giandhari, D. Doolabh, S. Pillay, E.J. San, N. Msomi, K. Mlisana, A. von Gottberg, S. Walaza, M. Allam, A. Ismail, T. Mohale, A.J. Glass, S. Engelbrecht, G. Van Zyl, W. Preiser, F. Petruccione, A. Sigal, D. Hardie, G. Marais, M. Hsiao, S. Korsman, M.-A. Davies, L. Tyers, I. Mudau, D. York, C. Maslo, D. Goedhals, S. Abrahams, O. Laguda-Akingba, A. Alisoltani-Dehkordi, A. Godzik, C.K. Wibmer, B.T. Sewell, J. Lourenço, L.C.J. Alcantara, S.L.K. Pond, S. Weaver, D. Martin, R.J. Lessells, J.N. Bhiman, C. Williamson, T. de Oliveira, Emergence and rapid spread of a new severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2) lineage with multiple spike mutations in South Africa (preprint), Epidemiology (2020), https://doi.org/10.1101/2020.12.21.20248640.

[22] E. Callaway, The coronavirus is mutating — does it matter? Nature 585 (2020) 174–177, https://doi.org/10.1038/d41586-020-02544-6.

[23] P. Koehl, M. Levitt, Sequence variations within protein families are linearly related to structural variations, J. Mol. Biol. 323 (2002) 551–562, https://doi.org/10.1016/S0022-2836(02)00971-3.

[24] H. Liao, W. Yeh, D. Chiang, R.L. Jernigan, B. Lustig, Protein sequence entropy is closely related to packing density and hydrophobicity, Protein Eng. Des. Sel. 18 (2005) 59–64, https://doi.org/10.1093/protein/gzi009.

[25] R. Rao, N. Bhattacharya, N. Thomas, Y. Duan, X. Chen, J. Canny, P. Abbeel, Y. S. Song, Evaluating Protein Transfer Learning with TAPE, 2019.

[26] A. Elnaggar, M. Heinzinger, C. Dallago, G. Rehawi, Y. Wang, L. Jones, T. Gibbs, T. Feher, C. Angerer, M. Steinegger, D. Bhowmik, B. Rost, ProtTrans: towards cracking the language of life's code through self-supervised learning (preprint), Bioinformatics (2020), https://doi.org/10.1101/2020.07.12.199554.

[27] K.E. ArunKumar, D.V. Kalaga, C.M.S. Kumar, M. Kawaji, T.M. Brenza, Forecasting of COVID-19 using deep layer recurrent neural networks (RNNs) with gated recurrent units (GRUs) and long short-term memory (LSTM) cells, Chaos, Solit. Fractals 146 (2021) 110861, https://doi.org/10.1016/j.chaos.2021.110861.

[28] V.K.R. Chimmula, L. Zhang, Time series forecasting of COVID-19 transmission in Canada using LSTM networks, Chaos, Solit. Fractals 135 (2020) 109864, https://doi.org/10.1016/j.chaos.2020.109864.

[29] K. Sarkar, S. Khajanchi, J.J. Nieto, Modeling and forecasting the COVID-19 pandemic in India, Chaos, Solit. Fractals 139 (2020) 110049, https://doi.org/10.1016/j.chaos.2020.110049.

[30] R. Magar, P. Yadav, A. Barati Farimani, Potential neutralizing antibodies discovered for novel corona virus using machine learning, Sci. Rep. 11 (2021) 5261, https://doi.org/10.1038/s41598-021-84637-4.

[31] Y. Wang, P. Yadav, R. Magar, A.B. Farimani, Bio-informed Protein Sequence Generation for Multi-Class Virus Mutation Prediction, bioRxiv, 2020, https://doi.org/10.1101/2020.06.11.146167.

[32] Z. Memon, S. Qureshi, B.R. Memon, Assessing the role of quarantine and isolation as control strategies for COVID-19 outbreak: a case study, Chaos, Solit. Fractals 144 (2021) 110655, https://doi.org/10.1016/j.chaos.2021.110655.

[33] P.C.L. Silva, P.V.C. Batista, H.S. Lima, M.A. Alves, F.G. Guimarães, R.C.P. Silva, COVID-ABS: an agent-based model of COVID-19 epidemic to simulate health and economic effects of social distancing interventions, Chaos, Solit. Fractals 139 (2020) 110088, https://doi.org/10.1016/j.chaos.2020.110088.

[34] K.S. Sharov, Creating and applying SIR modified compartmental model for calculation of COVID-19 lockdown efficiency, Chaos, Solit. Fractals 141 (2020) 110295, https://doi.org/10.1016/j.chaos.2020.110295.

[35] I. Cooper, A. Mondal, C.G. Antonopoulos, A SIR model assumption for the spread of COVID-19 in different communities, Chaos, Solit. Fractals 139 (2020) 110057, https://doi.org/10.1016/j.chaos.2020.110057.

[36] F. Ndaïrou, I. Area, J.J. Nieto, D.F.M. Torres, Mathematical modeling of COVID-19 transmission dynamics with a case study of Wuhan, Chaos, Solit. Fractals 135 (2020) 109846, https://doi.org/10.1016/j.chaos.2020.109846.

[37] R. Wang, J. Chen, K. Gao, Y. Hozumi, C. Yin, G.-W. Wei, Analysis of SARS-CoV-2 mutations in the United States suggests presence of four substrains and novel variants, Commun. Biol. 4 (2021) 1–14, https://doi.org/10.1038/s42003-021-01754-6.

[38] Z. Zhao, B.A. Sokhansanj, C. Malhotra, K. Zheng, G.L. Rosen, Genetic grouping of SARS-CoV-2 coronavirus sequences using informative subtype markers for pandemic spread visualization, PLoS Comput. Biol. 16 (2020), e1008269, https://doi.org/10.1371/journal.pcbi.1008269.

[39] Y. Shu, J. McCauley, GISAID: global initiative on sharing all influenza data – from vision to reality, Euro Surveill. 22 (2017), https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494.

[40] van der Maaten Laurens, Geoffrey Hinton, Visualizing Data using t-SNE, J. Mach. Learn. Res. 9 (2008) 2579–2605.

[41] G.E. Crooks, WebLogo: a sequence logo generator, Genome Res. 14 (2004) 1188–1190, https://doi.org/10.1101/gr.849004.

[42] C.E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (1948) 379–423, https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

[43] P.J.A. Cock, T. Antao, J.T. Chang, B.A. Chapman, C.J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, M.J.L. de Hoon, Biopython: freely available Python tools for computational molecular biology and bioinformatics, Bioinformatics 25 (2009) 1422–1423, https://doi.org/10.1093/bioinformatics/btp163.

[44] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T.L. Scao, S. Gugger, M. Drame, Q. Lhoest, A.M. Rush, HuggingFace's Transformers: State-Of-The-Art Natural Language Processing. ArXiv191003771 Cs, 2020.

[45] A. Bustamam, H. Tasman, N. Yuniarti, I. Frisca, Mursidah, Application of k-means clustering algorithm in grouping the DNA sequences of hepatitis B virus (HBV), in: Presented at the INTERNATIONAL SYMPOSIUM on CURRENT PROGRESS IN MATHEMATICS and SCIENCES 2016 (ISCPMS 2016): Proceedings of the 2nd International Symposium on Current Progress in Mathematics and Sciences 2016, Depok, Jawa Barat, Indonesia, 2017, 030134, https://doi.org/10.1063/1.4991238.

[46] S. Mannor, X. Jin, J. Han, X. Jin, J. Han, X. Jin, J. Han, X. Zhang, K-means clustering, in: C. Sammut, G.I. Webb (Eds.), Encyclopedia of Machine Learning, Springer US, Boston, MA, 2011, pp. 563–564, https://doi.org/10.1007/978-0-387-30164-8_425.

[47] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, G. Varoquaux, API design for machine learning software: experiences from the scikit-learn project, in: ECML PKDD Workshop: Languages for Data Mining and Machine Learning, 2013, pp. 108–122.

[48] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[49] P.J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20 (1987) 53–65, https://doi.org/10.1016/0377-0427(87)90125-7.

[50] J.A. Plante, Y. Liu, J. Liu, H. Xia, B.A. Johnson, K.G. Lokugamage, X. Zhang, A. E. Muruato, J. Zou, C.R. Fontes-Garfias, D. Mirchandani, D. Scharton, J.P. Bilello, Z. Ku, Z. An, B. Kalveram, A.N. Freiberg, V.D. Menachery, X. Xie, K.S. Plante, S. C. Weaver, P.-Y. Shi, Spike mutation D614G alters SARS-CoV-2 fitness, Nature 592 (2021) 116–121, https://doi.org/10.1038/s41586-020-2895-3.

[51] The CITIID-NIHR BioResource Covid-19 Collaboration, The Covid-19 Genomics UK (COG-UK) Consortium, D.A. Collier, A. De Marco, I.A.T.M. Ferreira, B. Meng, R. P. Datir, A.C. Walls, S.A. Kemp, J. Bassi, D. Pinto, C. Silacci-Fregni, S. Bianchi, M. A. Tortorici, J. Bowen, K. Culap, S. Jaconi, E. Cameroni, G. Snell, M.S. Pizzuto, A. F. Pellanda, C. Garzoni, A. Riva, A. Elmer, N. Kingston, B. Graves, L.E. McCoy, K.G. C. Smith, J.R. Bradley, N. Temperton, L. Ceron-Gutierrez, G. Barcenas-Morales, W. Harvey, H.W. Virgin, A. Lanzavecchia, L. Piccoli, R. Doffinger, M. Wills, D. Veesler, D. Corti, R.K. Gupta, Sensitivity of SARS-CoV-2 B.1.1.7 to mRNA vaccine-elicited antibodies, Nature 593 (2021) 136–141, https://doi.org/10.1038/s41586-021-03412-7.

[52] W. Zhang, B.D. Davis, S.S. Chen, J.M. Sincuir Martinez, J.T. Plummer, E. Vail, Emergence of a novel SARS-CoV-2 variant in southern California, J. Am. Med. Assoc. 325 (2021) 1324, https://doi.org/10.1001/jama.2021.1612.

[53] J. Wise, Covid-19: the E484K mutation and the risks it poses, BMJ n359 (2021), https://doi.org/10.1136/bmj.n359.

[54] Z. Liu, L.A. VanBlargan, L.-M. Bloyet, P.W. Rothlauf, R.E. Chen, S. Stumpf, H. Zhao, J.M. Errico, E.S. Theel, M.J. Liebeskind, B. Alford, W.J. Buchser, A.H. Ellebedy, D. H. Fremont, M.S. Diamond, S.P.J. Whelan, Identification of SARS-CoV-2 spike mutations that attenuate monoclonal and serum antibody neutralization, Cell Host Microbe 29 (2021) 477–488, https://doi.org/10.1016/j.chom.2021.01.014, e4.

[55] Y. Huang, C. Yang, X. Xu, W. Xu, S. Liu, Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19, Acta Pharmacol. Sin. 41 (2020) 1141–1149, https://doi.org/10.1038/s41401-020-0485-4.

[56] T.A. Hopf, J.B. Ingraham, F.J. Poelwijk, C.P.I. Schärfe, M. Springer, C. Sander, D. S. Marks, Mutation effects predicted from sequence co-variation, Nat. Biotechnol. 35 (2017) 128–135, https://doi.org/10.1038/nbt.3769.

[57] Rollins Nathan, Brock Kelly, Rollins Joshua, et al., SARS-CoV-2 proteins [WWW document], 2020. https://marks.hms.harvard.edu/sars-cov-2/Spike, 9.13.2021.

[58] J.J. Jacob, K. Vasudevan, A.K. Pragasam, K. Gunasekaran, B. Veeraraghavan, A. Mutreja, Evolutionary tracking of SARS-CoV-2 genetic variants highlights an intricate balance of stabilizing and destabilizing mutations (preprint), Genomics (2020), https://doi.org/10.1101/2020.12.22.423920.