Capturing SARS-CoV-2 Mutations with NLP

Onkar Mayekar *mayekar.onkar1@gmail.com*
Aatika S *sanusuperperfect01@gmail.com*
Vishw Vekariya *vishwvekariya@gmail.com*
Purushotham Vadde *purushotham.vadde93@gmail.com*

***Abstract*—The Transformer model has truly changed the way we work with text data and has yet a chance to prove itself for usages outside the world of language processing. Although it is considered only a tool for translation, this moderate attempt to catch mutations and build a vocabulary has shown that the capabilities of this architecture are remarkable. The results may be biased based on the input that we feed it, but having such a small evaluation loss while still being able to accurately capture mutations justifies our approach and hypothesis that NLP has a place in the Genetics world. Although this is a modest attempt to process natural languages, the next step would be to mix Transformers and GAN networks to create larger datasets to predict and identify new variants. GAN networks have been proven to generate new biological sequences, and similar models have been successful. This research highlights how different fields can combine to create a better understanding of distinct problems.**

I. Introduction

As of December 2019, the world has been facing with a new strain of coronavirus, severe acute respiratory syndrome–coronavirus 2 (SARS-CoV-2) [1], and since its emergence in Wuhan, China it immediately plunged the entire world into an unrelenting pandemic. This pandemic caused by the disease COVID-19 has required a substantial response by governments, hospitals, and all health authorities, and very easily weakened any beliefs of solution foreseen from advanced medicine by shoving the world into one of the most formidable and controversial pandemics. The spread of the virus has been declared as a global public health challenge and has been impacting nearly every aspect of society worldwide. As of March 11th, COVID-19 had been detected in 221 countries, with 117,799,584 confirmed cases and more than 2,615,018 deaths [2]. But this virus will not go away in near feature [3]. Hence, in the last year, many scientists have published multiple studies, papers clamming that they can predict the life span of this virus, but unfortunately many have been proven wrong. Most of the papers that shared different epidemiology models were under the impression that their models are wrong due to the lack of data validity, but this is not the case rather the enormous number of small factors that contribute to the spreading of a virus. Since most papers have tried to describe the world with mathematical models [4] (like SIR, SEIRD, etc.) this approach is always a kind of real-world simplification. If we aim to investigate a phenomenon or something that depends on several factors, the mathematical idea is to separate and simplify in order to understand the impact it has. Models and all kinds of static predictions are not perfect but have proven to be useful in understanding the behavior of the virus and have significantly helped health institutions to prepare their medical staff and increase the resources.

This global problem,SARS-CoV-2 virus, belongs to a family called coronaviruses, which according to the Baltimore Classification [5], belong to a group of viruses commonly referred as (+)ssRNA. This classification is based on the mRNA synthesis. The genetic material of the SARS-CoV-2 virus is a *Positive Sense Single Strand RNA*, where the genome functions as mRNA, so no transcription is required for translation. Most coronaviruses are RNA viruses that cause diseases (similar like the COVID-19) in mammals and birds, and can easily jump from one species to another. This jumping over is one of the enigmas for scientists who are trying to monitor the origin viruses that jump from wildlife to humans, a process which is called zoonotic spillover. [6] Each virus is consisted of particles, and the SARS-CoV-2 viral is structured like most coronaviruses, it has a shell and contains a protected genetic material in form of a single-stranded RNA that is long approximately 27–32 kb (1 kilo base = 1000 bp). All coronaviruses have similar genome organization and expression, and their genome is the largest of all RNA viruses.

Fig. 1. SARS-CoV-2 isolate Wuhan-Hu-1, complete genome [7]

The SARS-CoV-2 genome is composed of multiple structural and nonstructural proteins as shown in Figure 1. The nonstructural are at the 1a/b 5' end, referred to as the open reading frame ORF, the part that is transcribed from DNA / RNA which is consisted of 16 nonstructural proteins labeled NSP1 to NSP16. The structural proteins are the nucleocapsid (N), the spike (S), the envelope (E) and the membrane (M). These structural proteins are encoded by other open reading frames located at the 3'-end.

II. Data

The spread of the SARS-Cov-2 virus is a global problem because viruses do not recognize interstate borders and do not distinguish who they infect. Although countries such as Italy and the America have gone through their greatest waves (December 2021), many countries such as India [8] and Brazil [9] are currently in great waves of new cases and are still experiencing their rise.

But why is the world so interested in tracking down the origin of the virus occurring in their country? The answer to this question is that in order to stop a pandemic of this size one has to understand how the virus replicates is genetic material and how this virus mutates. Because viruses spread to people with all sorts of different DNA phenotypes, the more people it infects, the more likely it is to mutate, and it does this rapidly fast. The tracking of these mutations is of crucial significance because if we were to recognize the virus with all its mutations, we could develop drugs and vaccines that better target it and any other variant of it. If scientist know the of the whereabouts of future mutations before they appear, with antibody synthesis they could develop better vaccines that will recognize the virus and weaken it, a similar approach to how Western medicine is developing HIV drugs [10].

A. Mutations, Variants, Lineages

The approach proposed in this article is a language model that can foresee new mutations by creating a grammar of well know variants. In order to develop this kind of a method that can say whether an amino acid will mutate or not we first have to understand the logic behind mutations, what can they cause and how is the data organized for these amino-acid (aa) changes. Although the global problem of the virus is the rate of mutation, the solution for predicting each subsequent variant or group of sub-variants of the virus depends on biological differences in the genome itself. To us they may seem benign, but to scientists these biological differences in the genome play a key role in vaccine development, because they are thought to occur by chance, and yet they can change the behavior of the virus completely.

Visual representation (Figure 2) and definitions of the mutations terminology that we use:

- **Mutation**: is a change in the coding region itself which can be:
  - *point mutation* (base substitutions): small scale changes that occur in the RNA sequence of 1 to 2 nucleotides. This type of mutation is a nucleotide substitution.
  - *frame shift* (deletions or insertions): refers to the insertion or deletion of nucleotides in the RNA sequence.
- **Variant**: A genome that contains a particular set of mutations
- **Lineage**: All the descendants of a branch of a phylogenetic tree. Within a lineage, there may be additional mutations which revert some changes or accumulate new ones. We primarily rely on PANGO lineages [15] (Phylogenetic Assignment of Named Global Outbreak)

The genome of the SARS-Cov2 virus is a 30 kb long (RNA virus) and like all living organisms the virus is subject to natural selection and it changes. Although this virus has multiple coding regions and each plays a key role when it comes to replication of the virus, I decided to look at all the variants that occur in the **S gene** of the virus, known as the *glycoprotein* (or Spike protein). The selection of this gene is based on its importance [11] to the binding of the host cell. The first contact that a cell has with a viral particle is directed by this gene and its understanding its rate of mutation is potentially crucial.

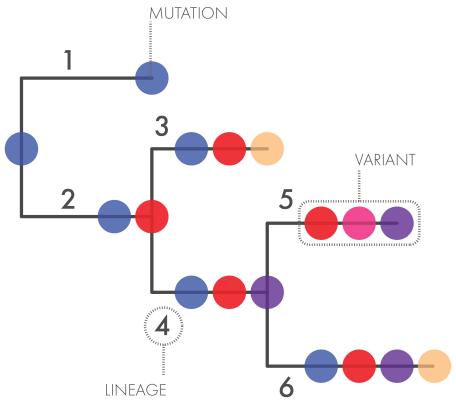

Fig. 2. Glossary of Mutations Terms

B. Data acquisition

The proposed thought of building a language structure in a manner demonstrated the dataset, its last debut is a consequence of the manner in which these models learn and prepare. For the piece of making a dataset with organically right variations, gigantic affirmation goes to the GISAID stage [12], who have put forth enormous attempt to coordinate the information and make wonderful documentation of its elements. There are a few segments on the fundamental GISAID site and every one of them can be helpful somehow or another for the dataset creation. Since our primary center was variations, the information was from the ”Arising Variations” area. This part of the page assists with observing new variations of Coronavirus that might become applicable because of indications of expanded multiplication (assessed by changing the quantity of destinations) in mix with likely consequences for receptor or immunizer restricting, remarked in CoVsurver [13]. At present, 124 amino corrosive changes and glycoprotein (Spike) cancellations happening in no less than 10 unique geographic areas have been distinguished in examinations to actuate immunizer escape, expanded ACE2 [14] restricting, or expanded protein articulation and steadiness. Spike proteins are considered as a component of the development of blends of possible variations to be followed.

Changes displayed in Figure 3, that are with the ”X” augmentation are unclear fundamentals of the relating pages, which may likewise incorporate Spike erase pages. The variations for every month (as per the date of assortment) are positioned by SxC, which is the result of the adjustment of the quantity of locales (contrasted with earlier months; like

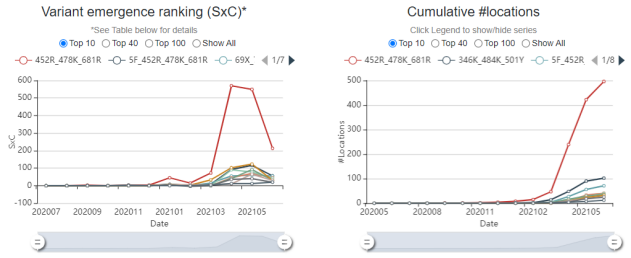

Fig. 3. Variant emergence ranking (SxC) (left), Number of cumulative locations (right); GISAID

the spread of S) and the quantity of important amino corrosive changes with expected impact to blends (C).

The page itself has metadata for each month starting from July 2020 - April 2021. The data is organized in tables for each month separately where there is a larger number of columns (some of which are features in our dataset).

Each of the columns represents a potential information for learning models but we use a reduce the columns dataset, to the following columns:

- **Variant** (string): The name is a combination of the amino acid changes and deletions in the Spike protein that occur in at least 10 different geographical locations and were identified in studies to cause antibody escape, increase ACE2 binding or increase Spike protein expression and stability are considered as part of combinations or constellations forming potential variants to be monitored. These 147 aa changes are the ones found in this "Variant" column
- **Top Lineage** (number + string): Lineage values are the lineage (origin) of each new variant. The lineages are all from the PANGO [15] platform and in the column are consisted of two parts:
  - Largest number of sequenced genomes from this lineage from all genomes
  - Name of the PANGO lineage
- **Co-occurring changes** (list of strings): This is a list of all other amino acid changes that co-occur in more than 75% of all isolates with the variant (combinations of mutations) in the "Variant" column and are listed in this "Co-occurring changes" column. This list contains mutations not only for the glycoprotein but for the other genes of the genome as well (NSP, M etc.) These mutations that are found from GISAID may still be interesting to researchers who are tracking other possible contributing characteristics to these variants. Each element of the list is consisted of:
  - *Protein name* (Spike or other proteins) = In which gene did the change in the virus genome occur (this is the same for everyone)
  - *Amino acid in reference genome* = As in the example above, the first row has T
  - *Position in glycoprotein sequence* = As in the example above, the first row has 19, that is the 19th position in the original glycoprotein
  - *Mutated protein in the new variant* = As in the example above, the first row has R, which means that if there is T19R it means that the protein T at position 19 is replaced by R in this variant
- **(SxC)** (int): Product of the change in the number of sites (Loc or short S) and the number of relevant amino acid changes with potential effect (aachanges or short C)
- **Timestamp** (string): Month and year for the variant

C. Data analysis

The approach that is used to build a gramar from the data is based on a famous technique from the filed of Natural Language Processing called **Transformers**. [16] The idea is to feed a bunch of "sentences" to this type of model in order for it to effectively learn which is a biologically logical (for it is logical grammar) amino acid. So, in order to do that we have to create:

- list of all completely unique emerging variants
- list of unique lineages

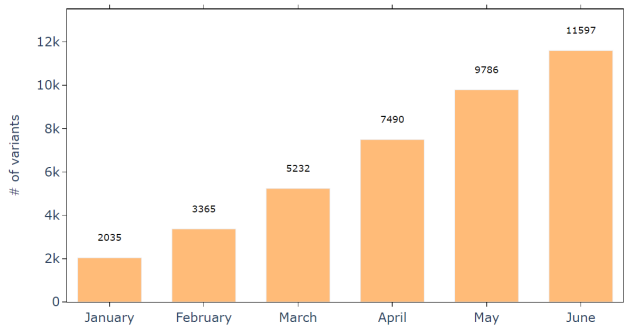

Fig. 4. Number of variants per month

The tables for each month are like the one displayed with Figure 4, and are in a period from July 20 '- June 21' (11 months in total). Figure 6 shows the total number of variants registered for months January 21' - June 21' and we can see that even after a year after the pandemic has started, the numbers between months have a linear growth. From January 21' - 2 035 it jumps to a frightening 11 597 in June 21', which only back ups the common genetic beliefs [17] that RNA viruses mutate a very often.

The total number of variants for the final dataset resulted in 11 937, and out of those variants only 2 224 are completely different. The reason for this reduced number of variants, after getting rid of the duplicates is that most of the variants are often separated into new emerging variants due to mutations to the other genes (proteins), not only the glycoprotein. Since the approach in this paper is to train a model that can distinctly tell apart amino acids of the S-gene, that is why we reduce the number of variants.

After analysing all of the variants we then read the reference gene for each variant and mutate it on the positions based from the "Variant" column and "Co-occurring changes list" column.

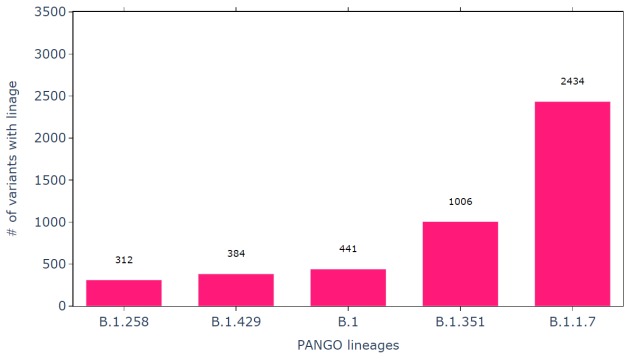

Fig. 5. Five most common lineages

Afterwards we write them in our first smaller training dataset. But, all of these variants have their most common lineage, which is the second value inside the cells from the "Top lineage" column. This column had a major data flaws because most of the rows had wrong values (non-existing lineages) and if total number of lineages is same as the number of variants (11 397) after getting rid of all the duplicates and nonsense inputs the total number of unique lineages came up to 101. Figure 6 and Figure 7 showcases the 5 most common lineages, with the number of their occurrences and an alignment chart.

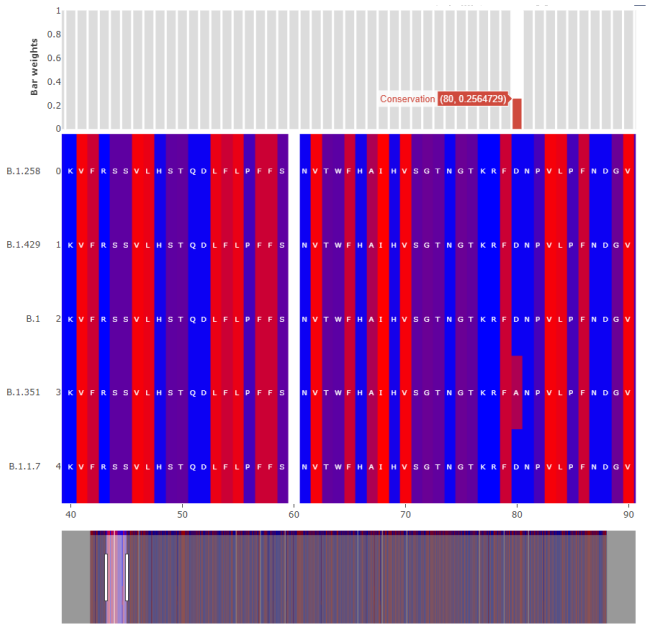

Fig. 6. Alignment chart of the five most common lineages (range base pairs 40-90)

In the number of total occurrences the clear leader is the linage B.1.17 [18] commonly know as the *Alpha variant*. As of 4 July 2021, 962,104 sequences in the B.1.1.7 lineage have been detected [19]. It was first identified in the UK in September 2020 and has since been detected in the US [20] and other countries. This variant is of growing concern because it has shown to be significantly more transmissible than other variants.

The analysis of the lineages finished with a search of all of the mutations since they are not included in our dataset. However, the Center for Viral Systems Biology (outbreak.info) [21] has all of the lineages and their mutations written in JSON files. After manually downloading all of the files, the last step is to mutate all of the amino acids in the reference genome peptide and write them all into our second smaller dataset. After this step the final dataset is consisted of **2 325 samples**.

III. Methods

The current dataset is just RNA sequences, that are read at the input of the model (from the S-gene region: 21563-25384). Although the problem can be solved by classification, for example getting a value for the Sxc column like a score this method gives creates a grammar which is a different approach.

A. Transformers architecture

Transformer models are based on sequence-to-sequence architecture (Seq2Seq), that has been used frequently in the world of natural language processing (NLP). [22], [23] In the background, there is a neural network that transforms a given sequence of elements, a sequence of words into a sentence, and transforms it into a new sequence - sentence. However, these models are primarily used for translation from one language to another, their way of learning led to the idea proposed in this article. Seq2Seq models are composed of *Encoder* and *Decoder*. The encoder reads an input sequence and maps it to a multidimensional vector. This abstract vector then feeds the Decoder which transforms it into an output sequence. This output can be in another language, symbol, copy of the input itself, etc. When I imagine the model, it reminds me of the following scenario: Encoder and Decoder speak two languages, one is their mother tongue different for both (for example German and Macedonian) while their second is an imaginary language and common to both. To translate something from German to Macedonian, the Encoder must convert from a German sentence to an imaginary one, and then the Decoder reads that imaginary one and translates it into Macedonian.

Now why are languages and this whole method useful for predicting virus variants? The answer is again a question: How do we know which part of the sequence is subject to change in the next variant? And here this is solved by the very way architecture teaches. If we imagine a scenario where neither the Encoder nor the Decoder have a common imaginary language, in order to learn it they have to train (the model is trained) with many samples. A simple choice for Encoder and Decoder from such an architecture are single LSTMs for each of the two.

B. Attention

The paper 'Consideration Is All You Want' [24] is presumably the best prologue to the clever engineering called Transformer. As the title demonstrates, it utilizes the consideration system we saw before. Like LSTM, Transformer is a design for

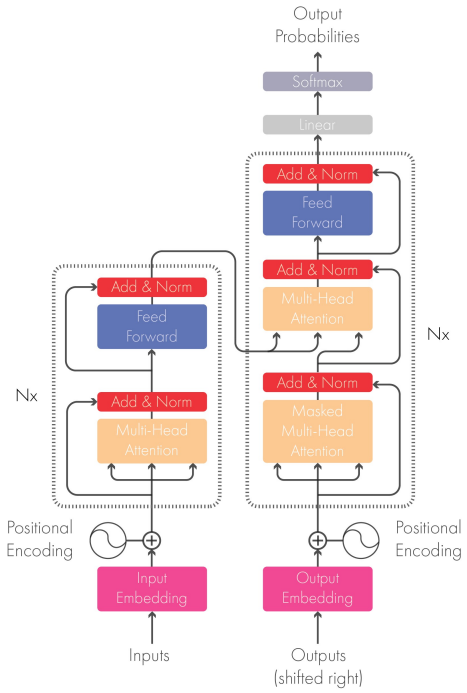

Fig. 7. The Transformer model architecture [24]

changing one arrangement into another with the assistance of two sections (Encoder and Decoder), however it contrasts from the recently depicted/existing succession to-grouping models since it infers no Intermittent Organizations (GRU, LSTM, and so on.).

To make the transformers more obvious, there is one specialized detail that is vital to supplement the thought and that is the *attention detail*. The consideration instrument of the actual model glances at the info arrangement and chooses at each step of realizing which different pieces of the succession are significant. This sounds a piece conceptual, yet a basic model makes a difference: At the point when you read this sentence, you generally center around the word you are perusing and yet your psyche keeps all the substance of the main watchwords in memory so it can give meaning and not be arbitrary.

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \qquad (1)$$

This mechanism works similarly to the sequence. For example, if we are talking about the translation from above, then this mechanism would be the Encoder when translating writes keywords related to the translation that are important for the semantics, and passes them to the Decoder as a plus before starting to translate. These new words make the Decoder's job easier because he knows exactly which parts of the sentence are important so that the context is not lost. Thus, in the learning process, the models put masks (context labels) that are placed on the input sequence itself in the first as a multi-head, and this is actually done to avoid potential "future" errors with the elements of the sequence (more about this in the other stages). These masks would be the mutations themselves for the input variant, which is our feature in the set of input sequences, and we place them so that the decoder can learn accurately in those places about the future sequence.

We should begin with the portrayal of the consideration instrument. It's not exceptionally confounded and can be depicted by condition (1). Q is a framework that contains the question (vector portrayal of single word in the grouping), K are the keys (vector portrayals of the multitude of words in the succession) and V are the qualities, which are again the vector portrayals of the relative multitude of words in the succession. For the encoder and decoder, multi-head consideration modules, V comprises of a similar word grouping than Q. Notwithstanding, for the consideration module that is considering the encoder and the decoder arrangements, V is not the same as the grouping addressed by Q.

After the multi-consideration heads in both the encoder and decoder, we have a pointwise feed-forward layer. This little feed-forward network has indistinguishable boundaries for each position, which can be depicted as a different, indistinguishable direct change of every component from the given succession.

C. Training

How to train such a 'beast'? Training and inferring on Seq2Seq models is a bit different from the usual classification problem. The same is true for Transformers.

Thus in the training process, the models put masks (context labels) that are placed on the input sequence itself in the first as a multi-head, and this is actually done to avoid potential "future" errors with the sequence elements. These masks would be the mutations themselves for the input variant, which is our feature in the set of input sequences, and we place them in order for the decoder to learn accurately in those places about the future sequence.

We are training a new language model from scratch using the Python libraries *Transformers* and *Tokenizers*. We choose to train a byte-level Byte-pair encoding tokenizer (the same as GPT-2), with the same special tokens as RoBERTa [25]. We then pick a vocabulary size of 22 (20 amino acids + token for a beginning and an end). After the vocabulary is set the library saves both a vocab.json, which is a list of the most frequent tokens ranked by frequency, and a merges.txt list of merges. After this the last couple of steps are tokenizing the inputs and spliting the dataset with a 85 (train) - 15 (eval) split, having 1977 samples for training and 348 for evaluation loss.

IV. Results

For trainging the model we used a Google *Colab enviroment* [26], with a Nvidia K80s, with 24 GB of GDDR5 memory. The training took 2h 12min to complete and resulted with a evaluation loss of only 6% ($evaluationloss = 0.0655$).

Beside taking a gander at the preparation and eval misfortunes going down, the least demanding method for checking whether our language model is picking up anything intriguing

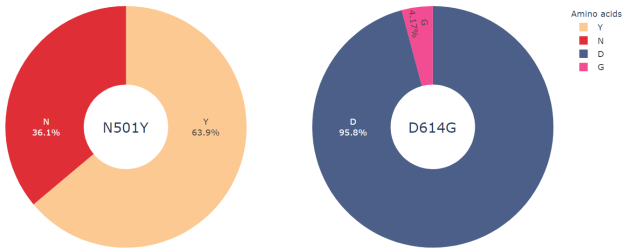

Fig. 8. Frequency of amino acid in dataset sequences at locations 501 and 614 in the Spike protein

is by means of the *FillMaskPipeline*. Pipelines are basic coverings around tokenizers and models, and the 'fill-veil' one will allow you to include a succession containing a concealed token (here, ¡mask¿) and return a rundown of the most likely filled groupings, with their probabilities.

Presently we begin concealing the amino acids areas we want to endlessly check whether our model catches them. We began by veiling the primary area, 0 which is 100% the amino corrosive M, and our model got it with a $probscore = 0.9999$. Yet, this was a simple one so we took a stab at testing one of the 2 most deadly changes [27], which are the N501Y and the D614G.

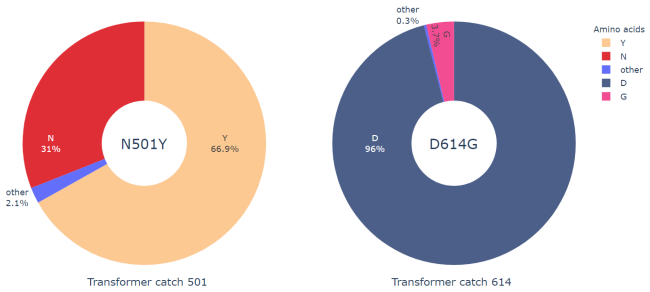

Fig. 9. Transformer catches for frequency of amino acid in dataset sequences at locations 501 and 614 in the Spike protein

Figure 9 shows their recurrence determined from our arrangements dataset.The scores that the model returned are shown in Figure 10, and demonstrate this approach can be promising for possibly getting transformations in the Spike protein. The recurrence of our dataset scores N with 63.9% and Y with 36.1 %, while he viral transformer scores N with 66.9% and Y with 31%.

V. Discussion

The Transformer model has truly changed the way we work with text data and has yet a chance to prove itself for usages outside the world of language processing. Although it is considered only as a tool for translation, this moderate attempt to catch mutations and build a vocabualry has shown that the capabilitites of this architecture are remarkable. The results may be a little bias on the input that we feed it, but to have such a small evaluation loss and still beeing able to catch with close accuracy the mutations justified our approach and hypothesis that NLP has a place in the Genetics world. Although this is a modest attempt to process natural languages, the next step would be to mix Transformers and GAN networks [28] to create larger datasets in order to predict not only catch new variants. GAN networks have been proven to stamp new biologically sequences [29] and similar models have been successful as well. Let these results paint an interesting picture about how easily different fields can mix and match to create better understanding of totally distinct problems.

VI. Conclusion

This study demonstrates the potential of Transformer models as powerful tools for analyzing genetic data, particularly in identifying mutations. By leveraging the Transformer model's ability to capture complex patterns in text, we have shown its capability to process genetic sequences in a way that may reveal meaningful biological insights. Our findings support the hypothesis that NLP methods, traditionally used in language processing, can be effectively applied to genetics, a field that similarly relies on sequence interpretation.

While the current model's accuracy and performance are promising, they are limited by the dataset's size and complexity. Moving forward, integrating Generative Adversarial Networks (GANs) with Transformers offers an exciting opportunity to expand datasets and improve the detection of novel genetic mutations. GANs, which have proven successful in generating biologically plausible data, could significantly enhance the model's robustness and generalizability.

Ultimately, this work contributes to a growing interdisciplinary approach to data science and genomics, underscoring the compatibility of methods across distinct scientific domains. As computational tools like Transformers continue to evolve, their applications across fields will likely become even more impactful, paving the way for innovative solutions to complex challenges in genomics and beyond.

References

[1] K. Yuen, Z. Ye, S. Fung, C. Chan and D. Jin, "SARS-CoV-2 and COVID-19: The most important research questions", Cell & Bioscience, vol. 10, no. 1, 2020. Available: 10.1186/s13578-020-00404-4 [Accessed 5 July 2021].

[2] "outbreak.info", outbreak.info, 2021. [Online]. Available: https://outbreak.info/situation-reports?pango=B.1.1.7&loc=GBR&loc=USA&loc=USA_US-CA&selected=GBR. [Accessed: 01- Jun- 2021].

[3] "Will the Coronavirus Ever Go Away? Here's What a Top WHO Expert Thinks", Time, 2021. [Online]. Available: https://time.com/5805368/will-coronavirus-go-away-world-health-organization/. [Accessed: 01- Jun- 2021].

[4] I. Korolev, "Identification and estimation of the SEIRD epidemic model for COVID-19", Journal of Econometrics, vol. 220, no. 1, pp. 63-85, 2021. Available: 10.1016/j.jeconom.2020.07.038 [Accessed 5 July 2021].

[5] G. Mahmoudabadi and R. Phillips, "A comprehensive and quantitative exploration of thousands of viral genomes", eLife, vol. 7, 2018. Available: 10.7554/elife.31955 [Accessed 5 July 2021].

[6] R. Plowright et al., "Pathways to zoonotic spillover", Nature Reviews Microbiology, vol. 15, no. 8, pp. 502-510, 2017. Available: 10.1038/nrmicro.2017.45 [Accessed 5 July 2021].

[7] "Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, co - Nucleotide - NCBI", Ncbi.nlm.nih.gov, 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/nuccore/NC_045512.2/. [Accessed: 05- Jul- 2021].

[8] [8] www.dw.com), "India tops 200,000 dead amid coronavirus surge — DW — 28.04.2021", DW.COM, 2021. [Online]. Available: https://www.dw.com/en/india-tops-200000-dead-amid-coronavirus [Accessed: 05- Jul- 2021].

[9] "Covid: Brazil has more than 4,000 deaths in 24 hours for first time", BBC News, 2021. [Online]. Available: https://www.bbc.com/news/world-latin-america-56657818. [Accessed: 07- Apr- 2021].

[10] T. Schlub et al., "Fifteen to Twenty Percent of HIV Substitution Mutations Are Associated with Recombination", Journal of Virology, vol. 88, no. 7, pp. 3837-3849, 2014. Available: 10.1128/jvi.03136-13 [Accessed 5 July 2021].

[11] K. Holmes and R. Williams, "Background Paper Functions of Coronavirus Glycoproteins", Advances in Experimental Medicine and Biology, pp. 5-7, 1990. Available: 10.1007/978-1-4684-5823-7_2 [Accessed 5 July 2021].

[12] Elbe, S., and Buckland-Merrett, G. (2017) Data, disease and diplomacy: GISAID's innovative contribution to global health. Available: https://www.gisaid.org/

[13] "GISAID - CoVsurver mutations App", Gisaid.org, 2021. [Online]. Available: https://www.gisaid.org/epiflu-applications/covsurver-mutations-app/. [Accessed: 05- Jul- 2021].

[14] X. Xue et al., "Dynamics of binding ability prediction between spike protein and human ACE2 reveals the adaptive strategy of SARS-CoV-2 in humans", Scientific Reports, vol. 11, no. 1, 2021. Available: 10.1038/s41598-021-82938-2 [Accessed 5 July 2021].

[15] A. Rambaut et al., "Addendum: A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology", Nature Microbiology, vol. 6, no. 3, pp. 415-415, 2021. Available: 10.1038/s41564-021-00872-5.

[16] H. Models, "Transformers In NLP — State-Of-The-Art-Models", Analytics Vidhya, 2021. [Online]. Available: https://www.analyticsvidhya.com/blog/2019/06/understanding-transformers-nlp-state-of-the-art-models/. [Accessed: 05- Jul- 2021].

[17] S. Duffy, "Why are RNA virus mutation rates so damn high?", PLOS Biology, vol. 16, no. 8, p. e3000003, 2018. Available: 10.1371/journal.pbio.3000003.

[18] N. Davies et al., "Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England", Science, vol. 372, no. 6538, p. eabg3055, 2021. Available: 10.1126/science.abg3055 [Accessed 5 July 2021].

[19] "outbreak.info", outbreak.info, 2021. [Online]. Available: https://outbreak.info/situation-reports?pango=B.1.1.7&loc=GBR&loc=USA&loc=USA_US-CA&selected=GBR. [Accessed: 05- Jul- 2021].

[20] Galloway SE, Paul P, MacCannell DR, et al. Emergence of SARS-CoV-2 B.1.1.7 Lineage — United States, December 29, 2020–January 12, 2021. MMWR Morb Mortal Wkly Rep 2021;70:95–99. Available: http://dx.doi.org/10.15585/mmwr.mm7003e2external icon

[21] Julia L. Mullen, Ginger Tsueng, Alaa Abdel Latif, Manar Alkuzweny, Marco Cano, Emily Haag, Jerry Zhou, Mark Zeller, Emory Hufbauer, Nate Matteson, Kristian G. Andersen, Chunlei Wu, Andrew I. Su, Karthik Gangavarapu, Laura D. Hughes, and the Center for Viral Systems Biology outbreak.info. Available: https://outbreak.info/ (2020)

[22] D. Schutte et al., "Discovering novel drug-supplement interactions using a dietary supplements knowledge graph generated from the biomedical literature", arXiv.org, 2021. [Online]. Available: https://arxiv.org/abs/2106.12741v1. [Accessed: 05- Jul- 2021].

[23] S. Pingali, S. Yadav, P. Dutta and S. Saha, "Multimodal Graph-based Transformer Framework for Biomedical Relation Extraction", arXiv.org, 2021. [Online]. Available: https://arxiv.org/abs/2107.00596. [Accessed: 05- Jul- 2021].

[24] A. Vaswani et al., "Attention Is All You Need", arXiv.org, 2021. [Online]. Available: https://arxiv.org/abs/1706.03762. [Accessed: 05- Jul- 2021].

[25] Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach", arXiv.org, 2021. [Online]. Available: https://arxiv.org/abs/1907.11692. [Accessed: 05- Jul- 2021].

[26] Bisong E. (2019) Google Colaboratory. In: Building Machine Learning and Deep Learning Models on Google Cloud Platform. Apress, Berkeley, CA. https://doi.org/10.1007/978-1-4842-4470-8_7

[27] "The delta variant is the most dangerous SARS-CoV-2 mutation yet", The Economist, 2021. [Online]. Available: https://www.economist.com/graphic-detail/2021/06/16/the-delta-variant-is-the-most-dangerous-sars-cov-2-mutation-yet. [Accessed: 05- Jul- 2021].

[28] I. Goodfellow et al., "Generative Adversarial Networks", arXiv.org, 2021. [Online]. Available: https://arxiv.org/abs/1406.2661. [Accessed: 05- Jul- 2021].

[29] D. Repecka et al., "Expanding functional protein sequence spaces using generative adversarial networks", Nature Machine Intelligence, vol. 3, no. 4, pp. 324-333, 2021. Available: 10.1038/s42256-021-00310-5 [Accessed 5 July 2021].